SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 1–4, January 1996

001

AN UNUSUAL MINIMIZATION PRINCIPLE FOR PARABOLIC GRADIENT FLOWS*

LAWRENCE C. EVANS†

Abstract. We show that for a general parabolic system, generated as the gradient flow of a rank-one convex energy functional, the energy at any time is less than or equal to the energy "smoothly sampled" at earlier times. This generalizes the classical assertion that energy cannot increase along flows.

Key words. gradient flow, rank-one convexity, energy decrease

AMS subject classifications. 35K55, 49M10

This short note shows how a small modification of a computation due to Ball and Murat [2] yields an unusual dynamic minimization principle for smooth solutions of certain parabolic partial differential equation (PDE) systems, corresponding to gradient flows governed by appropriate energy functionals.

The set-up is this. Let U denote a smooth, bounded, open subset of $\mathbb{R}^n$. We define the *energy* of a mapping $\mathbf{v} : U \to \mathbb{R}^m$ to be

$$I[\mathbf{v}] = \int_U F(D\mathbf{v}, \mathbf{v}, x)dx, \tag{1.1}$$

where $F : M^{m\times n} \times \mathbb{R}^m \times U \to \mathbb{R}$ is a given smooth function, $F = F(P, z, x)$, and $M^{m\times n}$ denotes the space of real $m \times n$ matrices. We write $x = (x_1, \dots, x_n)$, $\mathbf{v} = (v^1, \dots, v^m)$, and

$$D\mathbf{v} = ((v^i_{x_\alpha}))_{\substack{1\le i\le m\\ 1\le\alpha\le n}}.$$

Below, we implicitly sum repeated Greek indices from 1 to n and Latin indices from 1 to m.

We are concerned with the system of PDE

$$u^i_t - \frac{\partial}{\partial x_\alpha}\left(\frac{\partial F}{\partial p^i_\alpha}(D\mathbf{u}, \mathbf{u}, x)\right) + \frac{\partial F}{\partial z^i}(D\mathbf{u}, \mathbf{u}, x) = 0 \qquad (i = 1, \dots, m) \tag{1.2}$$

in $U \times [0, \infty)$, which is the gradient flow on $L^2(U)$ generated by $I[\cdot]$. Let us henceforth suppose $\mathbf{u} = \mathbf{u}(x, t)$ to be a smooth solution of (1.2), subject to the time-independent boundary conditions

$$\mathbf{u} = \mathbf{g} \quad \text{on } \partial U \times [0, \infty), \tag{1.3}$$

$\mathbf{g} : \partial U \to \mathbb{R}$ being given. We also suppose that the mapping $P \mapsto F(P, z, x)$ is *rank-one convex* for each z, x; this means $f(t) = F(P + t\xi \otimes \eta, z, x)$ is convex for each $P \in M^{m\times n}$, $\xi \in \mathbb{R}^m$, $\eta \in \mathbb{R}^n$, $z \in \mathbb{R}^m$, $x \in U$. The system (1.2) is then parabolic, at least in some weak sense.

The minimization principle is given below.

* Received by the editors May 11, 1994; accepted for publication May 24, 1994. This research was supported in part by National Science Foundation grant DMS–9203440.

† Department of Mathematics, University of California, Berkeley, CA 94720.

THEOREM. *Fix a time $t_0 > 0$ and suppose $\theta : \overline{U} \to [0, t_0]$ is a smooth function. Define*

$$\mathbf{v}(x) = \mathbf{u}(x, \theta(x)) \qquad (x \in \overline{U}). \tag{1.4}$$

Then

$$I[\mathbf{u}(\cdot, t_0)] \leq I[\mathbf{v}]. \tag{1.5}$$

The case $\theta(\cdot) \equiv t \leq t_0$ becomes the classical assertion that the energy $I[\cdot]$ is nonincreasing in time. More generally, (1.5) says that no matter how we "sample" the values of $\mathbf{u}$ at times previous to t_0, we cannot lower the energy below that of $\mathbf{u}(\cdot, t_0)$.

In the illustration (Fig. 1) below, the energy of $\mathbf{u}$ at the top is less than or equal to the energy of $\mathbf{u}$ computed along any curved surface (= the graph of $\theta(\cdot)$), as drawn.

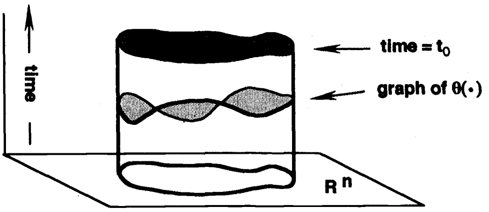


FIG. 1.

Proof. In view of (1.4),

$$v^i_{x_\alpha} = u^i_{x_\alpha} + u^i_t \theta_{x_\alpha} \qquad (1 \leq i \leq m,\ 1 \leq \alpha \leq n).$$

Therefore

$$\begin{aligned} I[\mathbf{v}] &= \int_U F(D\mathbf{u} + \mathbf{u}_t \otimes D\theta, \mathbf{v}, x)dx \\ &\geq \int_U F(D\mathbf{u}, \mathbf{v}, x) + \frac{\partial F}{\partial p^i_\alpha}(D\mathbf{u}, \mathbf{v}, x) u^i_t \theta_{x_\alpha}\, dx, \end{aligned}$$

the inequality a consequence of the rank-one convexity. Hence

$$\begin{aligned} I[\mathbf{v}] - I[\mathbf{u}(\cdot, t_0)] \geq \int_U & F(D\mathbf{u}, \mathbf{v}, x) + \frac{\partial F}{\partial p^i_\alpha}(D\mathbf{u}, \mathbf{v}, x) u^i_t \theta_{x_\alpha} \\ & - F(D\mathbf{u}(\cdot, t_0), \mathbf{u}(\cdot, t_0), x)dx, \end{aligned} \tag{1.6}$$

where $D\mathbf{u}$ and $\mathbf{u}_t$ are evaluated at $(x, \theta(x))$ in the first two integrands. Similarly to

Ball–Murat [2], let us now compute

$$\begin{aligned}
&\frac{\partial}{\partial x_\alpha}\left(\int_{t_0}^{\theta(x)} \frac{\partial F}{\partial p^i_\alpha}(D\mathbf{u}(x,t),\mathbf{u}(x,t),x)u^i_t(x,t)dt\right)\\
&\qquad - \frac{\partial F}{\partial p^i_\alpha}(D\mathbf{u}(x,\theta(x)),\mathbf{v}(x),x)u^i_t(x,\theta(x))\theta_{x_\alpha}(x)\\
&\qquad = \int_{t_0}^{\theta(x)} \frac{\partial}{\partial x_\alpha}\left(\frac{\partial F}{\partial p^i_\alpha}(D\mathbf{u},\mathbf{u},x)\right)u^i_t(x,t)\\
&\qquad + \frac{\partial F}{\partial p^i_\alpha}(D\mathbf{u},\mathbf{u},x)u^i_{x_\alpha t}(x,t)dt\\
&\qquad = \int_{t_0}^{\theta(x)} |\mathbf{u}_t|^2 + \frac{\partial F}{\partial z^i}(D\mathbf{u},\mathbf{u},x)u^i_t + \frac{\partial F}{\partial p^i_\alpha}(D\mathbf{u},\mathbf{u},x)u^i_{x_\alpha t}dt \qquad \text{(by (1.2))}\\
&\qquad = \int_{t_0}^{\theta(x)} |\mathbf{u}_t|^2 + \frac{d}{dt}F(D\mathbf{u},\mathbf{u},x)dt\\
&\qquad = \int_{t_0}^{\theta(x)} |\mathbf{u}_t|^2dt + F(D\mathbf{u}(x,\theta(x)),\mathbf{v}(x),x) - F(D\mathbf{u}(x,t_0),\mathbf{u}(x,t_0),x).
\end{aligned}$$

Plugging this identity into (1.6), we discover

$$I[\mathbf{v}] - I[\mathbf{u}(\cdot,t_0)] \geq \int_U \frac{\partial}{\partial x_\alpha}\left(\int_{t_0}^{\theta} \frac{\partial F}{\partial p^i_\alpha}u^i_t dt\right) - \int_{t_0}^{\theta} |\mathbf{u}_t|^2 dt\, dx.$$

Since $\mathbf{u}_t = 0$ on ∂U according to (1.3), the integral of the divergence term is zero. Consequently

$$I[\mathbf{v}] - I[\mathbf{u}(\cdot,t_0)] \geq \int_U \int_{\theta}^{t_0} |\mathbf{u}_t|^2 dt\, dx \geq 0,$$

as $0 \leq \theta \leq t_0$ on U. □

Remarks. Ball and Murat's calculation [2], following Sivaloganathan [4], simplifies and clarifies aspects of classical field theory in the calculus of variations, to determine when a critical point of $I[\cdot]$ embedded in a one-parameter family of critical points is, in fact, a strong local minimizer. The theorem above is a kind of (very crude) dynamic analogue.

Brezis–Ekeland [3] contains an interesting and completely different minimization principle for gradient flows governed by convex energies, and Auchmuty [1] has a related minimax principle.

Example. As an illustration of the Theorem, take $m = 1$ and set

$$I[v] = \int_U \tfrac{1}{2}|Dv|^2 + H(v)\, dx, \tag{1.7}$$

where $H : \mathbb{R} \to \mathbb{R}$. The corresponding gradient flow is the semilinear heat equation

$$u_t - \Delta u + h(u) = 0 \tag{1.8}$$

on $U \times [0,\infty)$, for $h = H'$.

Let $u^0(\cdot) = u(\cdot,0)$ denote the initial values. Now if

$$-\Delta u^0 + h(u^0) = 0 \qquad \text{in } U,$$

that is, if u^0 is a critical point of $I[\cdot]$, then $u(\cdot,t) = u^0(\cdot)$ for all $t \geq 0$ and assertion (1.5) provides no information. Suppose instead

$$-\Delta u^0 + h(u^0) < 0 \qquad \text{in } \overline{V}, \tag{1.9}$$

V denoting some smooth, open subregion of U. Now, if H is convex (1.9) implies the one-sided minimization principle

$$I[u^0] \leq I[v]$$

for any smooth function v such that $v \leq u^0$ and $v = u^0$ on $U - V$. (Proof: Multiply (1.9) by $u^0 - v \geq 0$, integrate by parts, and use convexity.) However, if H is not convex, as we hereafter assume, we can deduce no minimization principle from the differential inequality (1.9).

But now let us evolve u^0 under the flow governed by the PDE (1.8). In light of (1.8), there exists a small time $t_0 > 0$ such that $u_t > 0$ on $\overline{V} \times [0, t_0]$. The graph of u is consequently rising in V during the time interval $[0, t_0]$ and so sweeps out a region R between the graphs of $u(\cdot, 0)$ and $u(\cdot, t_0)$. (See Fig. 2, below.)

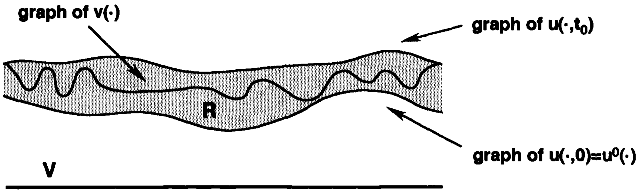


FIG. 2.

Let v be any smooth function whose graph lies within R and agrees with $u(\cdot, t_0)$ on $U - V$. Then according to the Theorem,

$$I[u(\cdot, t_0)] \leq I[v], \tag{1.10}$$

since we can find a smooth mapping $\theta : U \rightarrow [0, t_0]$ such that $v(x) = u(x, \theta(x))$.

In other words, *even for a nonconvex H, the gradient flow evolution causes $u(\cdot, t)$ to become a local one-sided minimizer in any region in which the graph is moving.*

REFERENCES

[1] G. AUCHMUTY, *Saddle points and existence uniqueness for evolution equations*, preprint.

[2] J. BALL AND F. MURAT, *Remarks on rank-one convexity and quasiconvexity*, in Ordinary and Partial Differential Equations, B. D. Sleeman and R. J. Jarvis, eds., Pitman Research Notes, Pitman, Boston, 1991.

[3] H. BREZIS AND I. EKELAND, *Un principle variational associé à certaines équations paraboliques* I *et* II, C. R. Acad. Sci. Paris Sér. I Math., 282 (1976), pp. 971–974 and 1197–1198.

[4] J. SIVALOGANATHAN, *Implications of rank-one convexity*, Ann. Inst. H. Poincaré Anal. Non Linéare, 5 (1988), pp. 99–118.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 5–39, January 1996

EXISTENCE OF PERIODIC SOLUTIONS FOR EQUATIONS OF EVOLVING CURVES*

YOSHIKAZU GIGA† AND NORIKO MIZOGUCHI‡

Abstract. Evolution equations of curvatures of convex curves are considered by the Gauss map parametrization. A time periodic unstable solution is constructed for a reasonable class of time periodic data. Our solution is arranged to satisfy a constraint so that it yields *closed*, embedded, convex curves moving periodically in time (up to translation) whose normal speed equals the curvature minus a given time periodic function depending on curves only through its normals. For curvatures of periodically evolving curves a priori lower and upper bounds depending only on periodic data are obtained. A new penalty method is introduced so that our solution satisfies the constraint. Solutions of penalized equations are constructed by adapting the degree theory.

AMS subject classifications. 35K55, 35K65, 53A04, 73B40

Key words. Harnack's inequality, penalty method, periodic solutions, evolving curves

1. Introduction.

We are concerned with positive, periodic solutions of a quasilinear parabolic equation for curvatures u of evolving curves

$$u_t = u^2(u_{xx} + u - f) \quad \text{in} \quad K, \tag{1.1}$$

with $K = (\mathbf{R}/2\pi\mathbf{Z}) \times (\mathbf{R}/T\mathbf{Z})$ for given $T > 0$. Here $f = f(x,t)$ is a given continuous function on K, i.e., $f \in C(K)$. In particular, f is 2π-periodic in space, i.e., $f(x + 2\pi, t) = f(x,t)$ and T-periodic in time i.e., $f(x, t+T) = f(x,t)$ for all $(x,t) \in \mathbf{R}^2$. Our goal is to construct a positive solution of (1.1) satisfying a constraint

$$\int_0^{2\pi} \frac{e^{ix}}{u(x,t)} dx = 0 \quad \text{for all} \quad t \in \mathbf{R} \quad (i = \sqrt{-1}). \tag{1.2}$$

1.1. Main existence theorem. *Suppose that f is in $C(K)$ together with its time derivative f_t. Then there is a positive function u in $\bigcap_{p>1} W_p^{2,1}(K) \subset C(K)$ (which implies $u_x \in C(K)$) such that u solves* (1.1) *with the constraint* (1.2) *if f is positive on K and satisfies*

$$\int_0^{2\pi} f(x,t) e^{ix} dx = 0 \quad \textit{for all} \quad t \in \mathbf{R}. \tag{1.3}$$

If f is smooth, so is u.

Here $W_p^{2,1}(K)$ denotes the space of all functions g in $L^p(K)$ together with g_x, g_{xx}, and g_t and $L^p(K)$ denotes the space of locally L^p functions in $\mathbf{R}^2$ which is 2π-periodic in x and T-periodic in t. The main existence theorem asserts the existence of a positive solution u, 2π-periodic in x and T-periodic in t.

The restriction (1.3) is necessary if u satisfies (1.2). Indeed, dividing both sides of (1.1) by u^2 yields

$$(-u^{-1})_t = u_{xx} + u - f \quad \text{in} \quad K. \tag{1.4}$$

* Received by the editors March 22, 1994; accepted for publication March 24, 1994.

† Department of Mathematics, Hokkaido University, Sapporo 060, Japan. The research of this author was partially supported by the Inamori Foundation.

‡ Department of Mathematics, Tokyo Gakugei University, Koganei, Tokyo 184, Japan.

Multiplying (1.4) by e^{ix} and integrating by parts over $(0, 2\pi)$ yields

$$-\frac{d}{dt}\int_0^{2\pi}\frac{e^{ix}}{u}dx = -\int_0^{2\pi} fe^{ix}dx \tag{1.5}$$

since $(e^{ix})_{xx} + e^{ix} = 0$. If u satisfies the constraint (1.2), f must satisfy (1.3).

The positivity assumption on f cannot be dropped completely even if f is smooth so that u is smooth. For example let $f \leq 0$ on K. Then at the minimizer $(x_0, t_0) \in K$ of u the equation (1.1) yields

$$0 = u_t(x_0, t_0) = u^2(x_0, t_0)(u_{xx}(x_0, t_0) + u(x_0, t_0) - f(x_0, t_0)) \geq u^3(x_0, t_0) > 0,$$

so there are no positive solutions of (1.1) for nonpositive (smooth) f.

Our main result yields the existence of a periodic-in-time solution (up to translation) for an evolution equation of curves whose normal speed equals the curvature minus a given time periodic function depending on curves through its normals. Let $\{\Gamma_t\}$ be a smooth one-parameter family of closed, embedded curves bounding a domain in the plane. Let $\mathbf{n}$ denote the inward unit normal vector field on Γ_t. Let V denote the normal velocity of Γ_t in the direction of $\mathbf{n}$. We consider an equation for Γ_t of the form

$$V = k - q(\mathbf{n}, t), \tag{1.6}$$

where k is the inward curvature and q is a given function. The equation (1.6) is an example of a curvature flow equation with anisotropy [Gu]. If Γ_t is convex, one can parametrize Γ_t by a Gauss map by introducing θ, $0 \leq \theta \leq 2\pi$, such that $\mathbf{n} = (\cos\theta, \sin\theta)$. The evolution of curvature k is expressed as

$$k_t = k^2(V_{\theta\theta} + V)$$

if we use θ-coordinate [Gu]. Applying this identity to (1.6) yields an evolution equation of curvature

$$k_t = k^2(k_{\theta\theta} + k - (Q_{\theta\theta} + Q)) \quad \text{with} \quad Q(\theta, t) = q(\cos\theta, \sin\theta, t), \tag{1.7}$$

where k and Q are 2π-periodic in θ. We next recover (1.6) from (1.7). For k a curve is given by the Gauss map

$$\mathbf{Z}(\theta, t) = \left(\int_0^\theta \frac{\sin\sigma}{k(\sigma, t)}d\sigma, -\int_0^\theta \frac{\cos\sigma}{k(\sigma, t)}d\sigma\right).$$

If k solves (1.7), then integrating by parts yields

$$\begin{aligned}\frac{\partial \mathbf{Z}}{\partial t} = ((k - Q)\cos\theta - (k_\theta - Q_\theta)\sin\theta - (k - Q)|_{\theta=0},\\ (k - Q)\sin\theta + (k_\theta - Q_\theta)\cos\theta - (k_\theta - Q_\theta)|_{\theta=0}).\end{aligned}$$

Translate $\mathbf{Z}$ by

$$\mathbf{X}_0(t) = \left(\int_0^t (k - Q)(0, \tau)d\tau, \int_0^t (k_\theta - Q_\theta)(0, \tau)d\tau\right),$$

so that new curve $\mathbf{X}(\theta,t) = \mathbf{Z}(\theta,t) + X_0(t)$ fulfills

$$V = \mathbf{n} \cdot \frac{\partial \mathbf{X}}{\partial t} = (\cos\theta, \sin\theta) \cdot \frac{\partial \mathbf{X}}{\partial t} = k - q.$$

We have thus obtained the curve

$$\Gamma_t = \{\mathbf{X}(\theta,t); 0 \le \theta \le 2\pi\}$$

satisfying (1.6). The equations (1.6) and (1.7) are equivalent through $\mathbf{X}$. However, for Γ_t closed we need $\mathbf{X}(0,t) = \mathbf{X}(2\pi,t)$, which is equivalent to the constraint

$$\int_0^{2\pi} \frac{e^{i\theta}}{k(\theta,t)} d\theta = 0.$$

If we set $u = k$, $x = \theta$, this is nothing but the constraint (1.2). Since the condition (1.3) is automatically satisfied for $f = Q_{\theta\theta} + Q$, the main existence theorem yields a periodic-in-time solution Γ_t (up to translation in space) of (1.6).

1.2. Existence of periodically evolving curves. *Suppose that Q is in $C(K)$ together with $Q_{\theta\theta}$, Q_t, $Q_{\theta\theta t}$ and that*

$$Q_{\theta\theta} + Q > 0 \quad \text{on} \quad K. \tag{1.8}$$

Then there are a constant vector $\mathbf{c} \in \mathbf{R}^2$ and a closed curve evolution Γ_t solving

$$V = k - q(\mathbf{n}, t), \tag{1.9}$$

$$\Gamma_{t+T} = \Gamma_t + \mathbf{c} \tag{1.10}$$

for all $t \in \mathbf{R}$. The curvature of Γ_t is always positive and the quantities in (1.9) are continuous. If Q is smooth, so is Γ_t.

In general, translation by $\mathbf{c}$ is necessary. For example, take

$$Q(\theta,t) = \cos\theta + 1$$

so that $Q_{\theta\theta} + Q = 1$. Suppose that Γ_t solves (1.9), (1.10) for some $\mathbf{c}$. Then it is easy to see that $\hat{\Gamma}_t = \Gamma_t + \mathbf{e}_1 t$ with $\mathbf{e}_1 = (1,0)$ solves $V = k - 1$ with T-periodic k. If $f = 1$ in (1.1), by a uniqueness result summarized in Lemma 1.3 the unique solution of (1.1) with (1.3) is $u \equiv 1$. Thus $\hat{\Gamma}_t$ is a circle C of radius 1 and $V = k - 1 = 0$. We now observe that under the relation $Q(\theta,t) = \cos\theta + 1$ if Γ_t solves (1.9), (1.10) (with positive curvature) then $\Gamma_t = C - \mathbf{e}_1 t$ up to a constant vector (independent of time). Thus $\Gamma_{t+T} = \Gamma_t - \mathbf{e}_1 T$ does not agree with Γ_t so translation is necessary in (1.10).

It turns out that the method developed here applies to a general parabolic equation

$$V = a(\mathbf{n})k - q(\mathbf{n},t)$$

including (1.9). This will be discussed in our forthcoming paper [GM].

1.3. Lemma on time-independent solutions. *Suppose that f in (1.1) is time independent. Then so is a positive solution u. If f is time independent there is at most one positive solution u of (1.1) satisfying (1.2).*

We shall prove this uniqueness result in the appendix. The condition (1.8) on Q is equivalent to saying that the curvature of Frank diagram

$$\mathcal{F} = \{(p_1,p_2) \in \mathbf{R}^2; r = 1/Q(\theta), (p_1,p_2) = r(\cos\theta, \sin\theta)\}$$

is positive. This is also discussed in the appendix.

There seems to be no literature on the existence of nontrivial periodic solutions of quasilinear equations, in particular of equations of evolving curves like (1.6), although there are several results for semilinear equations. Besides, it turns out that the periodic solution we seek is (Lyapunov) unstable. In fact all periodic, convex solutions for a general parabolic surface evolution equation are unstable unless the equation depends on location of surface explicitly. This was recently proved by the first author and Yama-uchi [GY]. Even for semilinear equations several methods for constructing periodic solutions only work for stable ones. For example in [AHM] stable periodic solutions were constructed by using subsolutions and supersolutions. For a class of semilinear equations a possibly unstable periodic solution was first constructed in [Am] as the "third" solution. Existence of unstable periodic solutions were recently proved by Hirano and the second author [HM1, HM2] for a class of semilinear parabolic equations. Their method is based on the Leray–Schauder degree theory. Since our equation is quasilinear and not semilinear, their theory does not apply directly to our setting. However we adapt the idea to construct solutions for approximate equations.

Since the equation (1.1) is degenerate at $u = 0$, we wonder whether positive solutions are bounded away from zero. Fortunately, under the constraint (1.2) we have a priori positive lower bounds for u (or curvature).

1.4. Theorem on bounds for curvature. *There are positive constants δ and M depending only on T, $||f||_\infty$, $||f_t||_\infty$, and $\min_K f > 0$ such that if $u \in C^\infty(K)$ satisfies* (1.1) *and* (1.2) *with $u > 0$ and f, $f_t \in C(K)$ with $f > 0$ on K, then $\delta \leq u \leq M$ on K. The constant M is independent of $\min_K f$. Here $||f||_\infty$ denotes the maximum norm of f in $C(K)$.*

To obtain a priori bounds we derive Harnack-type inequalities in our setting. It roughly asserts that

(i) $u(x,t)$ and $u(x,s)$ are comparable for all $t, s \in \mathbf{R}$ provided that u has an upper a priori bound.

(ii) $u(x,t_0)$ is not small compared with the maximum $u(x_0,t_0)$ of u over K.

Such estimates are derived by a differential identity as in [Ga2]. The derivation of the Harnack inequality from differential identities stems from [LY]. For further development of the Harnack inequality and its applications, especially for geometric evolution equations, the reader is referred to [Ha], a recent article of Andrews [And], and references therein. We note that a higher-dimensional version of (1.1) with $f = 0$ is found in [And] as the harmonic mean curvature flow equation.

The equation (1.1) admits a couple of integral bounds of solutions. If maximum of u is large, the Harnack inequality (ii) says its integral must be also large. Thus integral bounds yield an upper bound for solutions. The constraint is not invoked to get an upper bound but without (1.2) we do not get a lower bound. Indeed, suppose that $f = 1$. Then by Remark 1.3 every positive solution of (1.1) is independent of t and it solves

$$u_{xx} + u = 1 \quad \text{in} \quad \mathbf{R}/2\pi\mathbf{Z}.$$

Of course, $u_\sigma(x) = (1-\sigma)\cos x + 1$ $(2 > \sigma > 0)$ is a positive solution but $\{u_\sigma\}$ is not bounded away from zero.

To get a lower bound we integrate (1.4) over $(0,T)$ and obtain

$$U_{xx} + U > 0 \quad \text{in} \quad \mathbf{R}/2\pi\mathbf{Z} \quad \text{with} \quad U(x) = \int_0^T u(x,t)dt.$$

The constraint (1.2) implies

$$\int_0^{2\pi} \sin(x-\zeta)u^{-1}dx = 0 \quad \text{for all} \quad \zeta \in \mathbf{R}.$$

The key observation is that the integrals where the integrand is negative and positive are balanced or both integrals have the same growth order as u tends to zero. By the Harnack inequality (i) such a growth balance is true for U. We shall prove that if U satisfies $U_{xx} + U > 0$ together with integral balance, then U must have a lower bound away from zero. This is in turn gives a lower bound for u by (i). A rigorous proof is given in §4. Although bounds obtained here do not directly apply to construct solutions, the idea of proof is fundamental to finding uniform bounds for solutions of approximate equations.

1.5. Penalty method and approximate equations. We seek a solution of (1.1) satisfying the nonlinear constraint (1.2). Since not all solutions satisfy (1.2), we shall select the desired one by introducing a kind of penalty method. It is heuristically explained as follows. For small $\varepsilon > 0$ we consider a penalized equation

$$(1.11) \qquad u_t = u^2\left(u_{xx} + u + \frac{\varepsilon}{u} - f\right) \quad \text{in} \quad K.$$

For a positive solution u^ε of this equation we observe that (1.3) implies

$$-\frac{d}{dt}\int_0^{2\pi} \frac{e^{ix}}{u^\varepsilon}dx = \varepsilon \int_0^{2\pi} \frac{e^{ix}}{u^\varepsilon}dx$$

in the same way used to derive (1.5). Since u^ε is periodic in time, this implies that our approximate solution $u^\varepsilon > 0$ satisfies the constraint (1.2). It is possible to construct a positive solution of (1.11) mainly because $1/u$ has a strong stabilizing effect near $u = 0$; the derivative of $1/u$ tends to minus infinity as $u \to 0$. However, there is a serious drawback for (1.11). Since we do not know that $f - \varepsilon/u$ is positive, it seems to be impossible to derive a uniform bound from below for u^ε by the method sketched in the previous section. To overcome this difficulty a naive idea may be to replace (1.11) by

$$(1.12) \qquad u_t = u^2\left(u_{xx} + u + \frac{\varepsilon}{u \vee m\varepsilon} - f\right), \quad 0 < \frac{1}{m} < \min f$$

so that $f - \varepsilon(u \vee m\varepsilon)^{-1}$ is positive, where $a \vee b = \max(a,b)$. The solution u^ε of this equation has a priori bounds and it is not difficult to prove that $\{u^\varepsilon\}$ has a convergent subsequence and the limit satisfies (1.1) and (1.2). However, it seems to be difficult to construct a positive solution of (1.12) because the new term $\varepsilon(u \vee m\varepsilon)^{-1}$ is constant near $u = 0$ and has no stabilizing effect for small u.

We adapt the method of finding unstable solutions for semilinear equations developed by Esteban [E1], [E2] as well as Hirano and the second author [HM1], [HM2]. Unfortunately, their method does not apply to solve (1.12). We explain the reason by sketching our method. For fixed $b > 0$ we consider a uniformly parabolic equation

$$(1.13) \qquad u_t = a(u)(u_{xx} - bu + h) \quad \text{in} \quad K$$

where u has now no sign conditions and h is for the moment a given data. If a is continuous, one can prove that the solution operator $S : h \mapsto u$ is well defined in

$C(K)$. If $h = H(u)$ depends on u, a solution of (1.13) with $h = H(u)$ is interpreted as a fixed point of $S \circ H$. If we prove that

$$\deg(I - S \circ H, 0, B_R(0)) = 0,$$
$$\deg(I - S \circ H, 0, B_r(0)) = 1$$

for large ball $B_R(0)$ and small ball $B_r(0)$ centered at zero in $C(K)$, then there is a solution of (1.13) with $h = H(u)$ in $B_R(0) \setminus B_r(0)$. This method is used by Esteban [E1], [E2] for semilinear parabolic equations with superlinear nonlinearity. To implement this program for our equation we should at least modify our equation (1.12) so that it is uniformly parabolic for all $u \in C(K)$. However, not all modification of the u^2 term is good to derive Harnack-type inequalities. Also we should be careful to choose approximate equations so that degree of $I - S \circ H$ has desired properties. We should construct a homotopy of equations in a clever way. For example, to prove the degree in a large ball, equations appearing in the homotopy should be arranged so that solutions have an upper bound uniformly in the homotopy parameter. Our choice of the approximate equation is

(1.14)
$$u_t = (u + \varepsilon^2)^2(u_{xx} - bu + h) \text{ with } h = H(u) = bu + \frac{u^2}{(u+\varepsilon^2)^2}\left(u + \frac{\varepsilon}{\xi(u+\varepsilon^2)} - f\right).$$

Here ξ is a mollified function of $u \vee m\varepsilon$. The multiplier u^2 in H is necessary to calculate local degree near zero since it vanishes faster than u. The multiplier $(u + \varepsilon^2)^2$ is good both for Harnack inequalities and uniform parabolicity. The shift term ε^2 should vanish faster than the parameter ε of penalization so that the limit of approximate solution satisfies the constraint (1.2).

A penalty method is adapted in various evolution equations to introduce constraints of solutions. Rather than to present a huge list we point out one typical example for the harmonic gradient flow equations for mapping into a unit sphere. The requirement of values of mapping u is considered as the constraint $|u| = 1$. A global weak solution was constructed independently by Chen [Ch], Keller et al. [KRS], and Shatah [S] by using a penalty method; see also [CS] for a generalization.

The initial value problem for (1.7) with $q = 0$ was derived in [Ga1] and extensively studied by Gage and Hamilton [GH] for the curve shortening problem. Since a circle shrinks to a point in a finite time for the curve shortening equation ((1.6) with $q = 0$), the curvature may blow up in a finite time. Blow up profiles for convex immersed curves were classified by Angenent [An] based on results of [AL] and [EW] under the self-similar growth assumption for curvatures. It may happen that curvature growth is faster than self-similar rate. Its asymptotic profile is studied in [An] via (1.7) (with $q = 0$). Recently a more precise profile was obtained by Angenent and Velázquez [AV] by studying (1.6) itself. The initial-boundary value problem for a higher-dimensional version of (1.1) (with $f = 0$),

$$u_t = u^2(\triangle u + u), \quad u > 0,$$

in a bounded domain with zero boundary data was studied in [FM] and [Ga2] for positive initial data. The existence of blow up phenomena now depends on the first eigenvalue of the Laplace operator with zero boundary condition. These authors studied whether a solution blows up and they estimated the size of blow up sets. Recently, Wiegner [W] extended the first part

$$u_t = u^\gamma(\triangle u + u)$$

for $\gamma > 2$. There are interesting nonuniqueness results for $\gamma = 1$ by [U] (for the one-dimensional case) and [DL]. Among other results these authors proved that weak solutions may not be unique if initial data takes zero in some open set of the domain because of degeneracy of the equation. They discussed a class of solutions so that uniqueness holds.

This paper is organized as follows. In §2 we derive the Harnack-type inequalities. In §3 an upper bound of solutions is obtained for a general equation so that it applies to our approximate equations. In §4 we establish a lower bound and prove Theorem 1.4. In §5 a lower bound is derived for a general equation. Section 6 is devoted to constructing solutions for approximate equations based on a degree theory. The main theorem is proved there. In §7 we study the property of the solution operator S defined through (1.12). For the existence of S it may be possible to apply the a priori estimate for quasilinear parabolic equations in [LSU]. However, since the space dimension is one, we present a simple proof based only on L^p estimates for linear parabolic equations. In the appendix we give a geometric interpretation of the assumption (1.8); we also give a proof of Lemma 1.3.

2. Harnack-type inequalities. We consider a positive smooth solution u of

$$u_t = u^\gamma(u_{xx} + g(u,x,t)) \quad \text{in} \quad Q = I \times (a,b] \tag{2.1}$$

with $\gamma \in \mathbf{R}$ and smooth g, where I denotes an open interval. We study the behavior of the function $z = (\log u)_t$ to show that the growth of $\log u$ does not become too negative. Such an analysis was done by [Ga2, §2] when $\gamma = 2$ and $g = u$. We adapt his method in our setting. As in [Ga2] our results also extend to the case of several space dimensions, where u_{xx} is replaced with Δu; see §2.5.

For $z = (\log u)_t = u_t/u$ a straightforward calculation yields

$$\begin{aligned} z_x &= \frac{u_{tx}}{u} - \frac{u_t u_x}{u^2}, \\ z_{xx} &= \frac{u_{txx}}{u} - \frac{u_x u_{tx}}{u^2} - \frac{u_{tx}u_x + u_t u_{xx}}{u^2} + 2\frac{u_t u_x^2}{u^3} \\ &= \frac{u_{txx}}{u} - \frac{2u_x z_x}{u} - \frac{u_t u_{xx}}{u^2}. \end{aligned}$$

We thus observe that

$$u_{txx} = u z_{xx} + 2u_x z_x + u_{xx} z.$$

Differentiating $z = u^{\gamma-1}(u_{xx} + g)$ in t yields

$$\begin{aligned} z_t &= u^{\gamma-1}(u_{xxt} + g_u u_t + g_t) + (\gamma - 1)u^{\gamma-2}u_t(u_{xx} + g) \\ &= u^{\gamma-1}(u_{xxt} + g_u uz + g_t) + (\gamma - 1)u^{\gamma-2}uzu^{1-\gamma}z. \end{aligned}$$

Using the expression of u_{xxt} by z's, we end up with a parabolic differential equation for z.

2.1. Differential identity. *Suppose that $u > 0$ solves $u_t = u^\gamma(u_{xx} + g)$ in Q. Then $z = u_t/u$ solves*

$$z_t = u^\gamma z_{xx} + 2u^{\gamma-1}u_x z_x + \gamma z^2 + u^{\gamma-1}(g_u u - g)z + g_t u^{\gamma-1} \quad \textit{in} \quad Q.$$

2.2. Estimate of minimum of z. *Suppose that $u > 0$ solves* (2.1) *in Q with $\gamma > 0$. Suppose that z takes its minimum over Q at $(x_0, t_0) \in Q$. Then*

$$z \geq -\gamma^{-1} u^{\gamma-1}(g_u u - g)_+ - \gamma^{-1/2} u^{(\gamma-1)/2} |g_t|^{1/2} \quad at \quad (x_0, t_0),$$

where f_+ denotes the positive part of f, i.e., $f_+ = f \vee 0$.

Proof. Applying the maximum principle for the differential identity of z yields

$$\gamma z^2 + u^{\gamma-1} G z + g_t u^{\gamma-1} \leq 0 \quad \text{at} \quad (x_0, t_0) \quad \text{with} \quad G = g_u u - g.$$

Since $\gamma > 0$, this quadratic inequality for z implies

$$\begin{aligned} z &\geq -(2\gamma)^{-1}\{u^{\gamma-1} G + (u^{2(\gamma-1)} G^2 - 4\gamma u^{\gamma-1} g_t)^{1/2}\} \\ &\geq -(2\gamma)^{-1}\{u^{\gamma-1} 2G_+ + 2\gamma^{1/2} u^{(\gamma-1)/2} |g_t|^{1/2}\}. \qquad \square \end{aligned}$$

We apply the estimate of $\min z$ to get Harnack-type inequalities. We consider a positive, smooth solution of

$$u_t = u^\gamma (u_{xx} + g(u, x, t)) \quad \text{in} \quad K, \tag{2.2}$$

where $K = (\mathbf{R}/2\pi\mathbf{Z}) \times (\mathbf{R}/T\mathbf{Z})$ and g is a smooth function in each variable.

2.3. Harnack's inequality in time direction. *Assume that $\gamma \geq 1$ and $\alpha \geq 0$. Assume that there are constants c_0, $c_1 > 0$ such that*

$$v g_v(v, x, t) - g(v, x, t) \leq c_0, \qquad |g_t(v, x, t)|^{1/2} \leq c_1 \tag{2.3}$$

for all $(v, x, t) \in (\alpha, \infty) \times K$. Assume that there is a constant $c_2 > 0$ such that $\max_K u \geq c_2$ *for all solutions u of* (2.2) *with $u > \alpha$. Then there is a positive constant $C = C(c_0, c_1, c_2, \gamma) > 0$ such that any solution u of* (2.2) *with $u > \alpha$ fulfills*

$$u(x, t) \leq u(x, t_0) \exp(-C M^{\gamma-1}(t - t_0)), \quad M = \max_K u \tag{2.4}$$

for all (x, t_0), $(x, t) \in K$ with $t_0 - T \leq t \leq t_0$.

Proof. Since K is compact, the minimum of z in K is attained. Using the estimate for $\min_K z$, together with (2.3), we observe that

$$\min_K z \geq -\gamma^{-1} c_0 M^{\gamma-1} - \gamma^{-1/2} c_1 M^{(\gamma-1)/2} \tag{2.5}$$

if $\gamma \geq 1$. Since $M \geq c_2 > 0$, this yields

$$\min_K z \geq -C M^{\gamma-1}$$

with $C = \gamma^{-1} c_0 + \gamma^{-1/2} c_1 c_2^{-(\gamma-1)/2}$. We now obtain a differential inequality

$$-(u^{-1})_t = u_t/u^2 = z/u \geq -C M^{\gamma-1}/u \quad \text{in} \quad K$$

which becomes

$$(u^{-1})_t / u^{-1} \leq C M^{\gamma-1} \quad \text{in} \quad K.$$

For fixed $(x, t_0) \in K$, integrating the differential inequality over (t, t_0) with $t_0 - T \leq t \leq t_0$ yields

$$\log u^{-1}(x, t_0) - \log u^{-1}(x, t) \leq C M^{\gamma-1}(t_0 - t),$$

which is the same as (2.4). □

2.4. Harnack's inequality in space direction. *Assume that $\gamma \geq 1$, $\alpha \geq 0$, and* (2.3) *holds for g. Let u be a solution of* (2.2) *with $u > \alpha$. Let $(x_0, t_0) \in K$ be a maximizer of u. Then*

$$u(x,t_0)^\gamma \geq M^\gamma - \gamma C_M (x-x_0)^2/2 \tag{2.6}$$

with

$$C_M = \frac{c_0}{\gamma} M^{\gamma-1} + \frac{c_1}{\gamma^{\frac{1}{2}}} M^{(\gamma-1)/2} + M^{\gamma-1} g_M, \quad M = \max_K u = u(x_0,t_0),$$
$$g_M = \max\{(g(v,x,t))_+; \alpha < v \leq M, (x,t) \in K\}.$$

Proof. Since $u^{\gamma-1}u_{xx} = z - u^{\gamma-1}g$, it follows from (2.5) that

$$u^{\gamma-1}u_{xx} \geq -C_M \quad \text{in} \quad K. \tag{2.7}$$

This, in particular, implies

$$(u^\gamma)_{xx} = (\gamma u^{\gamma-1}u_x)_x = \gamma(\gamma-1)u^{\gamma-2}u_x^2 + \gamma u^{\gamma-1}u_{xx} \geq -\gamma C_M. \tag{2.8}$$

Integrating over (x_0, x) yields

$$(u^\gamma)_x(x,t_0) \geq (u^\gamma)_x(x_0,t_0) - \gamma C_M(x-x_0) = -\gamma C_M(x-x_0)$$

since (x_0,t_0) is a critical point of u^γ. Integrating again over (x_0,x) yields (2.6). □

2.5. Remarks. The results in §§2.1–2.4 can be extended to a multidimensional problem. Indeed §§2.1 and 2.2 are still valid for

$$u_t = u^\gamma(\Delta u + g(u,x,t)) \quad \text{in} \quad Q = \Omega \times (a,b], \quad \Omega \subset \mathbf{R}^n \tag{2.9}$$

if the first and second derivative in x is replaced by the gradient and the Laplacian, respectively. We consider T-periodic positive solution u of (2.9) with $a = 0$, $b = T$. If u is periodic in space, the same technique yields Harnack's inequality (2.4) provided that $W = \Omega \times (\mathbf{R}/T\mathbf{Z})$ replaces K, where Ω is an n-dimensional torus. However, Harnack's inequality (2.6) should be altered for W. Instead of (2.7) and (2.8), we obtain

$$u^{\gamma-1}\Delta u \geq -C_M, \quad \Delta u^\gamma \geq -\gamma C_M \quad \text{in } W. \tag{2.10}$$

Contrary to the one-dimensional problem, (2.10) does not yield (2.6). Instead of (2.6) we observe that

$$(u^\gamma)_\sharp(r,t) \geq M^\gamma - \gamma C_M r^2/2n, \quad r = |x - x_0| \tag{2.11}$$

for the maximizer (x_0,t_0) of u over W. Here $f_\sharp(r)$ denotes the mean of f on the sphere of radius r centered at x_0, i.e.,

$$f_\sharp(r) = \int_{|\omega|=1} f(r\omega + x_0)d\omega a^{-1}, \quad a = \text{area of the unit sphere.}$$

Indeed, using the identity

$$\triangle_x(f_\sharp(r(x)) = \int_{|\omega|=1} (\triangle f)(r(x)\omega + x_0)d\omega a^{-1}, \quad r(x) = |x - x_0|, \tag{2.12}$$

we observe that $(2.10)_2$ yields

$$(r^{n-1}((u^\gamma)_\sharp)')' \geq -\gamma C_M r^{n-1}$$

since $\triangle_x g(r(x)) = r^{1-n}(r^{n-1}g')'$. Here, the left-hand side of (2.12) is the Laplacian of $f_\sharp(r(x))$ as a function of x. Since (x_0, t_0) is a critical point of $(u^\gamma)_\sharp$, integrating over $(0, r)$ yields $((u^\gamma)_\sharp)' \geq -\gamma C_M r/n$. Integrating over $(0, r)$ again yields (2.11).

Now we must prove (2.12). We may assume $x_0 = 0$. A direct calculation shows

$$\triangle_x(f(r(x)\omega)) = ((\omega \cdot \nabla)(\omega \cdot \nabla)f)(r(x)\omega), \tag{2.13}$$

$$[\nabla - \omega(\omega \cdot \nabla)] \cdot ([\nabla - \omega(\omega \cdot \nabla)]f) = [\triangle - (\omega \cdot \nabla)(\omega \cdot \nabla)]f \tag{2.14}$$

for a given unit n-vector ω. If the left-hand side of (2.14) is evaluated at $r(x)\omega$, it agrees with $\triangle_s g(\omega)$, $g(\omega) = f(r(x)\omega)$, where $\triangle_s$ is the Laplace–Beltrami operator on the unit spheres. By the divergence theorem, integrating (2.14) over the sphere yields

$$0 = \int_{|\omega|=1} [(\triangle - (\omega \cdot \nabla)(\omega \cdot \nabla))f](r(x)\omega)d\omega.$$

This together with (2.13) yields (2.12). □

3. Upper bounds. We shall derive an a priori upper bound for a positive smooth solution $u > \alpha > 0$ of

$$u_t = u^\gamma\{u_{xx} + \varphi(u)(u + \psi(u) - f(x,t))\} \quad \text{in} \quad K. \tag{3.1}$$

Here φ and ψ are smooth functions on (α, ∞) and f is smooth on K, i.e., $f \in C^\infty(K)$ with $f > 0$; $\alpha \geq 0$ and $\gamma \in \mathbf{R}$.

3.1. Lemma on integral bounds. *Assume that* $\psi \geq 0$, $f \geq 0$, *and that* $0 \leq \varphi \leq c_3$, $v - \varphi(v)v \leq c_4$ *on* (α, ∞), $(\alpha \geq 0)$ *with some positive constants* c_3 *and* c_4. *Then*

$$\iint_K u dx dt \leq 2\pi T(c_3 \|f\|_\infty + c_4) \equiv C_1,$$

$$\iint_K \frac{u_t^2}{u^\gamma} dx dt \leq c_3 C_1 \|f_t\|_\infty \equiv C_2$$

holds for all solutions $u \in C^\infty(K)$ *of* (3.1) *with* $u > \alpha$ *on* K.

Proof. Multiplying $u^{-\gamma}$ with (3.1) and integrating by parts on K yields

$$\begin{aligned}\iint_K \frac{u_t}{u^\gamma} dx dt &= \iint_K \{u_{xx} + \varphi(u)(u + \psi(u) - f)\} dx dt \\ &= \iint_K \{\varphi(u)u - \varphi(u)(f - \psi(u))\} dx dt.\end{aligned}$$

Since u is T-periodic, we see

$$\int_0^T \frac{u_t}{u^\gamma} dt = (1-\gamma)^{-1} \int_0^T (u^{1-\gamma})_t dt = 0 \quad \text{for all} \quad x \in \mathbf{R}.$$

We thus observe that

$$\iint_K \varphi(u)udxdt = \iint_K \varphi(u)(f - \psi(u))dxdt \leq \iint_K \varphi(u)fdxdt \leq 2\pi Tc_3||f||_\infty$$

since $\psi \geq 0$ and $\varphi \geq 0$. From $v - \varphi(v)v \leq c_4$ it follows that

$$\iint_K udxdt \leq \iint_K (\varphi(u)u + c_4)dxdt \leq C_1.$$

It remains to prove the second inequality of Lemma 3.1. Multiplying u_t/u^γ with (3.1) and integrating over K yields

$$\iint_K \frac{u_t^2}{u^\gamma}dxdt = \iint_K \{u_{xx}u_t + \varphi(u)(u + \psi(u) - f)u_t\}dxdt.$$

Integrating by parts, we see

$$\iint_K u_{xx}u_tdxdt = -\iint_K u_xu_{xt}dxdt = -\frac{1}{2}\int_0^T \frac{d}{dt}\int_0^{2\pi} u_x^2dxdt = 0$$

by T-periodicity of u_x. We also observe that

$$\int_0^T \varphi(u)(u + \psi(u))u_tdt = 0 \quad \text{for all} \quad x \in \mathbf{R}.$$

We thus conclude by integration by parts in t that

$$\iint_K \frac{u_t^2}{u^\gamma}dxdt = -\iint_K \varphi(u)fu_tdxdt = \iint_K \Phi(u)f_tdxdt$$

by T-periodicity of u, where Φ is a primitive of φ defined by

$$\Phi(v) = \int_0^v \varphi(r)dr$$

so that $\Phi(v)_t = \varphi(v)v_t$ and $\Phi(v) \leq c_3v$. It now follows that

$$\iint_K \frac{u_t^2}{u^\gamma}dxdt \leq ||f_t||_\infty \cdot \iint_K c_3udxdt \leq c_3C_1||f_t||_\infty.$$

This is the same as the second inequality of Lemma 3.1. □

3.2. Upper bound theorem. *Assume that* $1 \leq \gamma < 3$ *and that* $\alpha \geq 0$. *Assume the hypotheses in Lemma* 3.1 *on* f, ψ, φ. *Assume furthermore that*

$$\varphi'(v)(\psi(v) - f) + \varphi(v)\psi'(v) \leq 0, \tag{3.2}$$

$$0 \leq \varphi'(v)v^2 \leq c_5, \quad \varphi(v)(\psi(v) - \min_K f) \leq c_6(v + 1) \tag{3.3}$$

on (α, ∞) *with constants* c_5 *and* $c_6 > 0$. *Then there is a positive constant* $M_0 = M_0(c_j,\ 3 \leq j \leq 6,\ T,\ ||f||_\infty,\ ||f_t||_\infty,\ \gamma)$ *such that if* $u \in C^\infty(K)$ *with* $\min_K u > \alpha$ *solves* (3.1), *then* $\max_K u \leq M_0$.

Proof. We shall combine Harnack's inequality (2.6) with integral bounds. We set

$$g(v, x, t) = \varphi(v)(v + \psi(v) - f(x, t)) \tag{3.4}$$

so that

$$\begin{aligned} g_v v - g &= \{\varphi'(v)(v+\psi(v)-f)+\varphi(v)(1+\psi'(v))\}v - \varphi(v)(v+\psi(v)-f) \\ &\le \varphi'(v)v^2 + \varphi(v)(f-\psi(v)) \quad \text{by} \quad (3.2) \\ &\le c_5 + c_3\|f\|_\infty \equiv c_0 \quad \text{for} \quad v > \alpha \end{aligned}$$

since $\psi \ge 0$. Moreover,

$$|g_t| \le \varphi(v)\|f_t\|_\infty \le c_3\|f_t\|_\infty \equiv c_1^2 \quad \text{for} \quad v > \alpha.$$

Applying Harnack's inequality (2.6), we have

$$u(x,t_0)^\gamma \ge M^\gamma - \gamma C_M(x-x_0)^2/2 \quad \text{for all} \quad x \in \mathbf{R} \tag{3.5}$$

with $M = \max_K u = u(x_0,t_0)$ and

$$C_M = \frac{c_0}{\gamma}M^{\gamma-1} + \frac{c_1}{\gamma^{1/2}}M^{(\gamma-1)/2} + M^{\gamma-1}g_M.$$

Here

$$\begin{aligned} g_M &= \max\{g(v,x,t)_+; \alpha < v \le M,\ (x,t) \in K\} \\ &\le c_3 M + c_6(M+1) \quad \text{by} \quad (3.3)_2. \end{aligned}$$

It follows that

$$C_M \le 2c'M^\gamma \qquad (c' = c_3 + c_6)$$

for large M, say, $M \ge M_1 = M_1(c_0,c_1,c_3,c_6,\gamma)$.

We now integrate (3.5) (assuming $M \ge M_1$) to get

$$\begin{aligned} \int_0^{2\pi} u(x,t_0)^\gamma dx = \int_{x_0-\pi}^{x_0+\pi} u(x,t_0)^\gamma dx &\ge \int_{x_0}^{x_0+\sigma} (M^\gamma - \gamma c' M^\gamma (x-x_0)^2)dx \\ &= M^\gamma \sigma(1-\gamma c'\sigma^2/3) \quad \text{for} \quad 0<\sigma<\pi \\ &\ge M^\gamma\sigma/2 \quad \text{for small} \quad \sigma, \quad \text{say,} \quad \sigma \le (3/2\gamma c')^{1/2} \equiv \sigma_0. \end{aligned}$$

We shall fix $\sigma = \sigma_0$ in the rest of our argument. By Lemma 3.1 there is t_1 such that

$$\begin{aligned} \int_0^{2\pi} u(x,t_1)^\gamma dx &\le \frac{1}{T}\int_0^T \int_0^{2\pi} u(x,t)^\gamma dxdt \\ &\le \frac{M^{\gamma-1}}{T}\iint_K udxdt \le \frac{C_1 M^{\gamma-1}}{T}. \end{aligned}$$

We now observe that

$$\begin{aligned} \int_0^{2\pi} u(x,t)^\gamma dx &= \int_0^{2\pi} u(x,t_1)^\gamma dx + \int_0^{2\pi}\int_{t_1}^t \gamma u^{\gamma-1}u_t(x,\tau)d\tau dx \\ &\le \int_0^{2\pi} u(x,t_1)^\gamma dx + \gamma\left(\iint_K u^{3\gamma-2}dxdt\right)^{1/2}\left(\iint_K \frac{u_t^2}{u^\gamma}dxdt\right)^{1/2} \\ &\le \frac{C_1}{T}M^{\gamma-1} + \gamma C_1^{1/2}M^{(3\gamma-3)/2}C_2^{1/2} \quad \text{for all} \quad t \in \mathbf{R} \end{aligned}$$

since $\iint u_t^2/u^\gamma$ is estimated by Lemma 3.1. Setting $t = t_0$ now yields

$$\frac{M^\gamma \sigma_0}{2} \le \int_0^{2\pi} u(x,t_0)^\gamma dx \le \frac{C_1}{T} M^{\gamma-1} + \gamma C_1^{1/2} C_2^{1/2} M^{(3\gamma-3)/2}.$$

Since $\gamma < 3$, i.e., $(3\gamma - 3)/2 < \gamma$, this inequality for M yields a bound for M, say $M \le M_2 = M_2(\sigma_0, C_1, C_2, T, \gamma)$. We thus conclude that $M \le M_0$ by setting $M_0 = \max(M_1, M_2)$. □

3.3. Remark. The assumptions of Theorem 3.2 are clearly satisfied for $u_t = u^\gamma(u_{xx} + u - f)$ with $f > 0$, which is a particular example of (3.1) found by setting $\varphi \equiv 1$ and $\psi \equiv 0$. We study general equations because we need to have a bound of solutions for solutions of equations which appear in the homotopy of approximate equations. These technical looking assumptions are really useful in what follows.

3.4. Remark. We obtain an upper bound as in §3.2 for the multidimensional problem where K is replaced by W as in §2.5. Note that the integral bounds in §3.1 hold for W with no changes. The only place to be altered in the proof of Theorem 3.2 is the estimate of space integral of u^γ from below since (2.6) should be replaced by (2.11). By (2.11) it holds that

$$\begin{aligned}\int_\Omega u^\gamma(x,t_0)dx &\ge \int_{|x-x_0|<\sigma} u^\gamma(x,t_0)dx \ge \int_0^\sigma r^{n-1}(M^\gamma - \gamma c' M^\gamma r^2/n)dr \\ &\ge M^\gamma \sigma^n/2n \quad \text{for small} \quad \sigma.\end{aligned}$$

The remaining argument is the same as in the proof of Theorem 3.2.

4. Constraints and lower bounds. We consider a positive (periodic) solution $u \in C^\infty(K)$ of (1.1) with the constraint (1.2). Our goal in this section is to establish an a priori lower bound for u when $f > 0$. Since an upper bound is obtained in Theorem 3.2, this will complete the proof of Theorem 1.4. As noted in the introduction, the constraint (1.2) plays an important role for a lower bound. We first study the stationary problem of (1.1),

$$U_{xx} + U = F \quad \text{in} \quad \mathbf{T} = \mathbf{R}/2\pi\mathbf{Z}. \tag{4.1}$$

4.1. Lemma on distance of zeros. *For $a \in \mathbf{R}$ and $d > 0$ suppose that V solves*

$$V_{xx} + V \ge 0 \quad on \quad (a, a+d)$$

with $V(a) = V_x(a) = 0$ and $V(a+d) = 0$. Suppose that

$$V \ge 0 \quad on \quad (a, a+d) \quad and \quad V \not\equiv 0.$$

Then $d > \pi$. Here V_x is assumed to be Lipschitz on $[a, a+d]$.

Proof. We may assume $a = 0$. Since $V(0) = V(d) = 0$, integration by parts yields

$$\int_0^d \sin\left(\frac{\pi x}{d}\right)\left\{V_{xx} + \left(\frac{\pi}{d}\right)^2 V\right\} dx = 0.$$

Suppose that $\pi/d \ge 1$. Since $\sin(\pi x/d) \ge 0$, $V \ge 0$ on $(0,d)$, we now observe that

$$\int_0^d \sin\left(\frac{\pi x}{d}\right)(V_{xx} + V)dx \le \int_0^d \sin\left(\frac{\pi x}{d}\right)\left\{V_{xx} + \left(\frac{\pi}{d}\right)^2 V\right\} dx = 0.$$

Since $V_{xx}+V \geq 0$, this implies

$$V_{xx}+V=0 \quad \text{on} \quad (0,d).$$

By $V(0)=V(d)=0$ we see that

$$V(x)=R\sin(\pi x/d)$$

with some constant R. Since $V_x(0)=0$, this implies $V\equiv 0$, which contradicts $V\not\equiv 0$. We thus proved $\pi/d<1$. □

We next discuss the way to get a lower bound of solutions of (4.1) from constraints. For later convenience we state here a general version. Let $\mu^{\pm}=\{\mu_\varepsilon^{\pm}\}_{\varepsilon>0}$ be a sequence of nonincreasing, continuous, positive functions on $(0,\infty)$ such that $\mu_\varepsilon^{\pm}$ converges to some (nonincreasing, positive,) function $\mu_0^{\pm}$ on $(0,\infty)$. We assume that μ_0^- is continuous in $(0,\infty)$ and that the convergence $\mu_\varepsilon^-\to\mu_0^-$ is uniform in every compact subset of $(0,\infty)$ as $\varepsilon\to 0$. Let $h^-=\{h_\varepsilon^-\}_{\varepsilon>0}$ be a bounded sequence in $L^\infty(0,\infty)$ such that $0\leq h_\varepsilon^-\leq 1$. Assume that h_ε^- converges to $h_0^-\equiv 1$ *uniformly* in every compact subset of $(0,\infty)$ as $\varepsilon\to 0$. For $U>0$ in (4.1) and $\varepsilon\geq 0$ we set

$$A_\varepsilon^{\pm}(\zeta,U)=\int_0^{2\pi}\sin_{\pm}(x-\zeta)\mu_\varepsilon^{\pm}(U)h_\varepsilon^{\pm}(U)dx, \quad \zeta\in\mathbf{R}, \tag{4.2}$$

where $\sin_+ z=\max(\sin z,0)$, $\sin_- z=-\min(\sin z,0)$, and $h_\varepsilon^+\equiv 1$ for $\varepsilon\geq 0$.

4.2. Lemma on a lower bound for stationary solutions. *Let k_j $(0\leq j\leq 4)$ be a positive constant. Suppose that*

$$\int_0^1 \mu_0^-(x^2)dx=\infty. \tag{4.3}$$

Then there are positive constants ε_0, δ_0 depending only on k_j and $\mu^{\pm}$, h^- such that a positive solution $U\in C^2(\mathbf{T})$ of (4.1) satisfies $\min_{\mathbf{T}} U\geq\delta_0$ if U and ε $(0\leq\varepsilon<\varepsilon_0)$ fulfill the following properties:

(U1) $0\leq F\leq k_0$, *where* $F=U_{xx}+U$,

(U2) $k_1\leq\max_{\mathbf{T}} U\leq k_2$,

(U3) $A_\varepsilon^-(\zeta,U)\leq k_3A_\varepsilon^+(\zeta,U)+k_4$ *for all* $\zeta\in\mathbf{R}$.

Proof. We argue by contradiction. Suppose that for some k_j $(0\leq j\leq 4)$ there would exist sequences $\{U_n\}$, $\{\varepsilon_n\}$ $(\varepsilon_n\geq 0)$ satisfying (U1)–(U3) with $U=U_n$, $\varepsilon=\varepsilon_n$, $F=F_n$ such that

$$0<\min_{\mathbf{T}} U_n\to 0 \quad \text{and} \quad \varepsilon_n\to 0 \quad \text{as} \quad n\to\infty. \tag{4.4}$$

Let $x_n\in\mathbf{T}$ be a point such that $\min_{\mathbf{T}} U_n=U_n(x_n)$. We may assume $x_n\to x_0$ for some $x_0\in\mathbf{T}$ by taking a subsequence if necessary. Since F_n and U_n are bounded by (U1) and (U2), the equation (4.1) implies that $\{U_{nxx}\}$ is bounded in $C(\mathbf{T})$. By Ascoli–Arzela's theorem U_n converges to some function U in $C^1(\mathbf{T})$, by taking a subsequence if necessary. Since $\{U_{nxx}\}$ is bounded, U_x is Lipschitz in $\mathbf{T}$ and solves

$$U_{xx}+U\geq 0 \quad (\text{by } F_n\geq 0).$$

By (4.4) we obtain

$$U(x_0)=\lim_{n\to\infty}U_n(x_n)=0$$

since the convergence $U_n \to U$ is uniform. By (U2) we see $U \not\equiv 0$, i.e., there is $x_* \in \mathbf{R}$ with $U(x_*) > 0$. We set

$$a = \sup\{x \in \mathbf{R}; U(x) = 0, x < x_*\} \in \mathbf{R}.$$

Since U takes zero at x_0 this is well defined. By continuity of U_x we see $U(a) = U_x(a) = 0$ and $U(x) > 0$ for $a < x \leq x_*$. We now apply Lemma 4.1 to observe that the distance of another zero (bigger than a) of U and a is more than π. Thus there is a small $0 < \rho_0 < \pi/2$ such that if $0 < \rho < \rho_0$ then

$$\lambda_\rho = \inf\{U(x); a + \rho \leq x \leq a + \rho + \pi\} > 0.$$

We fix $0 < \rho < \rho_0$ and set $\zeta = a + \rho$. Since $U_n \to U$ is uniform, we have a uniform lower bound for U_n,

$$U_n \geq \kappa = \lambda_\rho/2 \quad \text{on} \quad [\zeta, \zeta + \pi] \tag{4.5}$$

for sufficiently large n.

Since U_{xx} is bounded, $U_x(a) = U(a) = 0$ implies that

$$0 \leq U(x) \leq R(x-a)^2 \qquad \text{for all} \quad x \in \mathbf{R} \tag{4.6}$$

with some $R > 0$. From a bound (4.5) it follows that

$$\overline{\lim}_{n\to\infty} A_n^+ = \overline{\lim}_{n\to\infty} \int_\zeta^{\zeta+\pi} \sin(x-\zeta)\mu_{\varepsilon_n}^+(U_n)dx \leq \int_\zeta^{\zeta+\pi} \sin(x-\zeta)dx\mu_0^+(\kappa) = 2\mu_0^+(\kappa),$$

where $A_n^\pm = A_\varepsilon^\pm(\zeta, U_n)$ with $\varepsilon = \varepsilon_n$. By (U3) this implies

$$\overline{\lim}_{n\to\infty} A_n^- \leq 2k_3\mu_0^+(\kappa) + k_4 \equiv L < \infty.$$

Using Fatou's lemma, we observe that

$$\begin{aligned}\int_a^\zeta (-\sin(x-\zeta))\mu_0^-(U)dx &\leq \underline{\lim}_{n\to\infty} \int_a^\zeta (-\sin(x-\zeta))\mu_{\varepsilon_n}^-(U_n)h_{\varepsilon_n}^-(U_n)dx \\ &\leq \underline{\lim}_{n\to\infty} A_n^- \leq L,\end{aligned}$$

since the uniform convergence $\mu_\varepsilon^- \to \mu_0^-$, $h_\varepsilon^- \to h_0^-$ on every compact set in $(0, \infty)$ implies that

$$\lim_{n\to\infty} \mu_{\varepsilon_n}^-(U_n(x)) = \mu_0^-(U(x)), \quad \lim_{n\to\infty} h_{\varepsilon_n}^-(U_n(x)) = 1$$

for $a < x < \zeta$. From (4.6) it follows that

$$\int_a^\zeta (-\sin(x-\zeta))\mu_0^-(R(x-a)^2)dx \leq L.$$

Since $-\sin(x-\zeta) > 0$ on $(\zeta - \pi, \zeta) \ni a$, this contradicts (4.3). We have thus obtained a lower bound δ_0 for U for small $\varepsilon \geq 0$. □

4.3. Corollary. *Let $f \in C(\mathbf{T})$ be a positive function and $\sigma > 0$. There are positive constants δ_0, M_0 (depending only on σ, $\|f\|_\infty$ and $\min_{\mathbf{T}} f$) such that if a positive function $u \in C^2(\mathbf{T})$ solves*

$$u_{xx} + u = f \quad \textit{in} \quad \mathbf{T} \quad \textit{with} \quad \int_0^{2\pi} e^{ix}u^{-\sigma}dx = 0,$$

then $\delta_0 \le u \le M_0$ *provided that* $1/2 \le \sigma$.

Proof. Although an upper bound M_0 is obtained by a general Theorem 3.2, we sketch the proof in this context which reflects an essential step of the proof of Theorem 3.2. Clearly $uu_{xx} = fu - u^2 \ge -M^2$ with $M = \max_{\mathbf{T}} u$. This implies that

$$(u^2)_{xx} = 2u_x^2 + 2uu_{xx} \ge -2M^2,$$

which yields

$$u^2(x) \ge M^2 - M^2(x-x_0)^2, \qquad M = \max_{\mathbf{T}} u = u(x_0).$$

We now integrate to get

$$M^2 \int_{x_0}^{x_0+1} (1-(x-x_0)^2)dx \le \int_0^{2\pi} u^2 dx \le M \int_0^{2\pi} udx.$$

This yields a bound M_0 of M since we observe that

$$\int_0^{2\pi} udx = \int_0^{2\pi} fdx \le 2\pi ||f||_\infty$$

by integrating $u_{xx} + u = f$ on $\mathbf{T}$.

By the maximum principle we have

$$\max u \ge \min f.$$

Since the constraint $\int e^{ix} u^{-\sigma} dx = 0$ is equivalent to

$$\int_0^{2\pi} \sin(x-\zeta) u^{-\sigma} dx = 0 \quad \text{for all} \quad \zeta \in \mathbf{R},$$

(U3) is fulfilled with $k_3 = 1$, $k_4 = 0$, $\mu_\varepsilon^\pm(v) = v^{-\sigma}$, $h_\varepsilon^\pm(v) \equiv 1$ for all $\varepsilon \ge 0$. We observe that (U2) is fulfilled with $k_1 = \min_{\mathbf{T}} f > 0$, $k_2 = M_0$. Since (4.3) is fulfilled for $\mu_0^-(v) = v^{-\sigma}$, we now apply Lemma 4.2 with $k_0 = ||f||_\infty$ to get a desired lower bound δ_0 of u. □

We now study the time-dependent problem (1.1) or its general form

$$u_t = u^\gamma(u_{xx} + u - f) \quad \text{in} \quad K. \tag{4.7}$$

For $u \in C(K)$ it is convenient to define

$$U(x) = \int_0^T u(x,t)dt.$$

Divide (4.7) by u^γ and integrate over $(0,T)$ to get

$$U_{xx} + U = F \quad \text{in} \quad \mathbf{T} \quad \text{with} \quad F = \int_0^T f(x,t)dt, \tag{4.8}$$

since u is T-periodic in time. Harnack's inequality (2.4) allows us to compare u and U.

4.4. Lemma. *Suppose that* $u \in C(K)$ *satisfies* (2.4). *Then there are* λ, $\Lambda > 0$ (*depending only on* C, γ, M, T) *such that* $\lambda u(x,t) \le U(x) \le \Lambda u(x,t)$ *for* $(x,t) \in K$.

Proof. Since u is T-periodic, integrating (2.4) on $(t_0 - T, t_0)$ yields

$$U(x) = \int_{t_0-T}^{t_0} u(x,t)dt \le u(x,t_0)\Lambda, \quad \Lambda = \int_{t_0-T}^{t_0} \exp(-CM^{\gamma-1}(t-t_0))dt;$$

Λ is indepenent of t_0. The inequality $\lambda u \le U$ is similarly obtained by integrating

$$u(x,t_0)\exp(CM^{\gamma-1}(t_0-t)) \le u(x,t) \quad \text{on} \quad t_0 \le t \le t_0 + T. \qquad \square$$

4.5. Proof of Theorem 1.4. Since an upper bound of a positive smooth solution of (1.1) or (4.7) with $1 \le \gamma < 3$ is obtained by Theorem 3.2, it suffices to establish a lower bound assuming the constraint (1.2).

By the maximum principle we see that

$$\max_K u \ge \min_K f \equiv c_2. \tag{4.9}$$

Since Harnack's inequality (2.4) is available for (4.7) with $c_0 = ||f||_\infty$, $c_1 = ||f_t||_\infty$, $\gamma = 2$, Lemma 4.4 now yields

$$\max_T U \ge U(x_0) \ge \lambda u(x_0,t_0) \ge \lambda c_2 \equiv k_1 \tag{4.10}$$

by taking (x_0,t_0) as a maximizer of u over K. From an upper bound M_0 of u, it follows that U fulfills (U2) with $k_2 = M_0 T$. Since $0 \le f \le ||f||_\infty$, we see that (U1) is fulfilled with $k_0 = T||f||_\infty$. If $U \ge \delta_0$ on K, Lemma 4.4 gives a lower bound $\delta = \Lambda^{-1}\delta_0$ for u.

Finally, we check (U3) with (4.3) by choosing $\mu_\varepsilon^\pm(v) = v^{-1}$, $h_\varepsilon^\pm(v) \equiv 1$ for $\varepsilon \ge 0$, so that (4.3) is fulfilled. We shall drop subscript ε since all quantities we handle are independent of ε. We estimate A^- by Lemma 4.4 to get

$$A^-(\zeta, U) \le \lambda^{-1} \int_0^{2\pi} \frac{\sin_-(x-\zeta)}{u(x,t)}dx.$$

By the constraint (1.2) we have

$$\int_0^{2\pi} \frac{\sin_-(x-\zeta)}{u(x,t)}dx = \int_0^{2\pi} \frac{\sin_+(x-\zeta)}{u(x,t)}dx, \quad \zeta \in \mathbf{R}$$

since $\int_0^{2\pi} \sin(x-\zeta)u^{-1}dx = 0$. Using Lemma 4.4 again, we get

$$A^-(\zeta, U) \le \lambda^{-1}\Lambda A^+(\zeta, U),$$

which is (U3) with $k_3 = \lambda^{-1}\Lambda$ and $k_4 = 0$. This enables us to apply Lemma 4.2 to get a lower bound δ_0 for U. If we examine properties of all constants, we conclude that the constant δ_0 has the desired property.

4.6. Generalization. If we examine the proof, it turns out that the assumptions in Theorem 1.4 are weakened. First of all we may replace (1.1) by (4.7) with $1 \le \gamma < 3$. Second, (1.2) may be replaced by

$$\int_0^{2\pi} \frac{e^{ix}}{u^\sigma(x,t)}dx = 0$$

for some $\sigma \ge 1/2$ and some t. Here $\mu_\varepsilon^\pm(v)$ should be $v^{-\sigma}$ in the proof.

5. Approximate penalized equations. To solve (1.1) with (1.2) for positive $f \in C^\infty(K)$ satisfying (1.3) we consider an approximate penalized equation. Let $m > 1$ be a large number such that

$$\min_K f - \frac{1}{m} \geq \frac{1}{2} \min_K f. \tag{5.1}$$

For $0 < \varepsilon,\ \varepsilon' < 1$ we take a smooth nondecreasing function $\xi = \xi_\varepsilon$ on $[0, \infty)$ such that

$$\xi(v) = v \quad \text{for} \quad v \geq m\varepsilon + \varepsilon', \tag{5.2}$$

$$v \vee m\varepsilon \leq \xi(v) \leq \ell(v \vee m\varepsilon) \quad \text{for} \quad v \geq 0 \tag{5.3}$$

$$\text{with some} \quad 1 < \ell < 2 \quad \text{independent of} \quad \varepsilon, \varepsilon', \text{and}\, v.$$

Our approximate penalized equation is of the form

$$w_t = (w + \varepsilon')^2 \left(w_{xx} + \frac{w^2}{(w+\varepsilon')^2} \left(w + \frac{\varepsilon}{\xi(w + \varepsilon')} - f \right) \right) \quad \text{in} \quad K. \tag{5.4}$$

The term $\varepsilon/\xi(w + \varepsilon')$ plays the role of a penalty term. (If we do not need to have smoothness of this term, we may take $\varepsilon/(w + \varepsilon') \vee m\varepsilon$.) The cutting off by $m\varepsilon$ of ξ is introduced so that $f - \varepsilon/\xi(w + \varepsilon')$ is always positive, which will be important to get a lower bound. The parameter ε' is taken so that equation (5.4) is uniformly parabolic, while ε is for penalization. To get a lower bound for w, the penalization effect must dominate the other approximation so ε' should tend to zero faster than ε when $\varepsilon \to 0$. For simplicity we shall take $\varepsilon' = \varepsilon^2$ for the rest of our argument. We set $w + \varepsilon^2 = u$ in (5.4) to get

$$u_t = u^2(u_{xx} + \varphi_\varepsilon(u)(u + \psi_\varepsilon(u) - f - \varepsilon^2)) \quad \text{in} \quad K \tag{5.5}$$

with

$$\varphi_\varepsilon(v) = \frac{(v - \varepsilon^2)^2}{v^2}, \; \psi_\varepsilon(v) = \frac{\varepsilon}{\xi(v)} \quad \text{for} \quad v > 0. \tag{5.6}$$

A direct calculation shows

$$\varphi_\varepsilon(v) = 1 - \frac{2\varepsilon^2}{v} + \frac{\varepsilon^4}{v^2} \tag{5.7}$$

$$\varphi'_\varepsilon(v) = \frac{2\varepsilon^2}{v^2}\left(1 - \frac{\varepsilon^2}{v}\right). \tag{5.8}$$

Since φ_ε and ψ_ε satisfy

(5.9)
$$0 \leq \varphi_\varepsilon(v) \leq 1, \quad 0 \leq \varphi'_\varepsilon(v)v^2 \leq 2\varepsilon^2 \leq 2, \quad v(1 - \varphi_\varepsilon(v)) \leq 2\varepsilon^2 \leq 2, \quad \text{for}\, v > \alpha = \varepsilon^2,$$

$$\min_K f - \psi_\varepsilon(v) > 0, \; \psi'_\varepsilon(v) \leq 0 \quad \text{for} \quad v > \alpha, \tag{5.10}$$

all assumptions in Theorem 3.2 are fulfilled with $c_3 = 1$, $c_4 = 2$, $c_5 = 2$, $c_6 = 0$; to get $(5.10)_1$ we use (5.1)–(5.3).

5.1. Upper bound proposition. *There is a positive constant* $M_0 = M_0(T, ||f||_\infty, ||f_t||_\infty)$ *such that if* $u \in C^\infty(K)$ *solves* (5.5) *with* φ_ε, ψ_ε *satisfying* (5.9), (5.10), *and*

$u > \varepsilon^2$ on K, then $u \le M_0$ on K. Here and hereafter, φ_ε and ψ_ε are assumed to be smooth on $(0, \infty)$.

As in §4 we set

$$U(x) = \int_0^T u(x,t)dt.$$

Applying the maximum principle for (5.5) we see

$$\max_K u \ge \min_K f - \psi_\varepsilon(u) \ge \min_K f - \frac{1}{m} \ge \frac{1}{2}\min_K f \equiv c_2$$

by (5.1)–(5.3). Since other assumptions for Harnack's inequality (2.4) hold by (5.9) and (5.10), U and u are compared by Lemma 4.4.

5.2. Comparison lemma. *Assume that φ_ε and ψ_ε fulfill* (5.9) *and* (5.10) *and that*

$$\min_K f - \psi_\varepsilon(v) \ge c_2 > 0 \quad \text{for} \quad v > \varepsilon^2$$

for some constant c_2. Then there are positive constants λ and Λ (depending only on T, $||f||_\infty$, $||f_t||_\infty$, c_2) such that if $u \in C^\infty(K)$ solves (5.5) *with $u > \varepsilon^2$ on K, then*

$$\lambda u(x,t) \le U(x) \le \Lambda u(x,t) \quad \textit{for} \quad (x,t) \in K. \tag{5.11}$$

We shall derive a uniform lower bound of u independent of ε by assuming (1.3) for f, i.e.,

$$\int_0^{2x} f(x,t)\sin(x-\zeta)dx = 0 \quad \text{for all} \quad \zeta, t \in \mathbf{R}. \tag{5.12}$$

5.3. Lemma on approximate constraints. *Assume that $f \in C^\infty(K)$ satisfies* (5.12). *Assume* $(5.9)_1$ *and*

$$1 - \varphi_\varepsilon(v) \le c_7\varepsilon^2 v^{-1} \quad \text{for} \quad v > \varepsilon^2 \tag{5.13}$$

with some $c_7 > 0$. Then if $u \in C^\infty(K)$ solves (5.5) *with $u > \varepsilon^2$ on K, we have*

$$\left| \int_0^T \int_0^{2\pi} \{\varphi_\varepsilon(u)\psi_\varepsilon(u) + (1 - \varphi_\varepsilon(u))(f + \varepsilon^2)\} \sin(x - \zeta)dxdt \right| \le 4Tc_7\varepsilon^2 \tag{5.14}$$

for all $\zeta \in \mathbf{R}$.

Proof. Since u is T-periodic in t, we have

$$\int_0^T \frac{u_t}{u^2}dt = \int_0^T \left(-\frac{1}{u}\right)_t dt = 0.$$

Since u is 2π-periodic in x, integrating by parts yields

$$\int_0^{2\pi} (u_{xx} + u)\sin(x - \zeta)dx = 0 \quad \text{for all} \quad \zeta \in \mathbf{R}.$$

Multiply (5.5) by u^{-2} to get

$$u_t u^{-2} = u_{xx} + u - (f + \varepsilon^2) + \varphi_\varepsilon(u)\psi_\varepsilon(u) + (1 - \varphi_\varepsilon(u))(f + \varepsilon^2) + (\varphi_\varepsilon(u) - 1)u.$$

Multiplying both sides by $\sin(x-\zeta)$ and integrating over K now yields

$$\int_0^T\int_0^{2\pi}\{\varphi_\varepsilon(u)\psi_\varepsilon(u)+(1-\varphi_\varepsilon(u))(f+\varepsilon^2)\}\sin(x-\zeta)dxdt$$
$$=\int_0^T\int_0^{2\pi}(1-\varphi_\varepsilon(u))u\sin(x-\xi)dxdt.$$

We use (5.13) to get (5.14). □

Since φ_ε in (5.6) satisfies (5.7), property (5.13) with $c_7=2$ holds for φ_ε defined by (5.6). By its form of φ_ε and ψ_ε in (5.6), it is clear that

(5.15)
$$\varphi_\varepsilon(v)\to 1,\ \psi_\varepsilon(v)/\varepsilon\to v^{-1}\ \text{uniformly on every compact subset of } (0,\infty)$$

as $\varepsilon\to 0$.

5.4. Lemma on constraints for the limit of approximate solutions. *Assume that φ_ε and ψ_ε fulfill* $(5.9)_1$, (5.13), *and* (5.15). *Assume that $u_\varepsilon\in C^\infty(K)$ $(u_\varepsilon>\varepsilon^2)$ solves* (5.5) *with $f=f_\varepsilon\in C^\infty(K)$ satisfying* (5.12). *Assume that $||f_\varepsilon||_\infty$ is bounded for $0<\varepsilon<1$. Suppose that u_ε converges to a positive $u\in C(K)$ uniformly in K. Then*

$$\int_0^T\int_0^{2\pi}\sin(x-\zeta)u^{-1}(x,t)dxdt=0\quad\text{for all}\quad \zeta\in\mathbf{R}. \tag{5.16}$$

Proof. Dividing (5.14) ($f=f_\varepsilon$, $u=u_\varepsilon$) by $0<\varepsilon<1$ and using (5.13), we observe that

(5.17)
$$\left|\int_0^T\int_0^{2\pi}\sin(x-\zeta)\varphi_\varepsilon(u_\varepsilon)\psi_\varepsilon(u_\varepsilon)\varepsilon^{-1}dxdt\right|\le c_7\varepsilon+c_7||f_\varepsilon+1||_\infty\int_0^T\int_0^{2\pi}\frac{\varepsilon}{u_\varepsilon}dxdt.$$

Since $u>0$, u_ε is bounded away from zero, say $u_\varepsilon\ge\delta_0>0$ uniformly for sufficiently small $\varepsilon>0$. The right-hand side of (5.17) now converges to zero as $\varepsilon\to 0$. By the convergence (5.15), sending ε to zero in (5.17) yields (5.16), since $u_\varepsilon\ge\delta_0$. □

The basic strategy to get a lower bound for u in (5.5) is similar to the proof of Theorem 1.4. This time we should check assumptions of Lemma 4.2 by using our approximate constraint (5.14). For this purpose we note a property of φ_ε of $(5.6)_1$. The assumptions on φ_ε in the following lemma are automatically satisfied for φ_ε of $(5.6)_1$.

5.5. Lemma. *Assume that φ_ε fulfills* (5.9) *and* (5.15). *Let Λ_0 be a positive number. There exists a sequence of continuous functions h_ε, $0\le h_\varepsilon\le 1$ on $(0,\infty)$ such that h_ε converges to one uniformly in every compact subset of $(0,\infty)$ as $\varepsilon\to 0$ and that*

$$\int_0^T\varphi_\varepsilon(v(t))dt\ge Th_\varepsilon(V),\quad V=\int_0^T v(t)dt$$

for all $v\in C[0,T]$ with $v>\varepsilon^2$ and $V\le\Lambda_0 v(t)$, $0\le t\le T$.

Proof. We set a continuous function

$$h_\varepsilon(\rho)=\varphi_\varepsilon(\rho\Lambda_0^{-1}\vee\varepsilon^2),\quad \rho>0$$

and observe that $h_\varepsilon \to 1$ uniformly in every compact subset of $(0,\infty)$ and that $0 \leq h_\varepsilon \leq 1$. Since φ_ε is nondecreasing on (ε^2,∞) by $(5.9)_2$, we thus conclude

$$\int_0^T \varphi_\varepsilon(v(t))dt \geq \int_0^T h_\varepsilon(V)dt = Th_\varepsilon(V). \qquad \square$$

We set

$$I_\varepsilon^\pm = I_\varepsilon^\pm(\zeta,u) = \varepsilon^{-1}\int_0^T\int_0^{2\pi}\{\varphi_\varepsilon(u)\psi_\varepsilon(u) + (1-\varphi_\varepsilon(u))(f+\varepsilon^2)\}\sin_\pm(x-\zeta)dxdt.$$

The approximate constraint (5.14) is rewritten as

$$|I_\varepsilon^+ - I_\varepsilon^-| \leq 4Tc_7\varepsilon. \tag{5.18}$$

5.6. Proposition on estimate of $A_\pm$. *Assume that $u \in C(K)$ satisfies the comparison* (5.11) *with U and $u > \varepsilon^2$ on K and that $f \geq 0$.*

(i) *Suppose that φ_ε fulfills* (5.9) *and* (5.15). *Suppose that $\psi_\varepsilon(v) \geq \varepsilon\nu_\varepsilon(v)/\ell$ for $v > 0$, where $\nu_\varepsilon(v) = 1/(v \vee m\varepsilon)$ with some $m > 0$, $\ell > 0$. Let h_ε be as in Lemma* 5.5 *and set $h_\varepsilon^- = h_\varepsilon$ and $\mu_\varepsilon^-(v) = \nu_\varepsilon(\lambda^{-1}v)$. Let A_ε^- be defined by* (4.2). *Then*

$$TA_\varepsilon^-(\zeta,U) \leq \ell I_\varepsilon^-(\zeta,u), \quad \zeta \in \mathbf{R}.$$

(ii) *Suppose that φ_ε satisfies* $(5.9)_1$ *and* (5.13). *Suppose that ψ_ε satisfies $\psi_\varepsilon(v) \leq \varepsilon v^{-1}$ for $v > 0$. Let A_ε^+ be defined by* (4.2) *with $\mu_\varepsilon^+(v) = v^{-1}$. Then*

$$I_\varepsilon^+(\zeta,u) \leq T\Lambda(1+\varepsilon c_7\|f+\varepsilon^2\|_\infty)A_\varepsilon^+(\zeta,U), \quad \zeta \in \mathbf{R}.$$

Proof. (i) Since $\ell\varepsilon^{-1}\psi_\varepsilon(u) \geq \nu_\varepsilon(u) \geq \mu_\varepsilon^-(U)$ we see

$$\begin{aligned} I_\varepsilon^- &\geq \varepsilon^{-1}\int_0^T\int_0^{2\pi}\varphi_\varepsilon(u)\psi_\varepsilon(u)\sin_-(x-\zeta)dxdt \quad (\text{by } f \geq 0, \varphi_\varepsilon \leq 1) \\ &\geq \ell^{-1}\int_0^{2\pi}\left(\int_0^T \varphi_\varepsilon(u)dt\right)\mu_\varepsilon^-(U)\sin_-(x-\zeta)dxdt. \end{aligned}$$

Apply Lemma 5.5 to get $\ell I_\varepsilon^- \geq TA_\varepsilon^-$.

(ii) We estimate φ_ε, $\psi_\varepsilon\varepsilon^{-1}$, $1-\varphi_\varepsilon$ from above to get

$$\begin{aligned} I_\varepsilon^+ &\leq \int_0^T\int_0^{2\pi}\left(\frac{1}{u} + \frac{\varepsilon c_7}{u}\|f+\varepsilon^2\|_\infty\right)\sin_+(x-\zeta)dxdt \\ &\leq T\Lambda(1+\varepsilon c_7\|f+\varepsilon^2\|_\infty)\int_0^{2\pi}\frac{1}{U}\sin_+(x-\zeta)dx \end{aligned}$$

which completes the proof. $\square$

5.7. Lower bound theorem. *Assume that $f \in C^\infty(K)$ fulfills* (5.12) *with $f > 0$. Let φ_ε and ψ_ε be defined by* (5.6) *with* (5.2), (5.3), *and $m > 1$. Then there are positive constants ε_0, δ depending only on T, $\|f\|_\infty$, $\|f_t\|_\infty$, $\min_K f$ such that if $u \in C^\infty(K)$ solves* $(5.5)_\varepsilon$ *with m satisfying* (5.1), *then $u \geq \delta$ on K for $0 < \varepsilon < \varepsilon_0$ provided that $u > \varepsilon^2$ on K.*

Proof. Since φ_ε and ψ_ε in (5.6) fulfill all assumptions on φ_ε and ψ_ε in §§5.1–5.6 including Proposition 5.6, we may apply all these results. By the comparison lemma (Lemma 5.2), it suffices to apply Lemma 4.2 to get a lower bound for U for sufficiently small ε.

We shall prove (U1)–(U3) in Lemma 4.2. Dividing (5.5) by u^2 and integrating over $(0,T)$ yields

$$\begin{aligned}
&0 = U_{xx} + U - F \quad \text{on} \quad \mathbf{T},\\
&F = \int_0^T \{(f + \varepsilon^2 - \psi_\varepsilon(u))\varphi_\varepsilon(u) + (1 - \varphi_\varepsilon(u))u\}dt.
\end{aligned}$$

Since $f - \psi_\varepsilon(u) > 0$ by (5.10) and $\varphi_\varepsilon(u) \leq 1$ we see $F \geq 0$. Moreover, by (5.13) with $c_7 = 2$, we observe that

$$F \leq \int_0^T \left(\|f + \varepsilon^2\|_\infty + \frac{2\varepsilon^2}{u}u\right) dt = T(\|f\|_\infty + 3) \equiv k_0 \quad (0 < \varepsilon < 1),$$

which proves (U1). By (5.1)–(5.3) applying the maximum principle yields

$$\max_K u (= u(x_0, t_0)) \geq \frac{1}{\sqrt{2}} \min_K f \equiv c_2$$

as in §5.1. By Lemma 5.2 we see

$$\max_{\mathbf{T}} U \geq U(x_0) \geq \lambda u(x_0, t_0) \geq \frac{\lambda}{2} \min_K f \equiv k_1 > 0$$

(cf. the proof of (4.10)). Since $u \leq M_0$ by Proposition 5.1, (U2) is now fulfilled with $k_2 = TM_0$.

It remains to show (U3) since $\mu_\varepsilon^\pm$, h_ε^- satisfy the desired properties including (4.3). Applying the approximate constraint (5.18) to Proposition 5.6 we observe that

$$\begin{aligned}
A_\varepsilon^-(\zeta, U) &\leq T^{-1}\ell I_\varepsilon^-(\zeta, u) \leq 2T^{-1} I_\varepsilon^-(\zeta, u)\\
&\leq 2T^{-1}(I_\varepsilon^+(\zeta, u) + 8T\varepsilon) \leq 2T^{-1} I_\varepsilon^+(\zeta, u) + 2 \cdot 8\\
&\leq 2\Lambda(1 + 2(\|f\|_\infty + 1))A_\varepsilon^+(\zeta, U) + 16
\end{aligned}$$

where we have used $0 < \varepsilon < 1$, $\ell < 2$, $c_7 = 2$. We have thus proved (U3). The proof is now complete. □

6. Existence of periodic solutions. We shall construct a positive solution $w \in C^\infty(K)$ of the approximate penalized equation (5.4) (with $\varepsilon' = \varepsilon^2$) for arbitrary positive $f \in C^\infty(K)$. We always assume (5.1)–(5.3) and positivity of $f \in C^\infty(K)$ in this section. Also we fix $\varepsilon > 0$ except in the proof of the main theorem given in §6.8. We begin with a solvability result for a uniformly parabolic equation since (5.4) is uniformly parabolic for $w > 0$. For $\omega > 0$, $T > 0$ we set

$$K_\omega = (\mathbf{R}/\omega\mathbf{Z}) \times (\mathbf{R}/T\mathbf{Z}).$$

We consider a uniform parabolic equation

$$(6.1) \qquad v_t = a(v)(v_{xx} - bv + h) \quad \text{in} \quad K_\omega.$$

6.1. Unique solvability lemma. *Assume that b is a positive constant and that a is a continuous function on $\mathbf{R}$ such that $a(\sigma) \geq a_0 > 0$ for all $\sigma \in \mathbf{R}$ with some constant a_0. For each $h \in C(K_\omega)$ there is a unique solution $v \in \bigcap_{q>1} W_q^{2,1}(K_\omega) \subset C(K_\omega)$ of (6.1). Moreover, the solution operator $h \mapsto v$ is a continuous, compact operator from $C(K_\omega)$ into itself. There exist $\theta_0 = \theta_0(a_0, ||h||_\infty, b, \omega, T) > 0$ and $C_0 = C_0(a_0, ||h||_\infty, b, \omega, T) > 0$ such that estimate*

$$||v||_{W_p^{2,1}} \leq C_0 ||h||_\infty$$

holds for all $2 \leq p \leq 2 + \theta_0$ and $h \in C(K_\omega)$.

We postpone the proof until the next section.

6.2. Mappings. We next interpret a solution w of (5.4) as a fixed point of a mapping. Let $b > 0$ be taken so that

$$\phi(w,x,t) \equiv bw_+ + \frac{(w_+)^2}{(w_+ + \varepsilon^2)^2}\left(w_+ + \frac{\varepsilon}{\xi(w_+ + \varepsilon^2)} - f(x,t)\right) \geq 0, \qquad w_+ = w \vee 0,$$

for all $w \in \mathbf{R}$, $(x,t) \in K$, and $\phi > 0$ on K if $w > 0$. It is possible to choose $b > 0$ because $(w_+)^2$ vanishes *faster* than w_+. Indeed, $w \leq 0$ implies $\phi \equiv 0$. If $w \geq \max_K f > 0$, then $\phi > 0$ for any $b > 0$ since $\xi > 0$. If otherwise (but $w > 0$), we observe that

$$\begin{aligned}\phi(w,x,t) &= \frac{w}{(w+\varepsilon^2)^2}\left\{b(w+\varepsilon^2)^2 + w^2 + \left(\frac{\varepsilon}{\xi(w+\varepsilon^2)} - f\right)w\right\} \\ &\geq \frac{w}{(w+\varepsilon^2)^2}(2b\varepsilon^2 w - fw) > 0\end{aligned}$$

provided that $b > ||f||_\infty / 2\varepsilon^2$.

For this choice of b let S denote the solution operator of

$$v_t = (v_+ + \varepsilon^2)^2(v_{xx} - bv + h) \quad \text{in} \quad K,$$

i.e., $S(h) = v$. By Lemma 6.1 S is well defined as an operator from $C(K)$ into itself. Lemma 6.1 also yields

(i) $S : C(K) \to C(K)$ is compact and continuous.

(ii) $S(h)$ is Hölder continuous on K for $h \in C(K)$.

The second assertion follows from the Sobolev embedding since $S(h) \in W_p^{2,1}(K)$ for all $p > 1$. Using (ii) and the Schauder theory of linear parabolic equations [LSU] we observe that

(iii) $v = S(h) \in C^{2,1}(K)$ (i.e., $v, v_{xx}, v_t \in C(K)$) provided that h is Hölder continuous in K.

By the strong maximum principle for classical solutions we see

(iv) $v = S(h) > 0$ in K if $h \geq 0$ and $h \not\equiv 0$, where h is Hölder continuous.

For $u \in C(K)$ we define a continuous operator $H : C(K) \to C(K)$ by

$$H(u)(x,t) \equiv \phi(u(x,t), x, t).$$

Since ϕ is locally Lipschitz in w, we observe that

(v) if u is Hölder continuous in K so is $H(u)$.

Suppose that $w \in C(K)$ is a fixed point of $S \circ H$, i.e., $S \circ H(w) = w$. Then by (ii) w is Hölder continuous. This implies $S(H(w)) \in C^{2,1}(K)$ by (iii), (v). Since $\phi > 0$ for $w > 0$,

(vi) $H(u) \geq 0$ and $H(u) \not\equiv 0$ unless $u \equiv 0$, so $w = S(H(w)) > 0$ in K if $w \not\equiv 0$ by (iv).

The equivalence of (5.4) and $S \circ H(w) = w$ is formally easy. We now obtain a positive solution $w \in C^{2,1}(K)$ of (5.4). Of course, this solution is in $C^\infty(K)$ since $f \in C^\infty(K)$ by the higher regularity theory [LSU].

6.3. Lemma on the local degree near zero. *There is a constant* $r_0 > 0$ *such that the (Leray–Schauder) degree of* $I - S \circ H$ *of the value zero in* $B_r(0)$ *equals one for* $0 < r < r_0$, *i.e.,*

$$\deg(I - S \circ H, B_r(0), 0) = 1,$$

where I *denotes the identity operator and*

$$B_r(0) = \{h \in C(K); ||h||_\infty \leq r\}.$$

Proof. Since S is compact and S, H are continuous, the degree of $I - S \circ \tau H$ is well defined for $0 \leq \tau \leq 1$. Since $S(0) = 0$ by $b > 0$, the homotopy invariance of degree implies

$$\deg(I - S \circ H, B_r(0), 0) = \deg(I, B_r(0), 0) = 1$$

provided that

$$(S \circ \tau H)(u) \neq u \tag{6.2}$$

for all $u \in \partial B_r(0)$, $0 \leq \tau \leq 1$ if r is sufficiently small.

It remains to prove (6.2). We argue by contradiction. We may assume that there are $\{\tau_n\}$ $(0 < \tau_n \leq 1)$ and $\{u_n\}$ such that $u_n \not\equiv 0$ and

$$(S \circ \tau_n H)(u_n) = u_n \quad \text{with} \quad ||u_n||_\infty \to 0 \quad \text{as} \quad n \to \infty.$$

By §6.2 (iv)–(vi) we observe that $u_n > 0$ and solves

$$u_t = (u + \varepsilon^2)^2(u_{xx} - bu + \tau H(u)) \quad \text{in} \quad K \tag{6.3}$$

with $\tau = \tau_n$ in a classical sense. By (5.1)–(5.3) we see

$$f - \frac{\varepsilon}{\xi(u_n + \varepsilon^2)} \geq \frac{1}{2} \min_K f > 0.$$

Since $||u_n||_\infty \to 0$, this implies

$$G_n \equiv f - \frac{\varepsilon}{\xi(u_n + \varepsilon^2)} - u_n > 0 \quad \text{for large} \quad n.$$

Multiplying $(6.3)_\tau$ by $(u + \varepsilon^2)^{-2}$ and integrating over K with $u = u_n$ and $\tau = \tau_n$ yields

$$0 = \int_0^T \int_0^{2\pi} u_n \left\{ b(1 - \tau_n) + \frac{\tau_n u_n}{(u_n + \varepsilon^2)^2} G_n \right\} dxdt$$

as in the proof of Lemma 3.1. Since $0 < \tau_n \leq 1$, $b > 0$, and $u_n > 0$, this implies $G_n = 0$. This contradicts the positivity of G_n for large n. We have proved (6.2). □

6.4. Lemma on the degree in a large ball. *There is an* $R_0 > 0$ *such that*

$$\deg(I - S \circ H, B_R(0), 0) = 0 \quad \textit{for} \quad R > R_0.$$

Proof. We set

$$\Phi(w) = bw_+ + \frac{w^2}{(w_+ + \varepsilon^2)^2} w_+ + 1 \quad \text{for} \quad w \in \mathbf{R}$$

and its affine homotopy to H

$$\Phi_\tau(w) = \tau H(w) + (1 - \tau)\Phi(w), \ 0 \le \tau \le 1$$

for $w \in C(K)$. We first observe that

$$\deg(I - S \circ \Phi, B_R(0), 0) = 0 \quad \text{for} \quad R > 0.$$

Indeed, since Φ_τ also satisfies (v) and (vi) of §6.2 with $H = \Phi_\tau$, if $S \circ \Phi(u) = u$, then $u > 0$ and it is a $C^{2,1}$ solution of

$$u_t = (u + \varepsilon^2)^2 \left(u_{xx} + \frac{u^2}{(u + \varepsilon^2)^2} u + 1 \right) \quad \text{in} \quad K.$$

At the minimum point of u we see $u_t = 0$ and $u_{xx} \ge 0$ but this is impossible by the equation for u. So there is no fixed point of $S \circ \Phi$.

By the homotopy invariance of degree the proof will be complete if we prove the a priori bound (indepenent of τ) for w satisfying $S \circ \Phi_\tau(w) = w$ so that we set

$$R_0 = \sup\{||w||_\infty; S \circ \Phi_\tau(w) = w, 0 \le \tau \le 1\}.$$

As in §6.2, $S \circ \Phi_\tau(w) = w$ implies that w is positive and it is a $C^{2,1}$ solution of

$$w_t = (w + \varepsilon^2)^2 \{w_{xx} - bw + \Phi_\tau(w)\} \quad \text{in} \quad K. \tag{6.4}$$

Since f is smooth, so is w. The following a priori bound is sufficient for our purpose.

6.5. Upper bound lemma. *There is a constant* $R_1 > 0$ *such that if* $w \in C^\infty(K)$ *with* $w > 0$ *solves* (6.4), *then* $w \le R_1$ *in* K.

Proof. In (6.4) we set $w + \varepsilon^2 = u$ to get

$$u_t = u^2[u_{xx} + \varphi_\varepsilon(u)\{u + \tau\psi_\varepsilon(u) + (1 - \tau)/\varphi_\varepsilon(u) - \tau f - \varepsilon^2\}]$$

with φ_ε, ψ_ε satisfying (5.6). If we denote

$$\varphi = \varphi_\varepsilon, \quad \psi = \tau\psi_\varepsilon + (1 - \tau)/\varphi_\varepsilon,$$

then $u > \varepsilon^2$ solves

$$u_t = u^2(u_{xx} + \varphi(u)\{u + \psi(u) - \tau f - \varepsilon^2\}.$$

It is sufficient to check the assumptions of Theorem 3.2 by setting $\alpha = \varepsilon^2$. By (5.9) all we have to prove are

$$\varphi'(v)(\psi(v) - \tau f - \varepsilon^2) + \varphi(v)\psi'(v) \le 0, \quad \text{for } v > \varepsilon^2, \tag{6.5}$$

$$\varphi(v)(\psi(v) - \min_K(\tau f + \varepsilon^2)) \le c_6(v + 1), \quad \text{for } v > \varepsilon^2. \tag{6.6}$$

The left-hand side of (6.5) is

$$\begin{aligned}&\varphi'(\tau(\psi_\varepsilon - f) - \varepsilon^2 + (1-\tau)\varphi^{-1}) + \varphi(\tau\psi_\varepsilon' + (1-\tau)(\varphi^{-1})')\\ &\quad \le \varphi'\varphi^{-1}(1-\tau) + \varphi(\varphi^{-1})'(1-\tau) = (1-\tau)(\varphi'\varphi^{-1} - \varphi\varphi'\varphi^{-2}) = 0\end{aligned}$$

since $\varphi' \ge 0$, $\psi_\varepsilon - f < 0$, $\psi_\varepsilon' \le 0$ on $[\varepsilon^2, \infty)$ by (5.1)–(5.3); here $\varphi^{-1} = 1/\varphi$. The left-hand side of (6.6) is

$$\varphi(\tau(\psi_\varepsilon - \min_K f) + (1-\tau)\varphi^{-1} - \varepsilon^2) \le \varphi\varphi^{-1}(1-\tau) \le 1.$$

We have thus proved (6.5) and (6.6) (with $c_6 = 1$). □

6.6. Remark. For the semilinear equation

$$\begin{aligned} u_t &= \Delta u + m(t)u^p && \text{in} \quad (0,T)\times\Omega,\\ u &= 0 && \text{on} \quad (0,T)\times\partial\Omega \end{aligned}$$

with T-periodic m, Esteban [E1], [E2] obtained a positive T-periodic solution u provided that $1 < p < n/(n-2)$ where Ω is a smoothly bounded domain in $\mathbf{R}^n$. The key argument is to prove similar statements for Lemmas 6.4 and 6.5. The method to get an upper bound in [E1], [E2] is based on a blow up argument. This is used in [Gi] to get a bound for positive solutions of the initial-boundary value problem

$$u_t = \Delta u + u^p \quad \text{in} \quad (0,\infty)\times\Omega$$

with $u = 0$ on $(0,\infty)\times\partial\Omega$ for $1 < p < (n+2)/(n-2)$. However, it seems to be difficult to obtain an upper bound by this method in our problem. The application of Harnack's inequality to get a bound seems to be a new approach.

6.7. Existence Theorem for approximate penalized equation. *Assume that $f \in C^\infty(K)$ and $f > 0$. Then for every $\varepsilon > 0$ there is a positive solution $w \in C^\infty(K)$ of* (5.4) *with $\varepsilon' = \varepsilon^2$ and* (5.1)–(5.3).

Proof. By Lemmas 6.3 and 6.4 we have

$$\deg(I - S\circ H, B_R(0)\setminus B_r(0), 0) = -1$$

for $r < R$ with $r < r_0$, $R > R_0$. It now follows that there exists $w \not\equiv 0$ such that $S\circ H(w) = w$. This yields a desired solution w of equation (5.4) as remarked at the end of §6.2. □

6.8. Proof of Theorem 1.1. We approximate f by $f_\varepsilon \in C^\infty(K)$ satisfying (1.3) such that

$$f_\varepsilon \to f \text{ in } C(K) \text{ as } \varepsilon \to 0 \quad \text{with} \quad \|f_{\varepsilon t}\|_\infty \le 2\|f_t\|_\infty,\ \|f_\varepsilon\|_\infty \le 2\|f\|_\infty, 0 < \varepsilon < 1.$$

Of course, this is possible by convoluting f with a mollifier in space and a mollifier in time so that (1.3) is inherent to f_ε.

Let w_ε be a solution of the approximate penalized equation (5.4) (with $\varepsilon' = \varepsilon^2 > 0$ and $f = f_\varepsilon$) obtained in Theorem 6.7. By Proposition 5.1 and Theorem 5.7 we have $\delta \le w_\varepsilon + \varepsilon^2 \le M_0$ with M_0, $\delta > 0$ independent of sufficiently small $\varepsilon > 0$, say, $\varepsilon < \varepsilon_0$. This implies that w_ε solves

$$w_t = a_\varepsilon(w)(w_{xx} - w + F_\varepsilon) \quad \text{in} \quad K$$

with $a_\varepsilon(\sigma) = ((\sigma + \varepsilon^2)^2 \vee \delta^2) \wedge M_0^2$ and $||F_\varepsilon||_\infty \leq N_0$ for $0 < \varepsilon < \varepsilon_0$, where N_j is a constant depending only on T, $||f||_\infty$, $||f_t||_\infty$, $\min_K f$. Since $a_\varepsilon(\sigma) \geq \delta^2$, from Lemma 6.1 it follows that

$$||w_\varepsilon||_{W_p^{2,1}} \leq C_0||F_\varepsilon||_\infty \leq N_1, \quad 0 < \varepsilon < \varepsilon_0$$

at least for some $p > 2$. Applying the Sobolev inequality yields a bound on the ν-Hölder norm of w_ε for some $0 < \nu < 1$,

$$||w_\varepsilon||_{C^\nu} \leq N_2, \quad 0 < \varepsilon < \varepsilon_0.$$

By the Banach–Alaoglu theorem and the Ascoli–Arzela theorem there is a subsequence (still denoted w_ε) and $u \in W_p^{2,1} \cap C^\nu$ such that

$$w_\varepsilon \to u \quad \text{weakly in} \quad W_p^{2,1}, \quad w_\varepsilon \to u \quad \text{strongly in } C(K).$$

Letting $\varepsilon \to 0$ in (5.4) now yields (1.1) for u since $\varepsilon/\xi(w_\varepsilon + \varepsilon^2) \to 0$ by $w_\varepsilon + \varepsilon^2 \geq \delta$. The higher regularity

$$\begin{aligned} &u \in W_p^{2,1}(K) \quad \text{for} \quad \text{all } p > 1 \quad \text{or} \\ &u \in C^\infty(K) \quad \text{for} \quad f \in C^\infty(K) \end{aligned}$$

follows from the standard linear L^p and Schauder regularity theory [LSU] since the equation (1.1) is uniformly parabolic if $\delta^2 \leq u \leq M^2$.

It remains to prove the constraint (1.2). As in the introduction, if u solves (1.1) and f satisfies (1.3), then

$$-\frac{d}{dt}\int_0^{2\pi} \frac{\sin(x-\zeta)}{u(x,t)}dx = -\int_0^{2\pi} f\sin(x-\zeta)dx = 0, \quad t, \zeta \in \mathbf{R}.$$

Applying Lemma 5.4 yields

$$\int_0^T \int_0^{2\pi} \frac{\sin(x-\zeta)}{u(x,t)}dxdt = 0.$$

We thus conclude that

$$\int_0^{2\pi} \frac{\sin(x-\zeta)}{u(x,t)}dx = 0 \quad \text{for all} \quad t, \zeta \in \mathbf{R}$$

which is the same as (1.2).

7. Unique solvability of a class of quasilinear equations. Our goal in this section is to prove Lemma 6.1. We first prove uniqueness of solution by reflecting Lipschitz continuity of solutions w.r.t. data in L^1. In the proof we use smoothed signature functions which were often used to prove L^1 contractivity for scalar conservation law; see Crandall [Cr]. We next establish a priori bounds for linear parabolic equations by a perturbation method. Such a method is often used in a different context; see Campanato [Ca] and Giga and Yoshida [GYo].

7.1. Uniqueness lemma. *Suppose that $v_i \in C(K_\omega)$ solves (6.1) with $h = h_i \in C(K_\omega)$ with $i = 1, 2$, where $b > 0$. Suppose that $v_i \in W_q^{2,1}(K_\omega)$ for some $1 < q < \infty$ for $i = 1, 2$. Then*

$$b\int_0^T \int_0^\omega |v_1 - v_2|dxdt \leq \int_0^T \int_0^\omega |h_1 - h_2|dxdt.$$

In particular, the solution of (6.1) *is unique in* $C(K_\omega) \cap W_q^{2,1}(K_\omega)$.

Proof. Divide $(6.1)_i$ by $a(v_i)$ and subtract one equation from the other to get

$$A(v_1)_t - A(v_2)_t - w_{xx} + bw + k = 0 \quad \text{in} \quad K = K_\omega \tag{7.1}$$

with $w = v_1 - v_2$, $k = h_2 - h_1$, and $A(s) = \int_0^s a^{-1}(r)dr$. Let $\operatorname{sgn}_\rho(s)$, $\rho > 0$ denote the piecewise linear continuous function such that $\operatorname{sgn}_\rho(s) = 1$ for $s \geq \rho$, $\operatorname{sgn}_\rho(s) = -1$ for $s \leq -\rho$, and sgn_ρ is linear for $|s| \leq \rho$. Multiplying (7.1) by $\operatorname{sgn}_\rho(w)$ and integrating it over K yields

(7.2)
$$0 = \int_0^T \int_0^\omega \{(A(v_1) - A(v_2))_t \operatorname{sgn}_\rho(w) - w_{xx} \operatorname{sgn}_\rho w + bw \operatorname{sgn}_\rho w + k \operatorname{sgn}_\rho w\} dxdt.$$

Note that

$$-\int_0^\omega w_{xx} \operatorname{sgn}_\rho w dx = \int_0^\omega w_x \operatorname{sgn}'_\rho(w) w_x dx \geq 0$$

since $\operatorname{sgn}'_\rho \geq 0$. Sending ρ to 0 in (7.2) yields

$$\int_0^T \int_0^\omega \{(A(v_1) - A(v_2))_t \operatorname{sgn}(w) + b|w|\} dxdt \leq \int_0^T \int_0^\omega |k| dxdt.$$

It remains to prove

$$\int_0^T \int_0^\omega \{(A(v_1) - A(v_2))_t \operatorname{sgn}(w)\} \, dxdt = 0. \tag{7.3}$$

Since $a(r) > 0$ we observe that $\operatorname{sgn} w = \operatorname{sgn}(W)$ with

$$W = A(v_1) - A(v_2).$$

Thus

$$\int_0^T \int_0^\omega W_t \operatorname{sgn} W dxdt = \int_0^T \int_0^\omega |W|_t dxdt = 0,$$

since W is T-periodic in time. This is the same as (7.3). □

We next derive a priori estimates for linear parabolic equations of the form

$$u_t = a(x,t)(u_{xx} - bu + h) \quad \text{in} \quad K_\omega \tag{7.4}$$

with a constant $b > 0$. We use a perturbation method.

7.2. Linear estimate lemma. *Let a_i be a positive constant* $(i = 0, 1)$. *There exist positive constants $\theta_1 = \theta_1(a_i, b, \omega, T)$ and $C_* = C_*(a_i, b, \omega, T)$ such that the following estimate holds. For $a, h \in C(K_\omega)$ with $a_0 \leq a \leq a_1$ there is a unique solution $u \in W_p^{2,1}(K_\omega)$ satisfying*

$$||u||_{W_p^{2,1}} \leq C_* ||h||_\infty, \quad \text{for} \quad 2 \leq p \leq \theta_1 + 2. \tag{7.5}$$

The solution u always satisfies

$$||u||_\infty \leq b^{-1} ||h||_\infty. \tag{7.6}$$

Proof. We may assume $a_1 = 1$ by dilation of the time variable. We may assume that $a, h \in C^\infty(K)$ with $K = K_\omega$. Indeed, for $a, h \in C(K)$ there are $a_\varepsilon, h_\varepsilon \in C^\infty(K)$ such that $a_\varepsilon \to a$, $h_\varepsilon \to h$ in $C(K)$ as $\varepsilon \to 0$ and that $a_0 \le a_\varepsilon \le a_1 = 1$. There is a unique solution u_ε of (7.4) with $a = a_\varepsilon$, $h = h_\varepsilon$. By our results for u_ε we have

$$||u_\varepsilon||_{W_p^{2,1}} \le C_* ||h_\varepsilon||_\infty, \quad ||u_\varepsilon||_\infty \le b^{-1}||h_\varepsilon||_\infty \tag{7.7}$$

for $2 \le p \le 2 + \theta$. We thus observe that

$$u_\varepsilon \to u_0 \text{ weakly in } W_p^{2,1} \text{ and } u_\varepsilon \to u_0 \text{ strongly in } C(K)$$

as $\varepsilon \to 0$ by taking a subsequence if necessary. If $w \in W_2^{2,1}(K)$ solves

$$w_t = a(w_{xx} - bw) \quad \text{in} \quad K$$

then we see $w = 0$. Indeed, multiplying the equation by w_t/a and integrating by parts over K yields $w_t = 0$ so that $w = 0$. Thus, the solution $u \in W_2^{2,1}$ for (7.4) is unique. Since u_0 solves (7.4) it follows that $u = u_0$. Sending $\varepsilon \to 0$ in (7.7) yields (7.5), (7.6).

We first admit that for (smooth) a there is a unique solution $u \in W_2^{2,1}(K)$ of (7.4) for $h \in L^2(K)$. This solution u belongs to $C^\infty(K)$ if h is smooth by the standard regularity theory.

By the maximum principle for $u \in C^\infty(K)$, we see that

$$\min_K h \le bu \le \max_K h.$$

This yields (7.6).

We next consider the L^2 estimate of the linear equation with constant coefficients,

$$u_t = u_{xx} - bu + g \quad \text{on} \quad K. \tag{7.8}$$

Multiplying (7.8) with u_{xx} and integrating by parts over K yields

$$\int_0^T \int_0^\omega (u_{xx}^2 + bu_x^2)dxdt = -\int_0^T \int_0^\omega gu_{xx}dxdt.$$

Thus by the Cauchy inequality we have

$$\iint (u_{xx}^2 + bu_x^2)dxdt \le \frac{1}{2} \iint (u_{xx}^2 + g^2)dxdt,$$

which in particular yields estimates of the $L^2(K)$ norm

$$||u_{xx}||_2 \le 1 \cdot ||g||_2. \tag{7.9}$$

Let L_3 denote the operator norm of $g \mapsto u_{xx}$ from $L^3(K)$ into itself. By L^p theory [LSU] L_3 is finite and depends only on ω, T, and b. Applying Riesz's interpolation between $L^2(K)$ and $L^3(K)$ to (7.9) yields

$$||u_{xx}||_p \le L_p ||g||_p \tag{7.10}$$

with $L_p \to 1$ as $p \to 2$, $p > 2$, where $L_p = L_p(\omega, T, b)$.

We shall use a perturbation argument. Rewrite (7.4) as $u_t = u_{xx} - bu + g$ with $g = (a-1)(u_{xx} - bu + h) + h$ and apply (7.10) to get

$$||u_{xx}||_p \le L_p[||h||_p + (1-a_0)(||u_{xx}||_p + b||u||_p + ||h||_p].$$

We take θ_1 such that $2 \le p \le 2 + \theta_1$ implies

$$c = \sup\{L_p; 2 \le p \le 2+\theta_1\} < \frac{1}{1-a_0}$$

since $L_p \to 1$ as $p \to 2$. We thus have the estimate

$$||u_{xx}||_p \le c_* L_p(2||h||_p + b||u||_p) \quad \text{with} \quad c_* = (1 - c(1-a_0))^{-1}.$$

Since (7.6) yields

$$||u||_p \le S||u||_\infty \le Sb^{-1}||h||_\infty \quad \text{with} \quad S = S(T, \omega, p),$$

we now observe that

$$||u_{xx}||_p \le c_* L_p(2+1)S||h||_\infty \quad \text{for} \quad 2 \le p \le 2+\theta_1.$$

Interpolating (7.6) with this inequality, we conclude that

$$||u||_{W_p^{2,1}} \le C_* ||h||_\infty$$

for $2 \le p \le 2 + \theta_1$ by estimating u_t by the equation (7.4).

It remains to prove the unique existence of solution in $W_2^{2,1}$ of (7.4) for $h \in L^2$. As in the first paragraph of the proof, we have the uniqueness of solution of

$$-\varphi_t = (a\varphi)_{xx} - b(a\varphi) \quad \text{in} \quad K.$$

By the regularity theory only the L^2 solution is zero. This implies that the dual operator L^* of the bounded linear operator $L : u \mapsto h$ (defined by (7.4)) from $W_2^{2,1}$ to L^2 is injective. The closed range theorem now implies that L is surjective provided that the range of L is closed. Using (7.4) in the same way we did to obtain (7.9), we get

$$||u_{xx}||_2 + ||u_x||_2 \le c||Lu||_2.$$

Multiplying u_t by (7.4) and integrating now yields

$$||u_t||_2 \le C'||Lu||_2,$$

since $\int uu_t dt = 0$. Use (7.4) to get

$$||u||_{W_2^{2,1}} \le C''||Lu||_2,$$

which implies the range of L is closed, and the proof is now complete. □

7.3. A priori estimate. We consider

$$v_t = a_T(v)(v_{xx} - bv + h) \quad \text{in} \quad K = K_\omega \tag{7.11}$$

with $a_\tau(p) = \tau a(p) + (1-\tau)a_0$, $0 \leq \tau \leq 1$, where $a(p) \geq a_0$ is given in Lemma 6.1 and $b > 0$. The estimate (7.6) yields

$$||v||_\infty \leq b^{-1}||h||_\infty$$

for a $W_2^{2,1}$ solution v of (7.11) with $v \in C(K)$. We set

$$a_1 = \max\{a(p); |p| \leq b^{-1}||h||_\infty\} < \infty \tag{7.12}$$

to get

$$a_0 \leq a_\tau(v) \leq a_1$$

for a solution $v \in C(K)$ of (7.11). By Lemma 7.2 there are positive constants θ_1 and C_* such that for $2 \leq p \leq 2+\theta_1$ a solution $v \in C(K) \cap W_2^{2,1}(K)$ fulfills

$$||v||_{W_p^{2,1}} \leq C_*||h||_\infty. \tag{7.13}$$

Here θ_1 and C_* depend only on $||h||_\infty$, a_0, b, ω, T.

7.4. A homotopy of mappings. We fix p such that $2 < p < 2+\theta_1$. Since $K = K_\omega$ is two dimensional, there is $0 < \nu < 1$ such that the inclusion $W_p^{2,1}(K)$ into $C^\nu(K)$ is continuous by the Sobolev inequality. We fix ν and set $X = C^\nu(K)$. For $(v,\tau) \in X \times [0,1]$, let $w = R(v,\tau)$ be the unique solution of

$$w_t = a_\tau(v)(w_{xx} - bw + \tau h) \quad \text{in} \quad K. \tag{7.14}$$

7.5. Proposition. *The mapping R gives a compact, continuous mapping from $X \times [0,1]$ into X. The image of R is contained in $\bigcap_{q>1} W_q^{2,1}(K)$. Moreover $R(v,0) = 0$ for all $v \in X$.*

Proof. For $v \in X$, (7.6) implies

$$a_0 \leq a_\tau(v) \leq a_1,$$

where a_1 is defined by (7.12). Suppose that $||v||_X \leq M$. Then $a(v)$ has a modulus of continuity m depending on v only through M. Since m is a modulus of continuity of a_τ for $0 \leq \tau \leq 1$, L^q theory of linear parabolic equation provides

$$||R(v,\tau)||_{W_q^{2,1}} \leq Z||h||_q \quad \text{for} \quad ||v||_X \leq M$$

for $1 < q < \infty$ with $Z = Z(a_0, a_1, q, M)$ where a_1 is as in (7.12). We observe that the image of R is contained in $W_q^{2,1}(K)$ for all q. Take q large so that $W_q^{2,1} \subset X$ is compact. Thus

$$\{R(v,\tau); ||v||_X \leq M, 0 \leq \tau \leq 1\}$$

is compact in X, so $R : X \times [0,1] \to X$ is compact.

From (7.14) it is clear that $R(v,0) = 0$.

Now we must prove the continuity of R. Suppose that $v_j \to v$ in X, $\tau_j \to \tau$ as $j \to \infty$. We apply L^q theory to estimate (for $1 < q < \infty$)

$$||u_j||_{W_q^{2,1}} \leq B, \quad u_j = R(v_j, \tau_j)$$

with $B = B(a_0, a_1, q, ||h||_\infty)$ since $\{v_j\}$ is bounded so that $a_{\tau_j}(v_j)$ has a modulus of continuity independent of j.

For sufficiently large q we observe that

$$(7.15) \qquad u_j \to u \quad \text{weakly in} \quad W_q^{2,1}(K) \quad \text{and} \quad u_j \to u \quad \text{in} \quad C^\nu(K)$$

for some $u \in X \cap W_q^{2,1}(K)$ by taking a subsequence if necessary. Sending $j \to \infty$ in

$$u_{jt} = a_{\tau_j}(v_j)(u_{jxx} - bu_j + \tau_j h)$$

now yields (7.14) with $w = u$. Since (7.14) admits a unique solution for given v, we see $u = R(v, \tau)$. We have thus proved that $R : X \times [0, 1] \to X$ is continuous. □

7.6. Proof of Lemma 6.1. Since ν is chosen so that $W_p^{2,1}(K) \subset C^\nu(K) = X$ is continuous, a priori estimate (7.13) yields

$$||v||_X \leq SC_*||h||_\infty$$

with $S = S(p, \nu, \omega, T)$. By Proposition 7.5 and this estimate the Leray–Schauder fixed point theorem yields $v \in X$ such that $R(v, 1) = v$. By (7.15) we see $v \in \bigcap_{q>1} W_q^{2,1}(K) \subset C(K)$. $R(v, 1) = v$ means v solves (6.1). A priori estimate of Lemma 6.1 follows from (7.13). Since the solution of (6.1) is unique, it remains to prove that $h \mapsto v$ is compact, continuous from $C(K)$ into itself.

The compactness follows from a priori estimates in Lemma 6.1 since K is two dimensional. Suppose that $h_j \to h$ in $C(K)$ and that v_j is a solution of (6.1) with $h = h_j$. By a priori estimates there is a v such that

$$v_j \to v \quad \text{weakly in} \quad W_p^{2,1} \quad \text{and} \quad v_j \to v \quad \text{in} \quad C(K)(j \to \infty)$$

by taking some subsequence if necessary, since K is two dimensional. Letting $j \to \infty$ in (6.1) with $v = v_j$, $h = h_j$ yields (6.1) for v and h. This implies the continuity of $h \mapsto v$.

8. Appendix.

We shall prove that (1.1) with constraint (1.2) admits at most one positive solution if f is time independent. We shall also give a geometric interpretation of the assumption (1.8).

8.1. Proof of Lemma 1.3. Multiplying (1.4) by u_t and integrating by parts over K yields

$$\int_0^T \int_0^{2\pi} \frac{u_t}{u^2}^2 dxdt = \int_0^T \int_0^{2\pi} \left\{ -\frac{1}{2}(u_x^2)_t + \frac{1}{2}(u^2)_t - fu_t \right\} dxdt = 0$$

if f is time independent. This implies $u_t = 0$ on K.

It remains to prove that

$$(8.1) \qquad u_{xx} + u = f(x) \quad \text{on} \quad \mathbf{T} = \mathbf{R}/2\pi\mathbf{Z}$$

admits at most one positive solution satisfying (1.2). Suppose that $u > 0$ solves (8.1) with (1.2). Then all positive solutions U of (8.1) are of the form

$$U(x; \xi, \eta) = u(x) + \xi \cos x + \eta \sin x$$

for some $\xi, \eta \in \mathbf{R}$. It is easy to see that

$$D = \{(\xi, \eta); U(x; \xi, \eta) > 0 \quad \text{for all} \quad x \in \mathbf{T}\} \subset \mathbf{R}^2$$

is open and convex. Since $-\log$ is strictly convex, so is the function

$$E(\xi,\eta)=\int_0^{2\pi}-\log U(x,\xi,\eta)dx \quad \text{on} \quad D.$$

The strict convexity of E on D implies that E has at most one critical point. Since

$$\frac{\partial E}{\partial \xi}=\int_0^{2\pi}-\frac{\cos x}{U}dx, \quad \frac{\partial E}{\partial \eta}=\int_0^{2\pi}-\frac{\sin x}{U}dx,$$

$U(x;\xi,\eta)$ satisfies (8.1) with (1.2) if and only if (ξ,η) is a critical point of E in D. By the choice of u, E takes its critical point at the origin, so there is no other critical point of E on D. In other words, there is no $U(x,\xi,\eta)$ satisfying (1.2) unless (ξ,η) is the origin. We have thus proved the uniqueness of positive solution u of (8.1) fulfilling the constraint (1.2).

8.2. Curvature of Frank diagram. *Let $\mathcal{F}$ be the Frank diagram of $Q>0$ as in §1.3. Let $\kappa(\theta)$ be the inward curvature of $\mathcal{F}$ at $p=(Q(\theta))^{-1}(\cos\theta,\sin\theta)$. Then*

$$\kappa = Q^3(Q+Q'')/(Q'^2+Q^2)^{3/2}. \tag{8.2}$$

In particular, (1.8) *is equivalent to saying that the curvature of $\mathcal{F}$ is positive.*

Proof. For $p\in\mathbf{R}^2$ we define $q(p)$ by

$$p=q(p)(Q(\theta))^{-1}(\cos\theta,\sin\theta),$$

where θ is the argument of p. Thus q is a well-defined positive function (except the origin) and is positively homogeneous of degree one. Moreover, $\mathcal{F}$ is a one-level set of q. We observe that

$$\kappa=\operatorname{div}\left(\frac{\nabla q}{|\nabla q|}\right)=\frac{1}{(q_1^2+q_2^2)^{3/2}}(q_2^2q_{11}-2q_1q_2q_{12}+q_1^2q_{22}), \tag{8.3}$$

where $q_i=\partial q/\partial p_i$, $i=1,2$.

We may assume that $\theta=0$ and $Q(0)=1$ by rotation and dilation of coordinates. Since q is homogeneous of degree one so that q_i is of degree zero, we see

$$q=q_1\cos\theta+q_2\sin\theta, \quad 0=q_{11}\cos\theta+q_{12}\sin\theta, \quad 0=q_{21}\cos\theta+q_{22}\sin\theta, \tag{8.4}$$

where functions are evaluated at $(\cos\theta,\sin\theta)$. From these identities it follows that $q_{21}(p_*)=0$, $q_{11}(p_*)=0$, and $q_1(p_*)=q(p_*)=1$ for $p_*=(1,0)$ since $Q(\theta)=q(\cos\theta,\sin\theta)$. Differentiate Q in θ and use $(8.4)_1$ to get

$$Q'(\theta)=-q_1\sin\theta+q_2\cos\theta,$$
$$Q''(\theta)+Q(\theta)=q_{11}\sin^2\theta-2q_{12}\cos\theta\sin\theta+q_{22}\cos^2\theta,$$

which yields

$$q_2(p_*)=Q'(0) \quad \text{and} \quad q_{22}(p_*)=Q''(0)+Q(0)=Q''(0)+1.$$

Plugging these values into (8.3) we obtain

$$\kappa(0)=\frac{1}{(1+Q'(0)^2)^{3/2}}(Q''(0)+1).$$

This completes the proof. □

REFERENCES

[AL] U. Abresch and J. Langer, *The normalized curve shortening flow and homothetic solutions*, J. Differential Geom., 23 (1986), pp. 175–196.

[AHM] N. D. Alikakos, P. Hess, and H. Matano, *Discrete order preserving semigroups and stability for periodic parabolic differential equations*, J. Differential Equations, 82 (1989), pp. 322–341.

[Am] H. Amann, *Periodic solutions of semilinear parabolic equations*, in Nonlinear Analysis, L. Cesari, R. Kannan, and H. F. Weinberger, eds., Academic Press, New York, 1978, pp. 1–29.

[And] B. Andrews, *Harnack inequalities for evolving hypersurfaces*, Math. Z., 217 (1994), pp. 179–197.

[An] S. Angenent, *On the formation of singularities in the curve shortening flow*, J. Differential Geom., 33 (1991), pp. 601–633.

[AV] S. Angenent and J. J. L. Velázquez, *Asymptotic behavior of singularities in the curve shortening flow*, in preparation.

[Ca] S. Campanato, *L^p regularity for weak solutions of parabolic systems*, Ann. Scuola Norm. Sup. Pisa, 7 (1980), pp. 65–85.

[Ch] Y. Chen, *The weak solutions to the evolution problem of harmonic maps,*, Math. Z., 201 (1989), pp. 69–74.

[CS] Y. Chen and M. Struwe, *Existence and partial regularity results for the heat flow for harmonic maps*, Math. Z., 201 (1989), pp. 83–103.

[Cr] M. G. Crandall, *The semigroup approach to first order quasilinear equations in several variables*, Israel J. Math., 12 (1972), pp. 108–132.

[DL] R. Dal Passo and S. Luckhaus, *A degenerate diffusion problem not in divergence form*, J. Differential Equations, 69 (1987), pp. 1–14.

[EW] C. Epstein and M. Weinstein, *A stable manifold theorem for the curve shortening equations*, Comm. Pure Appl. Math., 40 (1987), pp. 119–139.

[E1] M. Esteban, *On periodic solutions of superlinear parabolic problems*, Trans. Amer. Math. Soc., 293 (1986), pp. 171–189.

[E2] ———, *A remark on the existence of positive periodic solutions of superlinear parabolic problems*, Proc. Amer. Math. Soc., 102 (1988), pp. 131–136.

[FM] A. Friedman and B. Mcleod, *Blow-up of solutions of nonlinear degenerate parabolic equations*, Arch. Rational Mech. Anal., 96 (1986), pp. 55–80.

[Ga1] M. Gage, *Curve shortening makes convex curves circular*, Invent. Math., 76 (1984), pp. 357–364.

[Ga2] ———, *On the size of the blow-up set for a quasilinear parabolic equations*, Contemporary Mathematics, 127 (1992), pp. 41–58.

[GH] M. Gage and R. Hamilton, *The shrinkings of convex plane curves by the heat equation*, J. Differential Geometry, 23 (1986), pp. 69–96.

[Gi] Y. Giga, *A bound for global solutions of semilinear heat equations*, Commun. Math. Phys., 103 (1986), pp. 415–421.

[GM] Y. Giga and N. Mizoguchi, *Existence of periodically evolving convex curves moved by anisotropic curvature*, Proc. International Conference of Advances in Geometric Analysis and Continuum Mechanics, 1993, P. Concus and K. Lancaster, eds., to appear.

[GY] Y. Giga and K. Yama-uchi, *On instability of evolving hypersurfaces*, Differential Integral Equations, 7 (1994), pp. 863–872.

[GYo] Y. Giga and Z. Yoshida, *A dynamic free-boundary problem in plasma physics*, SIAM J. Math. Anal., 21 (1990), pp. 1118–1138.

[Gu] M. Gurtin, *Thermomechanics of Evolving Phase Boundaries in the Plane*, Oxford Press, Oxford, United Kingdom, 1993.

[Ha] R. Hamilton, *The Ricci flow on surfaces*, Contemp. Math., 71 (1988), pp. 237–261.

[He] P. Hess, *On positive solutions of semilinear periodic-parabolic problems in infinite-dimensional systems*, in Lecture Notes in Mathematics, 1076, Springer-Verlag, Berlin, New York, 1984, pp. 101–114.

[HM1] N. Hirano and N. Mizoguchi, *Existence of unstable periodic solutions for semilinear parabolic equations*, Banach Center Publ., to appear.

[HM2] ———, *Positive unstable solutions for semilinear parabolic equations*, Proc. amer. Math. Soc., 123 (1995), pp. 1487–1495.

[KRS] J. Keller, J. Rubinstein, and P. Sternberg, *Reaction-diffusion processes and evolution to harmonic maps*, SIAM J. Appl. Math., 49 (1989), pp. 1722–1733.

[LSU] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'zeva, *Linear and quasilinear equations of parabolic type*, in Transl. Math. Monographs, no. 23, American Mathematial Society, Providence, RI, 1968.

[LY] P. Li and S.-T. Yau, *On the parabolic kernel of the Schrödinger operator*, Acta Math., 156 (1986), pp. 153–201.

[S] J. Shatah, *Weak solutions and development of singularities in the $SU(2)\sigma$-model*, Comm. Pure Appl. Math., 41 (1988), pp. 459–469.

[U] M. Ughi, *A degenerate parabolic equation modelling the spread of an epidemic*, Ann. Mat. Pura Appl., 143 (1986), pp. 385–400.

[W] M. Wiegner, *Blow-up for solutions of some degenerate parabolic equations*, Differential Integral Equations, 7 (1994), pp. 1641–1647.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 40–55, January 1996

003

FULLY NONLINEAR STOCHASTIC PARTIAL DIFFERENTIAL EQUATIONS*

G. DA PRATO† AND L. TUBARO‡

Abstract. The authors study a class of fully nonlinear stochastic partial differential equations by the reduction to a family of deterministic fully nonlinear equations using the stochastic characteristic method.

Key words. stochastic partial differential equations

AMS subject classification. 60H15

1. Introduction.

We are concerned with the stochastic equation

$$(1)\qquad \begin{cases} du(t,\cdot) = L(t,\cdot,u,Du,D^2u)\,dt \;+ \langle b(t,\cdot)Du + h(t,\cdot)u, dW(t)\rangle, \\ u(0) = u_0 \end{cases}$$

where

$$L : [0,+\infty[\times\mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N \times \mathbb{R}^{N^2} \to \mathbb{R},\;\; (t,x,u,p,q) \to L(t,x,u,p,q),$$

$$b : [0,T] \times \mathbb{R}^N \;\to L(\mathbb{R}^N;\mathbb{R}^M),\;\; (t,x) \to b(t,x),$$

$$h : [0,T] \times \mathbb{R}^N \to \mathbb{R}^M,\;\; (t,x) \to h(t,x)$$

are suitable functions, (see Hypothesis 2.1 below), and W is an $\mathbb{R}^M$-valued standard Brownian motion in a given probability space $(\Omega,\mathcal{E},\mathbb{P})$ adapted to a filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\geq 0}$ and such that, for any $t \geq 0$ and any $\varepsilon > 0$, $\mathcal{F}_t$ and $W(t+\varepsilon) - W(t)$ are independent. Moreover, $L(\mathbb{R}^N)$ and $L(\mathbb{R}^N;\mathbb{R}^M)$ denote, respectively, the space of the square matrices or order N and the space of the $M \times N$ rectangular matrices.

Problem (1) has been extensively studied in the semilinear case, where the operator L is simply

$$L(t,x,u,p,q) = \sum_{i,j=1}^{N} a_{ij}(t,x)q_{ij} + a_0(t,x,u,p),$$

and matrix $a(t,x)$ is positive definite; see, for instance, [4] and the references therein.

To our knowledge, in the quasilinear case

$$L(t,x,u,p,q) = \sum_{i,j=1}^{N} a_{ij}(t,x,p)q_{ij} + a_0(t,x,u,p),$$

only a few papers have been devoted to the subject. We recall [1] for quasilinear equations in divergence form, where a "splitting up" method was used, and [3] and [5], where an abstract quasilinear equation was solved, under suitable hypotheses, using a semigroup approach.

* Received by the editors October 8, 1993; accepted for publication (in revised form) May 24, 1994.

† Scuola Normale Superiore di Pisa, Piazza dei Cavalieri, 56126 Pisa, Italy.

‡ Dipartimento di Matematica, Facolta di Scienze Università di Trento, 38050 Povo, Trento, Italy.

In this paper we proceed, as in [15] and [19], by transforming problem (1) in a deterministic fully nonlinear equation for almost all $\omega \in \Omega$ (see §2 below). This equation can be solved in a maximal time interval by using recent results about (deterministic) fully nonlinear equations (see [11] and [13]). One can also show that the solution is global, provided a suitable (sharp) a priori estimate holds. We will give an application in §4.

We shall use the following notation. If φ is a mapping from $\mathbb{R}^N$ into $\mathbb{R}$, we shall denote its gradient by $D\varphi$ and the Jacobian matrix by $D^2\varphi$.

If α is a mapping from $\mathbb{R}^N$ into $\mathbb{R}^N$, we shall denote its first and second derivatives, provided they exist, by α' and α''. We shall denote the trace of a matrix α by Tr α and we denote the vector of components

$$(\text{TR}\ [\alpha'(x)\alpha(x)])_i \quad = \quad \text{Tr}\ [(\alpha'(x)e_i)\alpha(x)]$$

by $\text{TR}[\alpha'(x)\alpha(x)]$, where $e_1, \dots, e_n$ is the canonical basis of $\mathbb{R}^N$.

Moreover we shall denote the Banach space of all real bounded functions in $\mathbb{R}^N$ endowed with the "sup" norm by $C(\mathbb{R}^N)$, the subset of $C(\mathbb{R}^N)$ of all α-Hölder continuous functions by $C^\alpha(\mathbb{R}^N)$, and the subset of $C(\mathbb{R}^N)$ of all functions that are k-times differentiable with continuous and bounded derivatives of order less or equal to k by $C^k(\mathbb{R}^N), k = 1, 2 \dots.$

It is convenient to introduce two special classes of functions.

DEFINITION 1.1. (i) *A mapping*

$$a : [0, +\infty[\times \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N \times \mathbb{R}^{N^2} \to \mathbb{R}, \quad (t, x, u, p, q) \to a(t, x, u, p, q)$$

belongs to the class $(0,0)$ *if it is continuous and, for any* $T > 0$ *and* $r > 0$, *there exists a constant* $N_{T,r}$ *such that*

$$|a(t, x, u, p, q)| \leq N_{T,r}$$

for all $t \in [0, T]$, $x \in \mathbb{R}^N$, *and* u, p, *and* q *with norms not greater than* r.

(ii) *The mapping* a *belongs to the class* (α, β), *with* $\alpha \in]0, 1[$ *and* $\beta \in]0, 1[$, *if for any* $T > 0$ *and* $r > 0$, *there exists a constant* $M_{T,r}$ *such that*

$$|a(t, x, u, p, q) - a(s, x', u', p', q')| \leq M_{T,r}(|t-s|^\alpha + |x-x'|^\beta + |u-u'| + |p-p'| + |q-q'|)$$

for all $t, s \in [0, T]$, $x, x' \in \mathbb{R}^N$, *and* u, u', p, p', q, *and* q' *with norms not greater than* r.

Roughly speaking, a function $a(t, x, u, p, q)$ belongs to the class (α, β), if it is α-Hölder continuous in t, β-Hölder continuous in x, and locally Lipschitz continuous in u, p, and q uniformly in the other variables.

Finally we give the definition of solution of problem (1).

DEFINITION 1.2. *Let* T *be a stopping time with respect to the filtration* $\{\mathcal{F}_t\}_{t \geq 0}$. *A* strong solution *of problem* (1) *in* $[0, T]$ *is a mapping*

$$u : [0, T] \times \mathbb{R}^N \times \Omega \to \mathbb{R}, \quad (t, x, \omega) \to u(t, x),$$

such that the following hold:

(i) $u(t, \cdot)$ *is* $\mathcal{F}_t$*-Bochner measurable, for all* $t \geq 0$; *that is,* $u(t, \cdot)$ *is the a.s. limit of simple random variables with values in* $C^{2+\beta}(\mathbb{R}^N)$.

(ii) *For all $x \in \mathbb{R}^N$, the real stochastic process $u(\cdot, x)$ is such that*

$$L(t, x, u, Du, D^2u) \in L^1([0,T] \times \Omega),$$

$$b(\cdot, x)Du, \ h(\cdot, x)u \in L^2([0,T] \times \Omega; \mathbb{R}^M).$$

(iii) *For any $t \in [0,T]$, it holds that*

$$u(t, \cdot) = u_0 + \int_0^t L(s, \cdot, u, Du, D^2u)\, ds + \int_0^t \langle b(s, \cdot)Du(t, \cdot) + h(s, \cdot)u(t, \cdot), dW(s)\rangle$$

almost surely.

A *strong solution* of problem (1) in a stochastic interval $[0, \tau[$, is defined in the obvious way.

2. Reduction to a deterministic problem. In this section we transform problem (1) into a deterministic one. For the sake of clarity the transformation will be defined in two steps. We assume that the following hypothesis holds.

HYPOTHESIS 2.1. (i) *For some $\alpha, \beta \in]0,1[$, the mapping*

$$L : [0, +\infty[\times\mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N \times \mathbb{R}^{N^2} \to \mathbb{R}, \quad (t, x, u, p, q) \to L(t, x, u, p, q)$$

belongs to the class (α, β) together with its partial derivatives of the type[1] $D_x^h D_u^k D_p^l D_q^m L$, *for $|h| + |k| + |l| + |m| \le 2$.*

(ii) *There exists $\varepsilon > 0$ such that, for any $r > 0$, there is $C_r > 0$ satisfying*

$$|L(0, x, u, p, q) - L(0, y, u, p, q)| \le C_r|x - y|^{\beta+\varepsilon}$$

for all $x, y \in \mathbb{R}^N$ and u, p, and q with norms not greater than r.

(iii) *L belongs to the class $(0, 0)$ together with all its partial derivatives of the type $D_x^h D_u^k D_p^l D_q^m L$, with $|h| + |k| + |l| + |m| = 3$.*

(iv) *b and h are uniformly continuous and bounded in $[0,T] \times \mathbb{R}^N$ together with all their partial derivatives with respect to x, u of order less or equal to 4.*

(v) *All partial derivatives of b and h with respect to x, u of order less or equal to 4 are of class C^1 in time, uniformly in x.*

(vi) *For all $T, r > 0$, there exists $\nu_{T,r} > 0$ such that*

$$\frac{\partial L}{\partial q}(t, x, u, p, q) - \frac{1}{2}b(t, x)b^*(t, x) \ge \nu_{T,r}I,$$

for all $t \in [0,T]$, $x \in \mathbb{R}^N$, and u, p, and q with norms not greater than r.

In the following we set

$$\widetilde{L}(t, x, u, p, q) = L(t, x, u, p, q) - \frac{1}{2}\operatorname{Tr}\,[b(t, x)b^*(t, x)q].$$

Assumption 2.1(vi) implies that the nonlinear operator

$$u \to \widetilde{L}(t, x, u, Du, D^2u)$$

is elliptic.

[1] Here, following usual notation, the indices are multiindices.

We now define

$$y(t,x) = u(t,\xi(t,x)),$$

where ξ is the solution to the system

$$\text{(2)} \qquad \begin{cases} d\xi = -b(t,\xi)\, dW(t) + \text{ TR } [b'(t,\xi)b(t,\xi)]\, dt, \\ \xi(0) = x. \end{cases}$$

In virtue of Hypothesis 2.1(v) there exists a.s. the inverse of $\xi(t,\cdot)$ for all $t \geq 0$ which we denote by $\eta(t,\cdot)$ (see for instance [9]), so that

$$u(t,x) = y(t,\eta(t,x)), \;\; t \geq 0, x \in \mathbb{R}^N.$$

By the Itô–Ventzell formula (see [15], [14], [19]), it follows that y is the solution to the stochastic partial differential equation

$$\begin{aligned} dy(t,x) \;\; = \;\; & \widetilde{L}(t,\xi,u(t,\xi),Du(t,\xi),D^2u(t,\xi))\, dt \\ & - \langle b(t,\xi)Du(t,\xi),h(t,\xi)\rangle\, dt \, - \text{Tr}(Dh(t,\xi)\cdot b(t,\xi))\, u(t,\xi)\, dt \\ & + u(t,\xi)\langle h(t,\xi),dW(t)\rangle. \end{aligned}$$

Now we go to the second step by setting $v(t,x) = \rho(t,x)y(t,x)$, where

$$\rho(t,x) = \exp\left[-\int_0^t \langle h(s,\xi(s,x)),dW(s)\rangle + \frac{1}{2}\int_0^t |h(s,\xi(s,x))|^2 ds \right].$$

By using the fact that ρ is the solution to the stochastic differential equation

$$d\rho = -\rho\langle h,dW(t)\rangle + |h|^2\rho\, dt, \;\; \rho(0) = 1,$$

it is not difficult to check that v fulfills

$$\text{(3)} \qquad \begin{aligned} D_t v(t,x) \;\; = \;\; & \rho(t,x)\, \widetilde{L}(t,\xi(t,x),u(t,\xi(t,x)),Du(t,\xi(t,x)),D^2u(t,\xi(t,x))) \\ & - \rho(t,x)\langle b(t,\xi(t,x))Du(t,\xi(t,x)),h(t,\xi(t,x))\rangle \\ & - \text{Tr}(Dh(t,\xi(t,x))\cdot b(t,\xi(t,x)))\, v(t,x). \end{aligned}$$

In order to get a (deterministic) partial differential equation for v, we now have to express u and its first and second derivatives in terms of v by using the identity

$$\text{(4)} \qquad u(t,x) = \frac{v(t,\eta(t,x))}{\rho(t,\eta(t,x))}.$$

Setting $\zeta = 1/\rho$ and recalling that, by [9], the flow $\eta(t,\cdot), t \geq 0$ is regular, we obtain by a straightforward computation

$$Du(t,x) = (\eta'(t,x))^* D\zeta(t,\eta(t,x))v(t,\eta(t,x)) + (\eta'(t,x))^* Dv(t,\eta(t,x))\zeta(t,\eta(t,x))$$

and

$$\begin{aligned} D^2u(t,x) &= (\eta'(t,x))^*D^2\zeta(t,\eta(t,x))\eta'(t,x)v(t,\eta(t,x)) \\ &\quad + (\eta'(t,x))^*D^2v(t,\eta(t,x))\eta'(t,x)\zeta(t,\eta(t,x)) \\ &\quad + 2(\eta'(t,x))^*D\zeta(t,\eta(t,x)) \otimes \eta'(t,x)Dv(t,\eta(t,x)) \\ &\quad +\langle D\zeta(t,\eta(t,x)), \eta''(t,x)(\cdot,\cdot)\rangle v(t,\eta(t,x)) \\ &\quad +\langle Dv(t,\eta(t,x)), \eta''(t,x)(\cdot,\cdot)\rangle \zeta(t,\eta(t,x)). \end{aligned}$$

It follows that

$$u(t,\xi(t,x)) = \zeta(t,x)v(t,x), \tag{5}$$

$$Du(t,\xi(t,x)) = (\eta'(t,\xi(t,x)))^*D\zeta(t,x)v(t,x) + (\eta'(t,\xi(t,x)))^*Dv(t,x)\zeta(t,x), \tag{6}$$

and

$$\begin{aligned} D^2u(t,\xi(t,x)) &= (\eta'(t,\xi(t,x)))^*D^2\zeta(t,x)\eta'(t,\xi(t,x))v(t,x) \\ &\quad + (\eta'(t,\xi(t,x)))^*D^2v(t,x)\eta'(t,\xi(t,x))\zeta(t,x) \\ &\quad + 2(\eta'(t,\xi(t,x)))^*D\zeta(t,x) \otimes \eta'(t,\xi(t,x))Dv(t,x) \\ &\quad +\langle D\zeta(t,x), \eta''(t,\xi(t,x))(\cdot,\cdot)\rangle v(t,x) \\ &\quad +\langle Dv(t,x), \eta''(t,\xi(t,x))(\cdot,\cdot)\rangle \zeta(t,x). \end{aligned} \tag{7}$$

By substituting (5)–(7) in (3), we finally obtain that v is the solution to the problem

$$\begin{cases} D_tv = \Lambda(t,x,v,Dv,D^2v), \\ v(0) = u_0, \end{cases} \tag{8}$$

where

(9)

$$\begin{aligned} \Lambda(t,x,v,p,q) = \rho\,(t,x)\widetilde{L}\Big(t,\ &\xi(t,x),\ \zeta(t,x)v,\ (\eta'(t,\xi(t,x)))^*D\zeta(t,x)v \\ &+(\eta'(t,\xi(t,x)))^*\zeta(t,x)p, \eta'(t,\xi(t,x)))^*D^2\zeta(t,x)\eta'(t,\xi(t,x))v \\ &+(\eta'(t,\xi(t,x)))^*q\eta'(t,\xi(t,x))\zeta(t,x) \\ &+2\,(\eta'(t,\xi(t,x)))^*D\zeta(t,x)\otimes\eta'(t,\xi(t,x))p \\ &+\langle D\zeta(t,x), \eta''(t,\xi(t,x))(\cdot,\cdot)\rangle v \\ &+\langle p, \eta''(t,\xi(t,x))(\cdot,\cdot)\rangle\zeta(t,x)\,\Big) \\ -\rho(t,x)\langle b(t,\xi(\,t,x))&\Big((\eta'(t,\xi(t,x)))^*D\zeta(t,x)v + (\eta'(t,\xi(t,x)))^*\zeta(t,x)p\Big),\ h(t,\xi(t,x))\rangle \\ -\mathrm{Tr}(Dh(t,\xi(t,x))&\cdot b(t,\xi(t,x)))\,v(t,x). \end{aligned}$$

Now, by using the Itô formula and by proceeding as in [15] and [19], we can prove the following proposition.

PROPOSITION 2.2. *Let $u_0 \in C^{2+\beta}(\mathbb{R}^N)$ and let $\tau \leq T$ be a stopping time with respect to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. If u is a strong solution of* (1) *in* $[0,\tau]$ *then the function $v(\cdot,\cdot) = \rho(\cdot,\cdot)\, u(\cdot,\xi(\cdot,\cdot))$ is a strict solution of* (8).

Conversely, if v belongs to $C([0,\tau];C^{2+\beta}(\mathbb{R}^N)) \cap C^1([0,\tau];C^{\beta}(\mathbb{R}^N))$ a.s. and is a strict solution of (8) *such that $v(t,\cdot)$ is $\mathcal{F}_t$-Bochner measurable for any $t \geq 0$, then u, defined by* (4)*, is a strong solution of* (1).

In order to state the Markov property we need to extend Proposition 2.2 to the case of initial datum given at time $s \geq 0$. This is a straightforward task. Then, with obvious notation, (4) is meant as

$$u(t,x,\omega;s,u_0(\cdot,\omega)) = \frac{v(t,\eta(t,\omega;s,x),\omega;s,u_0(\cdot,\omega))}{\rho(t,\eta(t,\omega;s,x),\omega;s)}. \tag{10}$$

In what follows, we will omit the ω-dependence for brevity.

PROPOSITION 2.3. *Assume that problem* (1) *has a unique strong solution u. Define $\mathcal{M}_s = \mathcal{F}_s \cap \{\tau > s\}$, $s \geq 0$ and let $\Psi\colon C^{2+\beta}(\mathbb{R}^N) \to \mathbb{R}$ be a bounded Bochner-measurable function. Then $u(t,\cdot;0,u_0)$ is $\mathcal{M}_t$-Bochner measurable and*

$$\mathbb{E}\left[\Psi(u(t,\cdot;0,u_0))|\mathcal{M}_s\right] = \mathbb{E}\left[\Psi(u(t,\cdot;s,\varphi))\right]\Big|_{\varphi=u(s,\cdot;0,u_0)}. \tag{11}$$

In formula (11) $u(s,\cdot;0,\varphi)$ is the the strong solution to problem (1) with deterministic initial datum $\varphi \in C^{2+\beta}(\mathbb{R}^N)$.

Proof. We exploit the fact that the solution u to problem (1) can be expressed by mean of the solution v to problem (8); then we can apply known results about deterministic evolution equations available for v, to state corresponding results for u. First we remark that the semigroup property of the process u follows from Proposition 2.2 and the uniqueness of problem (8).

Finally, let us prove the Markov property (11). Arguing as in the proof of Theorem 9.8 in [4], we have

$$\begin{aligned}\mathbb{E}\left[\Psi(u(t,\cdot;0,u_0)|\mathcal{M}_s\right] &= \mathbb{E}\left[\Psi(u(t,\cdot;s,u(s,\cdot;0,u_0)))\Big|\mathcal{M}_s\right]\\ &= \mathbb{E}\left[\Psi(u(t,\cdot;s,\varphi))\Big|\mathcal{M}_s\right]\Bigg|_{\varphi=u(s,\cdot;0,u_0)}\\ &= \mathbb{E}\left[\Psi(u(t,\cdot;s,\varphi))\right]\Big|_{\varphi=u(s,\cdot;0,u_0)}.\end{aligned}$$

Note that the second equality is a consequence of

$$\mathbb{E}\left[\Psi(u(t,\cdot;s,\mathcal{Z}))\Big|\mathcal{M}_s\right] = \mathbb{E}\left[\Psi(u(t,\cdot;s,\varphi))\Big|\mathcal{M}_s\right]\Bigg|_{\varphi=\mathcal{Z}},$$

where $\mathcal{Z}\colon\Omega \to C^{2+\beta}(\mathbb{R}^N)$ is $\mathcal{M}_s$-Bochner measurable. In fact, this can be easily checked when $\mathcal{Z}$ is a simple function. Now let Z be a general random variable and

$\{Z_n\}$ a sequence of simple functions coverging a.s. to Z; then we are allowed to pass to the limit for $n \to \infty$ on the equality

$$\mathbb{E}\left[\Psi(u(t,\cdot;s,\mathcal{Z}_n))\Big|\mathcal{M}_s\right] = \mathbb{E}\left[\Psi(u(t,\cdot;s,\varphi))\Big|\mathcal{M}_s\right]\Bigg|_{\varphi=\mathcal{Z}_n},$$

thanks to continuous dependence on the initial data in equation (8) and recalling that (10) holds. □

3. Local existence and maximal solution. Here we are concerned with problem (8). We start by proving existence for almost any fixed $\omega \in \Omega$ in a maximal interval $[0,\tau(\omega)[$.

PROPOSITION 3.1. *Assume that Hypothesis* 2.1 *holds with* $\alpha \in]0,1/2[$, *and that* $u_0 \in C^{2+\beta+\varepsilon}(\mathbb{R}^N)$ *a.s.; moreover, let* u_0 *be* $\mathcal{F}_0$*-measurable. Then for all* $\omega \in \Omega$, *there exists a maximal interval* $[0,\tau(\omega)[$, *depending on* $u_0(\omega)$, *and a unique regular solution* v *of problem* (8). *Moreover,* $v(t)$ *is* $\mathcal{F}_t$*-Bochner measurable for all* $t \geq 0$ *and* τ *is a stopping time.*

Proof. Consider problem (8) for any fixed $\omega \in \Omega$. To solve it we are going to use Theorem A.2 in Appendix A. We set

$$X = C^\beta(\mathbb{R}^N), \quad D = C^{2+\beta}(\mathbb{R}^N)$$

and consider the mapping $F : [0,T] \times D \to X$, defined by

$$F(t,v) = \Lambda(t,x,v,Dv,D^2v), \ \forall\, t \geq 0. \tag{12}$$

Now we check that Hypothesis A.1 is fulfilled.

Hypothesis A.1(i) follows from Hypothesis 2.1, Proposition B.6 and standard arguments.

Hypothesis A.1(ii), with $\theta = \min\{\alpha, 1/2\}$, follows again from Hypothesis 2.1 and Proposition B.6, recalling that the trajectories of Brownian motion are θ-Hölder continuous for any $\theta < 1/2$.

We now consider Hypothesis A.1(iii). Let $t \geq 0$ and $v_0 \in C^{2+\beta}(\mathbb{R}^N)$. Then we have

$$\begin{aligned}F_v(t,v_0)v = {} & D_v\Lambda(t,\cdot,v_0,Dv_0,D^2v_0)v + D_p\Lambda(t,\cdot,v_0,Dv_0,D^2v_0)Dv \\ & + D_q\Lambda(t,\cdot,v_0,Dv_0,D^2v_0)D^2v.\end{aligned}$$

We claim that the linear operator $F_v(t,v_0)$ is uniformly elliptic. Due to formula (9) and Hypothesis 2.1(vi), it is enough to show that $\|\eta'(t,\xi(t,x))^{-1}\|$ is uniformly bounded in t and x. This follows from Hypothesis 2.1(v) and from the identity

$$\eta'(t,\xi(t,x))^{-1} = \xi'(t,x).$$

Now we notice that, due to the regularity of its coefficients, the realization of the operator $F_v(t,v_0)$ in $C(\mathbb{R}^N)$ is the generator of an analytic semigroup on $C(\mathbb{R}^N)$, thanks to [17]. By interpolation arguments (see, e.g., [12, §3.1]), it follows that $F_v(t,v_0) : D \to X$ generates an analytic semigroup in X and that

$$D_{F_v(t,v_0)}(\alpha,\infty) = C^{\beta+2\alpha}(\mathbb{R}^N), \alpha \in]0,1[, \ \beta + 2\alpha \text{ not integer.}$$

So, A.1(ii) follows. Moreover, thanks to 2.1(i) and (ii), we have

$$F(0,u_0) \in C^{\beta+\varepsilon}(\mathbb{R}^N) = D_{F_v(t,v_0)}(\varepsilon/2,\infty),$$

so that A.1(v) is satisfied with $\alpha = \varepsilon/2$. Finally, A.1(iv) follows from the Schauder's Theorem. So, Hypothesis A.1 is satisfied, with $\theta = \min\{\alpha, 1/2, \varepsilon/2\}$.

Now, fix $T > 0$, then, by Theorem A.2 there is a unique solution of problem (8) in a maximal interval, $[0,\tau(\omega)[$ included in $[0,T]$. Since the proof of Theorem A.2 (see [11]), is based on an iteration procedure which involves $\mathcal{F}_T$-Bochner measurable functions, then the solution v is $\mathcal{F}_T$-Bochner measurable too. Now, fix $T_1 \in [0,T[$, then, by the same argument and by the uniqueness, the maximal solution in $[0,T_1]$ is $\mathcal{F}_{T_1}$-Bochner measurable and defined in $[0,T_1 \wedge \tau(\omega)[$. It follows that $T_1 \wedge \tau$ is $\mathcal{F}_{T_1}$-Bochner measurable and so τ is a stopping time. □

In order to prove global existence (that is $\tau = T$ a.s.) for problem (1), we need global existence for problem (8) a.s. We now show that to accomplish this, it is enough to prove an a priori estimate for the norm $\|v(t,\cdot)\|_{C^{2+\beta+\varepsilon}(\mathbb{R}^N)}$ of the solution.

PROPOSITION 3.2. *Assume that Hypothesis* 2.1 *holds. Let* $u_0 \in C^{2+\beta+\varepsilon}(\mathbb{R}^N)$ *a.s. Let* $v(\cdot,\omega) : [0,\tau(\omega)[$ *be the maximal solution of problem* (8) *given by Proposition* 3.1. *Assume moreover that there exists a constant* $M > 0$ *such that*

$$\|v(t,\cdot)\|_{C^{2+\beta+\varepsilon}(\mathbb{R}^N)} \leq M, \text{ for all } t \in [0,\tau(\omega)[. \tag{13}$$

Then we have $\tau(\omega) = +\infty$.

Proof. We still put $X = C^\beta(\mathbb{R}^N)$ and $D = C^{2+\beta}(\mathbb{R}^N)$ and apply Theorem A.2. Thus, to have a global solution, we have to prove that the mapping

$$v : [0,\tau[\to D,$$

is uniformly continuous. By estimate (13), we have, in fact, where $B(0,T;X)$ denote the space of bounded function with value in the Banach space X,

$$v \in B(0,\tau; C^{2+\beta+\varepsilon}(\mathbb{R}^N)),$$

which implies, since $D_t v = F(t,v)$,

$$D_t v \in L^\infty(0,\tau; C^{\beta+\varepsilon}(\mathbb{R}^N)).$$

By an interpolation result (see [16, Prop. 2.7]), it follows that

$$v \in C^{1-\theta}([0,\tau]; C^{\beta+\epsilon+2\theta}(\mathbb{R}^N)).$$

Choosing $\theta = 1 - \varepsilon/2$, it follows

$$v \in C([0,\tau]; C^{2+\beta}(\mathbb{R}^N)),$$

which yields the conclusion. □

Remark 3.3. Sufficient conditions to get a priori estimates of type (13) for the deterministic problem (8) may be found in [6] and [8]. However, it is not easy to express such conditions in terms of the coefficients of the original problem (1).

Nevertheless, in the next section, we present an application where it is possible to prove global existence.

4. An application. We consider here the special case

$$\text{(14)} \qquad \begin{cases} du = a(t,x,u)u_{xx}dt + (b(t,x)u_x + h(t,x)u)dW(t), \\ u(0) = u_0, \end{cases}$$

where $W(t)$ is a scalar Brownian motion and a, b, and h satisfy the following hypothesis.

HYPOTHESIS 4.1. (i) *For any $T > 0$ the mappings*

$$a : [0,T] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \quad (t,x,u) \to a(t,x,u),$$

$$b : [0,T] \times \mathbb{R} \to \mathbb{R}, \quad (t,x) \to b(t,x),$$

$$h : [0,T] \times \mathbb{R} \to \mathbb{R}, \quad (t,x) \to h(t,x)$$

are of class C^∞ and are bounded together with all their derivatives.

(ii) *There exists $\nu > 0$ such that*

$$a(t,x,u) - \frac{1}{2}b^2(t,x) \geq \nu, \quad \forall\, t \geq 0, \quad x, u \in \mathbb{R}.$$

Under this hypothesis, we can easily see that, setting $N = 1$ and

$$L(t,x,u,p,q) = a(t,x,u)q,$$

hypothesis (2.1) holds.

Now we can prove the following result.

PROPOSITION 4.2. *Assume that Hypothesis* 4.1 *holds, and let $u_0 \in C^3(\mathbb{R})$. Then there exists a unique strong solution of problem* (14).

Proof. We first remark that in this case problem (2) becomes

$$\begin{cases} d\xi = -b(t,\xi)dW(t) + b(t,\xi)\, b_x(t,\xi)dt, \\ \xi(0) = x. \end{cases}$$

Proceeding as in §2, we set

$$u(t,x) = \frac{v(t,\eta(t,x))}{\rho(t,\eta(t,x))},$$

where $\eta(t,x)$ is the inverse of $\xi(t,x)$, and

$$\rho(t,x) = \exp\left\{-\int_0^t h(s,\xi(s,x))dW(s) + \frac{1}{2}\int_0^t h^2(s,\xi(s,x))ds\right\}.$$

Then v is the solution of the problem

$$\text{(15)} \qquad \begin{cases} v_t = R(t,x,v)v_{xx} + S(t,x,v)v_x + T(t,x,v), \\ v(0) = u_0, \end{cases}$$

where, setting

$$\widetilde{a}(t,x,v) = a(t,x,v) - \frac{1}{2}b^2(t,x),$$

we have

$$R(t,x,v) = \widetilde{a}\Big(t,\xi(t,x),\,\frac{1}{\rho(t,x)}v\Big)\ \eta_x^2(t,\xi(t,x)),$$

$$S(t,x,v) = \widetilde{a}\Big(t,\xi(t,x),\,\frac{1}{\rho(t,x)}v\Big)\ \Big(\eta_{xx}(t,\xi(t,x)) - 2\eta_x^2(t,\xi(t,x))\frac{\rho_x(t,x)}{\rho(t,x)}\Big)$$
$$-b(t,\xi(t,x))h(t,\xi(t,x))\eta_x(t,\xi(t,x)),$$

$$T(t,x,v) = \widetilde{a}\Big(t,\xi(t,x),\,\frac{1}{\rho(t,x)}v\Big)\ \Big(\eta_x^2(t,\xi(t,x))\frac{2\rho_x^2(t,x) - \rho(t,x)\rho_{xx}(t,x)}{\rho^2(t,x)}$$
$$-\eta_{xx}(t,\xi(t,x))\frac{\rho_x(t,x)}{\rho(t,x)}\Big)$$
$$-b(t,\xi(t,x))h_x(t,\xi(t,x)) + b(t,\xi(t,x))h(t,\xi(t,x))\eta_x(t,\xi(t,x))\frac{\rho_x(t,x)}{\rho(t,x)}.$$

In view of Proposition 3.1, given $\beta \in\,]0,1[$, problem (15) has a unique adapted solution in a maximal interval $[0,\tau[$. We want to prove that, with a suitable choice of β, there exists $K_1 > 0$ such that

$$\|v(t)\|_{C^{2+\beta}} \le K_1, \quad \forall\, t \in [0,\tau[. \tag{16}$$

We first remark that, by the maximum principle, there exists a positive constant $K(\|v_0\|_\infty)$ such that

$$\|v(t,\cdot)\|_\infty \le K(\|v_0\|_\infty),\ t \in [0,\tau[.$$

It follows that the operator

$$\mathcal{L}v = R(t,x,v)v_{xx} + S(t,x,v)v_x + T(t,x,v)$$

is an elliptic operator with L^∞ coefficients; so, from the well-known Krylov–Safonov theorem (see [7]), v belongs to $C^\beta([0,\tau[;\mathbb{R}^N)$ for some $\beta \in\,]0,1[$. Consequently, $\mathcal{L}$ is an elliptic operator with Hölder-continuous coefficients. Now the required a priori estimate (16) follows from the classical parabolic Schauder theory, which can be found in [8]. □

Appendix A. Fully nonlinear equations. Let D and X be Banach spaces such that

$$D \subset X,$$

with the embedding being continuous but not necessarily dense. We are given a mapping

$$F : [0,+\infty[\times D \to X,\ \ (t,v) \to F(t,v),$$

and we are concerned with the initial value problem

$$\text{(A.17)} \qquad \begin{cases} v'(t) = F(t, v(t)), & t \geq 0, \\ v(0) = v_0 \in D. \end{cases}$$

We shall consider only regular solutions. Let J be an interval in $[0, +\infty[$ such that $\min J = 0$, and fix $\theta \in]0,1[$. By definition, a *solution* of (A.17) on J is a function v such that, for some $\theta \in]0,1[$,

(i) $v \in C^{1+\theta}(J_1; X) \cap C^{\theta}(J_1; D)$, for any closed and bounded subinterval J_1 of J, and

(ii) $v'(t) = F(t, v(t)), t \in J$, and $v(0) = v_0$,

We assume the following hypothesis

HYPOTHESIS A.1. (i) *F is continuous together with its partial derivatives F_v and F_{vv}.*

(ii) *For every $T > 0$ and $v_0 \in D$, $F(\cdot, v_0)$ and $F_v(\cdot, v_0)$ are θ-Hölder continuous in $[0,T]$ locally uniformly with respect to v_0.*

(iii) *For all $(t,v) \in [0, +\infty[\times D$, $F_v(t,v)$ generates an analytic semigroup in X; that is, there are $\rho \in \mathbb{R}$, $\eta \in]\pi/2, \pi[$, and $M > 0$, possibly depending on t and v, such that the resolvent set of $F_v(t,v)$ contains the sector*

$$S_{\eta,\rho} = \{\lambda \in \mathbf{C} : \ \lambda \neq \rho, | \arg(\lambda - \rho) \ | < \eta\}$$

and

$$\|\lambda(\lambda - F_v(t,v))^{-1}\|_{L(X)} \leq M \quad \text{for } \lambda \in S_{\theta,\omega}.$$

(iv) *For all $(t,v) \in [0, +\infty[\times D$, there exists a constant $a > 0$ (possibly depending on t and v), such that*

$$a\|z\|_D \leq \|F_v(t,v) \cdot z\|_X \leq \frac{1}{a}\|z\|_D, \quad \textit{for all } z \in D.$$

(v) *We have*

$$F(0, v_0) \in D_A(\theta, \infty),$$

where $A = F_v(0, v_0)$ and $D_A(\theta, \infty)$ is the real interpolation space defined by

$$D_A(\theta, \infty) = \{z \in X : \sup_{\xi \in \ ]0,1]} \|\xi^{1-\theta} A e^{\xi A} z\|_X < +\infty\}.$$

The following result is proved in [11, Prop. 2.3 and 2.4].

THEOREM A.2. *Assume that Hypothesis* A.1 *holds. Then there exists a unique solution*

$$v(\cdot, v_0)$$

of problem (A.17) *in a maximal interval $[0, \tau_{v_0}[$, with $\tau_{v_0} > 0$.*

If, in addition, $v(\cdot, v_0) : [0, \tau_{v_0}[\to D$ is uniformly continuous, then $\tau_{v_0} = +\infty$.

Appendix B. Composition of Hölder-continuous functions. Let $\beta \in]0,1[$. For any function $u : \mathbb{R}^n \to \mathbb{R}, x \to u(x)$ we set

$$\|u\|_\infty = \sup\{|u(x)| : x \in \mathbb{R}^n\},$$

$$[u]_\beta = \sup\left\{\frac{|u(x)-u(y)|}{|x-y|^\beta} : x,y \in \mathbb{R}^n, x \neq y\right\},$$

$$\|u\|_\beta = \|u\|_\infty + [u]_\beta,$$

$$[u]'_\beta = \sup\left\{\frac{|u(x)-u(y)|}{|x-y|^\beta} : x,y \in \mathbb{R}^n, x \neq y,\ |x-y| \leq 1\right\},$$

$$\|u\|'_\beta = \|u\|_\infty + [u]'_\beta.$$

We remark that, since

$$[u]_\beta \leq [u]'_\beta + 2\|u\|_\infty,$$

the norms $\|\cdot\|_\beta$ and $\|\cdot\|'_\beta$ are equivalent.

If u is k times differentiable, $k \geq 1$, we set

$$\|u\|_{\beta,k} = \|u\|_\infty + \sum_{h=1}^{k} \|D^h\|_\infty + [D^k u]_\beta$$

and

$$C^{k+\beta}(\mathbb{R}^n) = \{u : \|u\|_{\beta,k} < +\infty\}.$$

We are given a mapping

$$\Lambda : [0,+\infty[\times \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N \times \mathbb{R}^{N^2} \to \mathbb{R},\quad (t,x,u,p,q) \to \Lambda(t,x,u,p,q),$$

and we set

$$\text{(B.18)}\quad F(t,u)(x) = \Lambda(t,x,u(x),Du(x),D^2u(x)),\ \forall\, u \in C^{2+\beta}(\mathbb{R}^N), t \in [0,T].$$

We are interested in the regularity properties of F.

LEMMA B.1. *Assume that* Λ *belongs to the class* $(0,0)$ *together with its partial derivatives* $D_{x_i}\Lambda$, $i = 1,\dots,N$, $D_u\Lambda$, $D_{p_i}\Lambda$, $i = 1,\dots,N$, *and* $D_{q_{i,j}}\Lambda$, $i,j = 1,\dots,N$. *Then, for any* $\beta \in]0,1[$, F *belongs to* $C([0,T] \times C^{2+\beta}(\mathbb{R}^N); C^\beta(\mathbb{R}^N))$.

Proof. We have to prove that $\|F(t,u) - F(t_0,u_0)\|_\beta \to 0$ when $t \to t_0$ and $u \to u_0$ in $C^{2+\beta}(\mathbb{R}^n)$. Since clearly

$$\lim_{\substack{t \to t_0 \\ u \to u_0}} \|F(t,u) - F(t_0,u_0)\|_\infty = 0,$$

it suffices to show that

$$\lim_{\substack{t \to t_0 \\ u \to u_0}} [F(t,u) - F(t_0,u_0)]'_\beta = 0.$$

Let $\|x-y\| \le 1$. We have to estimate

$$I = F(t,u)(x) - F(t_0,u_0)(x) - F(t,u)(y) + F(t_0,u_0)(y).$$

To simplify the next formulae, we introduce the notation

$$\lambda = \xi \ \left(t, x, u(x), Du(x), D^2u(x)\right) + (1-\xi) \ \left(t, y, u(y), Du(y), D^2u(y)\right)$$

and

$$\lambda_0 = \xi \ \left(t_0, x, u_0(x), Du(x), D^2u_0(x)\right) + (1-\xi) \ \left(t_0, y, u_0(y), Du_0(y), D^2u_0(y)\right).$$

Then we have

$$\begin{aligned} I = & \int_0^1 [D_x\Lambda(\lambda)(x-y) + D_u\Lambda(\lambda)(u(x)-u(y)) \\ & \quad + D_p\Lambda(\lambda)(Du(x)-Du(y)) + D_q\Lambda(\lambda)(D^2u(x)-D^2u(y))]d\xi \\ & - \int_0^1 [D_x\Lambda(\lambda_0)(x-y) + D_u\Lambda(\lambda_0)(u_0(x)-u_0(y)) \\ & \quad + D_p\Lambda(\lambda_0)(Du_0(x)-Du_0(y)) + D_q\Lambda(\lambda)(D^2u_0(x)-D^2u_0(y))]d\xi \\ = & \int_0^1 [D_x\Lambda(\lambda) - D_x\Lambda(\lambda_0)](x-y)d\xi \\ & + \int_0^1 [D_u\Lambda(\lambda) - D_u\Lambda(\lambda_0)](u_0(x)-u_0(y))]d\xi \\ & + \int_0^1 [D_p\Lambda(\lambda) - D_p\Lambda(\lambda_0)](Du_0(x)-Du_0(y))]d\xi \\ & + \int_0^1 [D_q\Lambda(\lambda) - D_q\Lambda(\lambda_0)](D^2u_0(x)-D^2u_0(y))]d\xi \\ & + \int_0^1 D_u\Lambda(\lambda)(u(x)-u(y)-u_0(x)+u_0(y))d\xi \\ & + \int_0^1 D_p\Lambda(\lambda)(Du(x)-Du(y)-Du_0(x)+Du_0(y))d\xi \\ & + \int_0^1 D_q\Lambda(\lambda)(D^2u(x)-D^2u(y)-D^2u_0(x)+D^2u_0(y))d\xi. \end{aligned}$$

It follows that

$$\begin{aligned} |I| \le & \int_0^1 |D_x\Lambda(\lambda) - D_x\Lambda(\lambda_0)||x-y|^\beta d\xi \\ & + \int_0^1 |D_p\Lambda(\lambda) - D_p\Lambda(\lambda_0)|[u]_\beta|x-y|^\beta d\xi \\ & + \int_0^1 |D_q\Lambda(\lambda) - D_q\Lambda(\lambda_0)|[u]_{\beta,1}|x-y|^\beta d\xi \\ & + \int_0^1 |D_q\Lambda(\lambda) - D_q\Lambda(\lambda_0)|[u]_{\beta,2}|x-y|^\beta d\xi \\ & + \int_0^1 |D_u\Lambda(\lambda_0)|[u-u_0]_\beta|x-y|^\beta d\xi \end{aligned}$$

$$+\int_0^1 |D_p\Lambda(\lambda_0)|[u-u_0]_{\beta,1}|x-y|^\beta d\xi$$

$$+\int_0^1 |D_q\Lambda(\lambda_0)|[u-u_0]_{\beta,2}|x-y|^\beta d\xi,$$

which yields the conclusion. □

The proof of the following lemma is similar and is left to the reader.

LEMMA B.2. *Assume that Λ belongs to the class $(0,0)$ together with its partial derivatives of the type $D_x^h D_u^k D_p^l D_q^m \Lambda$, with $|h|+|k|+|l|+|m| \leq 3$. Then for any $\beta \in]0,1[$, F is a mapping of class C^2 from $[0,+\infty[\times C^{2+\beta}(\mathbb{R}^n)$ into $C^\beta(\mathbb{R}^n)$.*

We now give sufficient conditions in order that $F(t,u)$ is α-Hölder continuous in t.

LEMMA B.3. *Assume that Λ belongs to the class (α,β) together with its partial derivatives $D_{x_i}\Lambda$, $i=1,\dots,N$, $D_u\Lambda$, $D_{p_i}\Lambda$, $i=1,\dots,N$, and $D_{q_{i,j}}\Lambda$, $i,j=1,\dots,N$. Then for all $u \in C^{\beta+2}(\mathbb{R}^n)$ there exists a constant $C(u)$ such that*

$$\|F(t,u)-F(s,u)\|_\beta \leq C(u)|t-s|^\alpha, \text{ for all } t,s \in [0,T]. \tag{B.19}$$

Proof. Let $u \in C^{\beta+2}(\mathbb{R}^n)$ and $r=\|u\|_{2,\beta}$. We first remark that

$$\|F(t,u)-F(s,u)\|_\infty \leq M_{T,r}|t-s|^\alpha, \text{ for all } t,s \in [0,T]. \tag{B.20}$$

Now we want to estimate $[F(t,u)-F(s,u)]'_\beta$. To do this, we have to estimate

$$J=\Lambda(t,x,u(x),Du(x),D^2u(x))-\Lambda(s,x,u(x),Du(x),D^2u(x))$$

$$-\Lambda(t,y,u(y),Du(y),D^2u(y))+\Lambda(s,y,u(y),Du(y),D^2u(y)).$$

We have

$$J=\Lambda(t,x,u(x),Du(x),D^2u(x))-\Lambda(t,y,u(x),Du(x),D^2u(x))$$

$$-\Lambda(s,x,u(x),Du(x),D^2u(x))+\Lambda(s,y,u(x),Du(x),D^2u(x))$$

$$+\Lambda(t,y,u(x),Du(x),D^2u(x))-\Lambda(t,y,u(y),Du(y),D^2u(y))$$

$$-\Lambda(s,y,u(x),Du(x),D^2u(x))+\Lambda(s,y,u(y),Du(y),D^2u(y)),$$

and so, setting

$$\mu_t=(t,\xi x+(1-\xi)y,u(x),Du(x),D^2u(x)),$$

$$\zeta_t=(t,y,\xi u(x)+(1-\xi)u(y),\xi Du(x)+(1-\xi)Du(y),\xi D^2u(x)+(1-\xi)D^2u(y)),$$

we have

$$J=\int_0^1 [D_x\Lambda(\mu_t)-D_x\Lambda(\mu_s)](x-y)d\xi$$

$$+\int_0^1 [D_u\Lambda(\zeta_t)-D_u\Lambda(\zeta_s)](u(x)-u(y))d\xi$$

$$+\int_0^1 [D_p\Lambda(\zeta_t)-D_p\Lambda(\zeta_s)](Du(x)-Du(y))d\xi$$

$$+\int_0^1 [D_q\Lambda(\zeta_t)-D_q\Lambda(\zeta_s)](D^2u(x)-D^2u(y))d\xi.$$

If $|x-y| \leq 1$ and $r = \|u\|_{2,\beta}$, we get

$$|J| \leq M_{T,r}(1+\|u\|_{2,\beta})|t-s|^{\alpha}|x-y|^{\beta},$$

which yields

$$[F(t,u)-F(s,u)]'_{\beta} \leq M_r(1+\|u\|_{\beta,2})|t-s|^{\alpha}. \qquad \square$$

The following result is proved similarly.

LEMMA B.4. *Assume that* Λ *belongs to the class* (α,β) *together with its partial derivatives of the type* $D_x^h D_u^k D_p^l D_q^m \Lambda$, *for* $|h|+|k|+|l|+|m| \leq 2$. *Then, for all* $r>0$ *and for all* $v \in C^{\beta+2}(\mathbb{R}^n)$ *such that* $\|v\|_{2,\beta} \leq r$, *there exists* $M_{T,r}^1$ *such that*

$$\|F_v(t,v)-F_v(s,v)\|_{\beta} \leq M_{T,r}^1|t-s|^{\alpha}, \text{ for all } t,s \in [0,T]. \tag{B.21}$$

We end this section by giving a sufficient condition in order that F fulfills Hypothesis A.1 (i) and (ii). For this, we need the following hypothesis.

HYPOTHESIS B.5. (i) Λ *belongs to the class* (α,β) *together with its partial derivatives of the type* $D_x^h D_u^k D_p^l D_q^m \Lambda$, *for* $|h|+|k|+|l|+|m| \leq 2$.

(ii) Λ *belongs to the class* $(0,0)$ *together with its partial derivatives of the type* $D_x^h D_u^k D_p^l D_q^m \Lambda$, *for* $|h|+|k|+|l|+|m| = 3$.

Then the following proposition holds.

PROPOSITION B.6. *Let* F *be defined by* (B.18) *and assume that Hypothesis* B.5 *holds. Then* F *fulfills Hypothesis* A.1(i) *and* (ii).

REFERENCES

[1] A. BENSOUSSAN, *Some existence results for stochastic partial differential equations*, in Stochastic Partial Differential Equations and Applications, G. Da Prato and L. Tubaro, eds., Pitman Res. Notes Math. Ser., 268 (1992), pp. 37–53.

[2] P. CANNARSA AND V. VESPRI, *Generation of analytic semigroups by elliptic operators with unbounded coefficients*, SIAM J. Math. Anal., 18 (1987), pp. 857–872.

[3] Y. L. DALECKY AND N. Y. GONCHARUK, *On a quasilinear stochastic equation of parabolic type*, Stochastic Anal. Appl., 12 (1994), pp. 103–129.

[4] G. DA PRATO AND J. ZABCZYK, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, 1992.

[5] N. Y. GONCHARUK, *On a class of quasilinear stochastic equation of parabolic type: Regular dependence of solutions on initial data*, preprint.

[6] N. V. KRYLOV, *Nonlinear Elliptic and Parabolic Equations of the Second Order*, Reidel, Dordrecht, 1987.

[7] N. V. KRYLOV AND M. V. SAFONOV, *An estimate of the probability that a diffusion process hits a set of positive measure*, Soviet Math. Dokl., 20 (1979), pp. 253–255.

[8] O. A. LADYZENSKAJA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and quasilinear equations of parabolic type*, Transl. Math. Monographs, 23 (1968).

[9] H. KUNITA, *Stocastic Flows and Stochastic Differential Equations,* Cambridge University Press, 1990.

[10] J. L. LIONS AND J. PEETRE, *Sur une classe d'espaces d'interpolation*, Institut des Hautes Études Scientifiques Publications Mathématiques, 19 (1964), pp. 5–68.

[11] A. LUNARDI, *An introduction to geometric theory of fully nonlinear parabolic equations*, in Qualitative Aspects and Applications of Nonlinear Evolution Equations, T. T. Li and P. de Mottoni, eds., World Scientific, Singapore, 1991, pp. 107–131.

[12] ———, *Interpolation spaces between domains of elliptic operators and spaces of continuous functions with applications to nonlinear parabolic equations,* Math. Nachr., 121 (1985), pp. 295–318.

[13] ———, *Analityc Semigroups and Optimal Regularity in Parabolic Problems*, Birkhäuser, Basel, 1995.

[14] E. Pardoux, *Applications of anticipating stochastic calculus to stochastic differential equations*, in Stochastic Analysis and Related Topics II, H. Korezlioglu and A. S. Ustunel, eds., Lecture Notes in Mathematics No. 1446, Springer-Verlag, Berlin, 1990, pp. 63–105.

[15] B. L. Rozovskii, *Stochastic Evolution Systems: Linear Theory and Applications to Nonlinear Filtering*, Kluwer–Nithoff, Boston, 1990.

[16] E. Sinestrari, *On the abstract Cauchy problem of parabolic type in spaces of continuous functions*, J. Math. Anal. Appl., 107 (1985), pp. 16–66.

[17] H. B. Stewart, *Generation of an analytic semigroup by strongly elliptic operators*, Trans. Amer. Math. Soc., 199 (1974), pp. 141–162.

[18] H. Tanabe, *On the equations of evolution in a Banach space*, Osaka J. Math., 12 (1960), pp. 363–376.

[19] L. Tubaro, *Some results on stochastic partial differential equations by the stochastic characteristic method*, Stochastic Anal. Appl., 62 (1988), pp. 217–230.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 56–69, January 1996

004

CAPILLARY WEDGES REVISITED*

PAUL CONCUS† AND ROBERT FINN‡

Abstract. Equilibrium capillary surfaces in zero gravity in cylindrical containers whose sections are (wedge) domains with corners are studied. Necessary and also sufficient conditions are developed for the existence or nonexistence of surfaces that are locally graphs over the base at the corner, with (prescribed) contact angles that may differ on the two sides. It is shown that the behavior can depart in significant qualitative ways from that which occurs when the two contact angles are the same. Conditions are derived under which such qualitative changes must occur, and illustrative examples are given.

Key words. capillarity, contact angle, free surface, mean curvature, microgravity, wedge domain

AMS subject classifications. 76B45, 53A10, 35B30, 49Q99

1. Introduction.

This paper is devoted to the results announced in our earlier note [1], concerning existence and nonexistence of capillary surfaces over domains with corners, when the data on the two sides of the corner may differ. The behavior of the solutions can differ in significant qualitative ways from that which occurs in the previously considered case of constant data; we are able to a large extent to characterize the conditions under which such qualitative changes must occur.

For background considerations, we refer the reader to our earlier papers [2], [3] and to Chapters 1, 5, and 6 in [4]. In general terms, we consider a cylindrical capillary tube Z with section Ω, closed at one end and partly filled with fluid in the absence of gravity, forming a free surface $\mathcal{S}$. We suppose the boundary Σ of Ω to be piecewise smooth and to have an isolated corner P of opening 2α, $0 < 2\alpha < \pi$, forming a local "wedge domain" at P; see Fig. 1. We seek conditions under which, for prescribed constant (contact) angles γ_1 and γ_2 in the interval $[0, \pi]$, there will exist an $\mathcal{S}$ that can be (locally) represented by a function $z = u(x, y)$ over a neighborhood Ω^* of P in Ω and which meets the sides Z_1 and Z_2, over adjacent segments Σ_1 and Σ_2 of $\partial\Omega$ that abut at P, in the angles γ_1 and γ_2.

Specifically, we seek a solution of

$$\operatorname{div} Tu = 2H \tag{1}$$

in some Ω^*, with

$$Tu = \frac{Du}{\sqrt{1 + |Du|^2}} \tag{2}$$

*Received by the editors April 14, 1993; accepted for publication (in revised form) May 27, 1994. This research was supported in part by the Applied Mathematical Sciences Subprogram of the Office of Energy Research, U.S. Department of Energy, under contract DE-AC03-76SF00098, National Aeronautics and Space Administration grants NAG3-1143 and NCC3-329, and National Science Foundation grant DMS91-06968.

†Lawrence Berkeley Laboratory and Department of Mathematics, University of California at Berkeley, Berkeley, CA 94720 (concus@lbl.gov).

‡Department of Mathematics, Stanford University, Stanford, CA 94305 (finn@cauchy.stanford.edu).

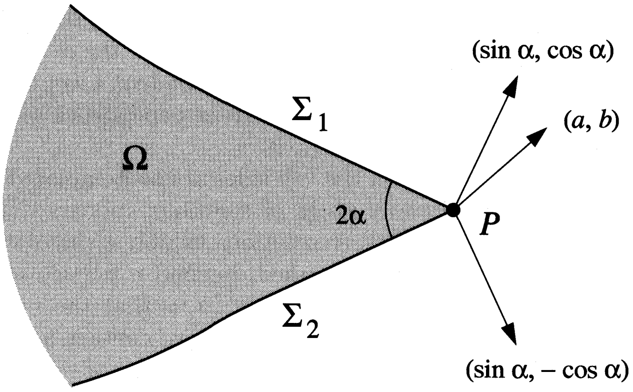


FIG. 1. *The wedge configuration.*

and H an arbitrary prescribed constant, such that

$$(3)\qquad \begin{aligned} \nu \cdot Tu &= \cos\gamma_1 \quad \text{on } \Sigma_1, \\ \nu \cdot Tu &= \cos\gamma_2 \quad \text{on } \Sigma_2, \end{aligned}$$

ν being unit exterior normal vector. Geometrically, H is the *mean curvature* of $\mathcal{S}$; when $H = 0$, $\mathcal{S}$ becomes a minimal surface. In a physical situation, H is determined by the global configuration of Ω and by the boundary conditions over the entire boundary. One sees easily that meaningful physical conditions can give rise to any desired value of H. It is worth noting that if $\gamma_1 = \gamma_2 \neq \pi/2$ over the entire boundary, then $H \neq 0$; if $\gamma_1 = \gamma_2 = \pi/2$ over the entire boundary, then the global problem admits the solution $u \equiv 0$ in Ω, which is unique up to an additive constant.

It is important to observe that in the statement of the problem, $\mathcal{S}$ is not assumed to be defined over P, and no growth conditions are imposed as P is approached from within Ω.

2. Relation to previous work. In earlier work [2], [3], we have shown that if $\gamma_1 = \gamma_2 = \gamma$, then a solution of the local problem (1)–(3) can exist only if $\alpha \geq \left|\frac{\pi}{2} - \gamma\right|$; if $H \neq 0$ and if Σ_1 and Σ_2 are linear segments, then this condition also suffices, while if $H = 0$, then $\alpha > \left|\frac{\pi}{2} - \gamma\right|$ suffices for existence. Again we emphasize that no growth restriction is required at P. Tam [5] showed that whenever a solution exists, then the surface $\mathcal{S}$ is continuous and has a continuous unit normal $\vec{N}$ up to P, see also Simon [6] for earlier work and Miersemann [7] and Lieberman [8] for further developments. This remarkable behavior is the underlying reason that Vreeburg obtained in [9] the identical expression $\alpha \geq \left|\frac{\pi}{2} - \gamma\right|$ as condition for existence of a surface with normal vector continuous to the vertex, without any use of the differential equation.

In the interim, Keller, King, and Merchant [10, §2] studied again the question of a capillary surface $u(x, y)$ defined in a wedge, with (possibly) differing angles γ_1, γ_2 on the two sides. Statements given in that paper conflict basically with a previous result of ours to which the authors refer and of which they assert a simplified proof. The new results of our present study disagree in turn with those announced in [10] for the corresponding general case. Our results are also to some extent at variance with the work by Vreeburg [9] indicated above; these differences are discussed in our paper [1].

It is a curious accident that, when the contact angles on the two wedge sides are equal, all the procedures lead to criteria that look formally similar to those we originally obtained. The criteria do nevertheless differ, even in the equal angle case, in an important respect. This discrepancy presumably lies at the source of the statement in the abstract of [10] that *"the height of the free surface at the corner tends to infinity as the wedge angle decreases to a critical value dependent upon the contact angle."* The statement is in conflict with the discontinuous dependence on the data that is characteristic for the behavior of the solutions.

When differing contact angles on the two sides are contemplated, the distinctions become still more marked. A whole range of solutions appears that is not encompassed in the results stated in [10]; these solutions exhibit a singular behavior at the vertex that has not been previously remarked, beyond a particular case that arose peripherally in [11]. For this reason, it appears to us that the proposal of [10], to use its criterion as a basis for experiments to measure contact angle, would lead to incorrect results in many cases. Examples of the new range of solutions are described in §5 below, and some numerically computed illustrations are given in [12]. It should be of considerable interest to test the results with physical experiments, which could be carried out in a suitable microgravity environment.

A more complete discussion of some of the above material appears in [13].

3. Conditions for existence. We proceed to discuss the problem posed in §1. In the (B_1, B_2) plane, we introduce the closed elliptical domain

$$\mathcal{E}: \ B_1^2 + B_2^2 + 2B_1B_2\cos 2\alpha \le \sin^2 2\alpha \tag{4}$$

inscribed in a square $\mathcal{Q}$ as indicated in Fig. 2. $\mathcal{E}$ cuts off domains $\mathcal{D}_1^+, \mathcal{D}_1^-$ of $\mathcal{Q}$ that are interior to the strip $\mathcal{A}: |B_1 - B_2| < 2\cos^2\alpha$, and domains $\mathcal{D}_2^+, \mathcal{D}_2^-$ of $\mathcal{Q}$ that are exterior to $\mathcal{A}$. Note that the lines $B_1 - B_2 = \pm 2\cos^2\alpha$ pass through the intersection points of $\partial\mathcal{E}$ with $\partial\mathcal{Q}$. We then have

THEOREM 1. *Set $B_1 = \cos\gamma_1$, $B_2 = \cos\gamma_2$. A necessary condition for existence of a solution surface $\mathcal{S}$: $u(x, y)$ of (1)–(3) with unit normal $\vec{N}$ continuous up to P is that the point (B_1, B_2) lie in $\mathcal{E}$; the boundary of $\mathcal{E}$ corresponds to those configurations for which $\mathcal{S}$ is vertical ($\vec{N}$ horizontal) at P. On $\partial\mathcal{E}\cap\partial\mathcal{D}_1^+$, there holds $\gamma_1+\gamma_2 = \pi-2\alpha$; on $\partial\mathcal{E}\cap\partial\mathcal{D}_1^-$, there holds $\gamma_1 + \gamma_2 = \pi + 2\alpha$. On $\partial\mathcal{E}\cap\partial\mathcal{D}_2^-$ and $\partial\mathcal{E}\cap\partial\mathcal{D}_2^+$, there hold, respectively, $\gamma_1 - \gamma_2 = \pi - 2\alpha$ and $\gamma_1 - \gamma_2 = -\pi + 2\alpha$. For existence of such a solution in a domain Ω^* of the type considered and for arbitrary H, it suffices that Σ_1 and Σ_2 be linear segments, and that (B_1, B_2) lie interior to $\mathcal{E}$. If $(B_1, B_2) \in \partial\mathcal{E}\cap\partial\mathcal{D}_1^+\cap\mathcal{A}$, then there is a solution (in some Ω^*) for any $H > 0$; if $(B_1, B_2) \in \partial\mathcal{E}\cap\partial\mathcal{D}_1^-\cap\mathcal{A}$, there is a solution for any $H < 0$.*

Any solution $u(x, y)$, corresponding to interior points of $\mathcal{E}$ or to points of $\partial\mathcal{E}$ interior to $\mathcal{A}$, is continuous and admits a continuous unit normal vector up to P.

Proof. Write $\vec{N} = \langle a, b, c\rangle$ with $c \le 0$, $a^2 + b^2 + c^2 = 1$. Referring to Fig. 1, we find

$$\begin{aligned}\cos\gamma_1 &= \ a\sin\alpha + b\cos\alpha,\\ \cos\gamma_2 &= \ a\sin\alpha - b\cos\alpha,\end{aligned}$$

and the first sentence of the necessary condition follows immediately from the observation that $a^2+b^2 \le 1$, equality holding if and only if $c = 0$. For any $(B_1, B_2) \in \partial\mathcal{E}\cap\partial\mathcal{D}_1^+$, there corresponds a unique (γ_1, γ_2) with γ_1, γ_2 in $[0, \pi]$. We rewrite

$$B_1^2 + B_2^2 + 2B_1B_2\cos 2\alpha = \sin^2 2\alpha \tag{5}$$

in the form

$$(1 - B_1^2)(1 - B_2^2) = (B_1B_2 + \cos 2\alpha)^2. \tag{6}$$

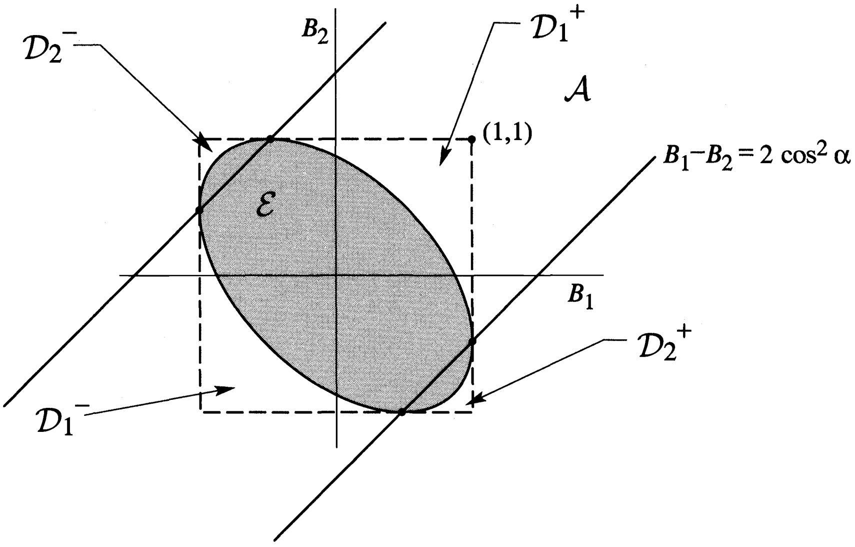


FIG. 2. *Elliptical domain $\mathcal{E}$, reference strip $\mathcal{A}$, square $\mathcal{Q}$, and domains $\mathcal{D}_i^{\pm}$; case $2\alpha < \frac{\pi}{2}$. If $2\alpha > \frac{\pi}{2}$, the directions of major and minor axes interchange.*

Since on the indicated arc we have

$$(7)\qquad (B_1 - B_2)^2 < 4\cos^4\alpha,$$

there follows from (5)

$$(8)\qquad B_1B_2 + \cos 2\alpha > 0.$$

Thus, using positive square roots, we obtain from (6)

$$(9)\qquad \sqrt{1-B_1^2}\sqrt{1-B_2^2} = B_1B_2 + \cos 2\alpha,$$

which is equivalent to

$$(10)\qquad \cos(\gamma_1 + \gamma_2) = \cos(\pi - 2\alpha),$$

so that either $\gamma_1 + \gamma_2 = \pi - 2\alpha$ or $\gamma_1 + \gamma_2 = \pi + 2\alpha$. But from (5), we find at the symmetry point $B_1 = B_2 = B > 0$ on $\partial\mathcal{E} \cap \partial\mathcal{D}_1^+$ that $B = \sin\alpha = \cos(\frac{\pi}{2} - \alpha)$. Thus, the former relation must hold at the symmetry point, and hence it holds throughout the arc. Similarly, on $\partial\mathcal{E} \cap \partial\mathcal{D}_1^-$, there holds $\gamma_1 + \gamma_2 = \pi + 2\alpha$.

On the remaining two arcs $\partial\mathcal{E} \cap \partial\mathcal{D}_2^-$ and $\partial\mathcal{E} \cap \partial\mathcal{D}_2^+$, we obtain by analogous reasoning that $\gamma_1 - \gamma_2 = \pi - 2\alpha$ and $-\pi + 2\alpha$, respectively.

To prove the sufficiency, observe that if (B_1, B_2) is interior to $\mathcal{E}$, then a, b, and c are uniquely determined by the conditions just given, and observe that $c < 0$. The plane Π through P with normal $\vec{N}$ then solves the problem when $H = 0$. If $H > 0$, then a lower hemisphere of radius $1/H$ and tangent to Π at P provides an explicit local solution, while if $H < 0$, then an upper hemisphere yields a solution (see Fig. 3).

If (B_1, B_2) is a boundary point of $\mathcal{E}$, then this procedure always fails when $H = 0$; if $H \neq 0$, then the procedure can under some circumstances yield a solution, provided the trace of Π on the plane of Ω does not enter the (closed) wedge domain. That is,

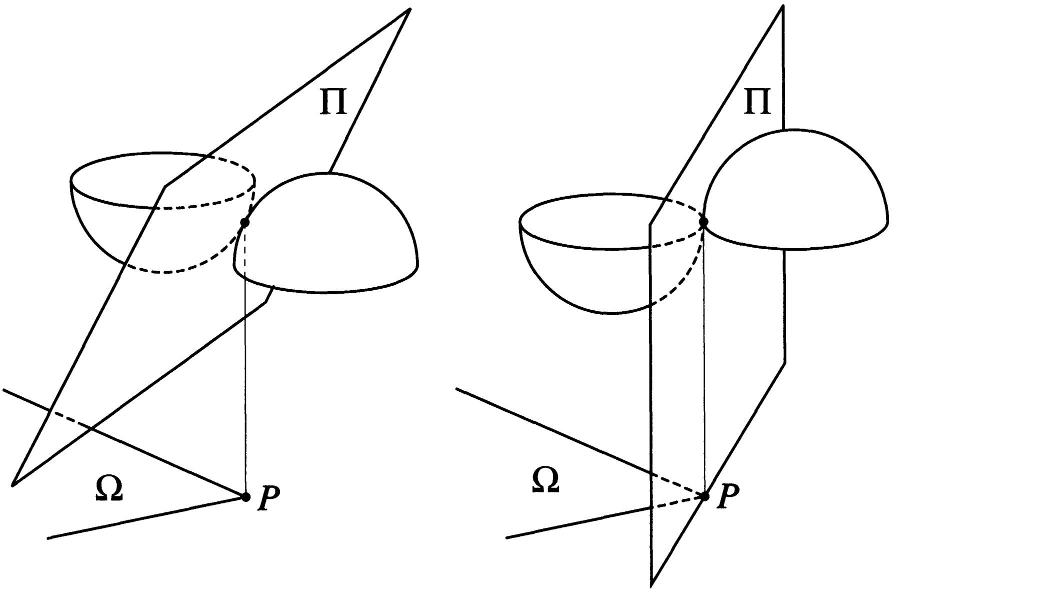


FIG. 3. *Covering of neighborhoods of P by hemispheres; when Π is vertical (corresponding to points on $\partial\mathcal{E}$), only one of the two hemispheres achieving the boundary data at P covers a neighborhood of P in Ω.*

the vector $\langle a, b\rangle$ of Fig. 1 must be a linear combination, with positive coefficients, of the two other vectors in the figure. Since $c = 0$ in this case, the condition becomes $|b| < \cos\alpha$, or equivalently $|B_1 - B_2| < 2\cos^2\alpha$; that is, $(B_1, B_2) \in \mathcal{A}$. But even with that restriction, not all possibilities can be achieved, as the signs of u_x, u_y now reverse for the two hemispheres tangent to Π at P that cover a neighborhood of P in Ω (see Fig. 3), and thus changing the sign of H also reverses the signs of u_x, u_y. Nevertheless, on $\partial\mathcal{E}\cap\partial\mathcal{D}_1^+$ the condition $\gamma_1+\gamma_2 = \pi - 2\alpha$ can be realized by an explicit construction with a lower hemisphere of arbitrary radius; the construction is indicated in Fig. 4.

Similarly, on $\partial\mathcal{E}\cap\partial\mathcal{D}_1^-$, there must hold $\gamma_1 + \gamma_2 = \pi + 2\alpha$, and an explicit construction can be achieved with an upper hemisphere of arbitrary radius.

With regard to the remaining two arcs $\partial\mathcal{E}\cap\partial\mathcal{D}_2^+$ and $\partial\mathcal{E}\cap\partial\mathcal{D}_2^-$, it will be shown below (in §5) that solutions exist, at least at the symmetry points of these arcs; these solutions are, however, not known explicitly.

The final statement of the theorem follows from the method of Tam [5], which applies without essential change to the extended situation considered here. □

It should be emphasized that we have not excluded the possibility of solutions with negative H achieving the data on the segment $\partial\mathcal{E}\cap\partial\mathcal{D}_1^+$ or with positive H achieving the data on the segment $\partial\mathcal{E}\cap\partial\mathcal{D}_1^-$.

In view of the four relations just obtained for γ_1 and γ_2 on $\partial\mathcal{E}$, we see that $\mathcal{E}$ appears as a rectangle in the γ_1, γ_2 coordinates, with sides inclined at 45° to the axes (Fig. 5).

4. Nonexistence. In the above discussion, the requirement that Σ_1 and Σ_2 be linear was introduced solely to facilitate a simple explicit sufficiency proof; it is not essential to the substance of the problem. It is less clear under what conditions solutions with discontinuities at P are excluded, as happens in the equal angle case. To study that point, we attempt to extend the method we introduced for that case to this more general situation. Following in general outline our earlier procedure, we

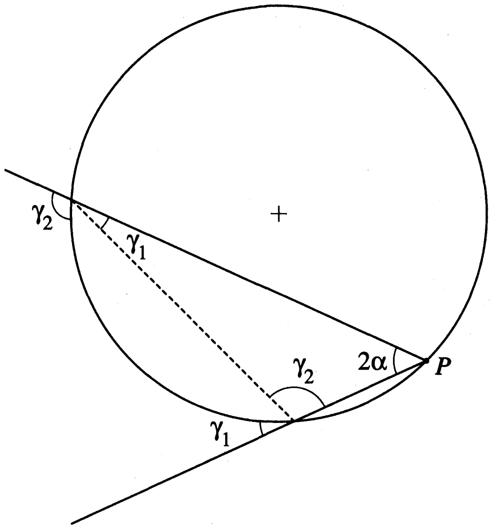


FIG. 4. *Construction of solution as lower hemisphere*; $\gamma_1 + \gamma_2 = \pi - 2\alpha$, $H > 0$.

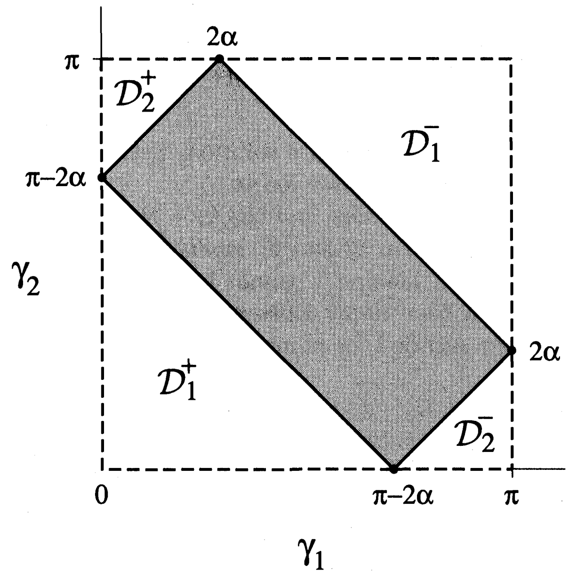


FIG. 5. *Image of $\mathcal{E}$ and of $\mathcal{Q}$ in* (γ_1, γ_2) *coordinates.*

apply Green's identity to (1) in the subdomain $\Omega^{(\Lambda)}$ indicated in Fig. 6, cut off by Γ and Λ (the segment Λ is introduced to exclude possible singularities at the vertex P). We obtain

$$(11) \qquad 2H\left|\Omega^{(\Lambda)}\right| = \left|\Sigma_1^{(\Lambda)}\right| \cos\gamma_1 + \left|\Sigma_2^{(\Lambda)}\right| \cos\gamma_2 + \int_\Gamma \nu \cdot Tu\, ds + \int_\Lambda \nu \cdot Tu\, ds.$$

The crucial observation in what follows is that $|\nu \cdot Tu| < 1$ for any function $u(x, y)$. This inequality permits us initially to move Λ to the vertex P, with the integral over that segment disappearing in the limit. Our next step is to replace $\nu \cdot Tu$ in the other integral by its positive and negative bounds and then to let Γ move to P by parallel translation. Referring to Fig. 6, we choose α_1 and α_2 in $\left(2\alpha - \frac{\pi}{2}, \frac{\pi}{2}\right)$ such that

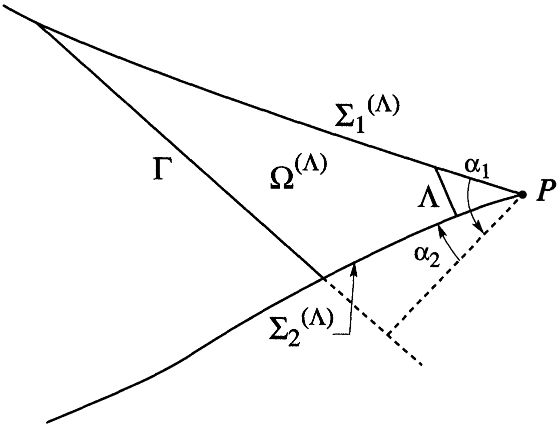


FIG. 6. *Configuration for Theorem 2. Note that* $\alpha_1 > 0$, $\alpha_2 < 0$; *both are in the range* $(2\alpha - \frac{\pi}{2}, \frac{\pi}{2})$.

$\alpha_1 + \alpha_2 = 2\alpha$. Since the area term in (11) tends to zero faster than any of the lengths and since Σ_1 and Σ_2 are asymptotically linear, we are led to the inequality

$$\left|\frac{\cos\gamma_1}{\cos\alpha_1} + \frac{\cos\gamma_2}{\cos\alpha_2}\right| \le \tan\alpha_1 + \tan\alpha_2$$

as a necessary condition for existence of a solution. Setting $A_1 = \cos\alpha_1$, $A_2 = \cos\alpha_2$ and introducing B_1, B_2 as above, we are led to

LEMMA 1. *If* α_1, α_2 *are as above and* $|B_2A_1 + B_1A_2| > \sin 2\alpha$ *then there is no solution to the problem, regardless of growth conditions at* P.

Clearly the conditions of Lemma 1 cannot be satisfied when (B_1, B_2) is interior to $\mathcal{E}$, as Theorem 1 would then imply existence of a solution. We ask whether the conditions are necessarily satisfied for points exterior to $\mathcal{E}$. In formal terms, we have the

QUESTION. *Given* (B_1, B_2) *disjoint from* $\mathcal{E}$, *do there exist* α_1 *and* α_2 *in* $\left(2\alpha - \frac{\pi}{2}, \frac{\pi}{2}\right)$ *such that* $\alpha_1 + \alpha_2 = 2\alpha$ *and* $|B_2A_1 + B_1A_2| > \sin 2\alpha$?

To answer the question, we first prove the following lemma.

LEMMA 2. *For* α_1 *and* α_2 *in* $(-\frac{\pi}{2}, \frac{\pi}{2})$, *the constraint* $\alpha_1 + \alpha_2 = 2\alpha$ *implies the relation*

$$A_1^2 + A_2^2 - 2A_1A_2\cos 2\alpha = \sin^2 2\alpha; \quad A_1, A_2 > 0 \tag{12}$$

describing that portion of an elliptical arc $\mathcal{F}$ *partly inscribed in a unit square in the* (A_1, A_2) *plane, that lies in the first quadrant* (*see Fig.* 7). *Conversely, whenever* (12) *holds there is a unique pair* α_1, α_2 (*up to permutation*) *such that* $\alpha_1 + \alpha_2 = 2\alpha$ *and* α_1 *and* α_2 *are in* $\left(2\alpha - \frac{\pi}{2}, \frac{\pi}{2}\right)$.

Proof. From $\alpha_1 + \alpha_2 = 2\alpha$, we find $A_1A_2 - \cos 2\alpha = \pm\sqrt{1 - A_1^2}\sqrt{1 - A_2^2}$, from which (12) follows on squaring both sides. Conversely, if (12) holds, it can be rewritten in the form just indicated. Choosing $\alpha_1 = \cos^{-1}(A_1)$ and $\alpha_2 = \cos^{-1}(A_2)$ in $[0, \pi/2)$, we find $\cos(\alpha_1 \pm \alpha_2) = \cos 2\alpha$, from which $\alpha_1 \pm \alpha_2 = \pm 2\alpha$. By changing the signs of α_1 or α_2 or both, we can arrange to have $\alpha_1 + \alpha_2 = 2\alpha$, with α_1 and α_2 in $(-\frac{\pi}{2}, \frac{\pi}{2})$. If $\alpha_1 < 2\alpha - \frac{\pi}{2}$ or $\alpha_2 < 2\alpha - \frac{\pi}{2}$, then $\alpha_1 + \alpha_2 < 2\alpha$. This contradiction completes the proof. □

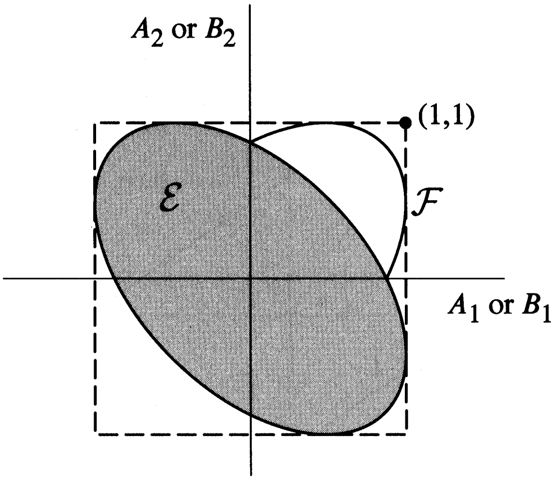


FIG. 7. *Elliptical domain $\mathcal{E}$, elliptical constraint arc $\mathcal{F}$; case $2\alpha < \frac{\pi}{2}$.*

We set

$$A_1 = \frac{x}{\cos\alpha} + \frac{y}{\sin\alpha}, \qquad A_2 = \frac{-x}{\cos\alpha} + \frac{y}{\sin\alpha}, \tag{13}$$

transforming the elliptical arc (12) into a circular arc $\mathcal{C}$ centered at the origin, of radius $R = \sin\alpha\cos\alpha$, and restricted to the upper sector between the lines

$$y = \pm x \tan\alpha \tag{14}$$

(see Fig. 8). The two lines L_1 and L_2 determined by $B_2A_1 + B_1A_2 = \pm\sin 2\alpha$ now become

$$\frac{B_1 - B_2}{\cos\alpha}x + \frac{B_1 + B_2}{\sin\alpha}y = \pm\sin 2\alpha. \tag{15}$$

The inequality $|B_2A_1 + B_1A_2| > \sin 2\alpha$ holds if and only if (x, y) lies outside the strip $\mathcal{W}$ bounded by the lines, and each line has distance

$$d = \frac{\sin^2 2\alpha}{2\sqrt{B_1^2 + B_2^2 + 2B_1B_2\cos 2\alpha}} \tag{16}$$

from the origin. When (B_1, B_2) is exterior to $\mathcal{E}$, we find

$$d < \sin\alpha\cos\alpha = R. \tag{17}$$

Despite the inequality in this last result, it can happen that $\mathcal{C}$ lies strictly interior to $\mathcal{W}$, as $\mathcal{C}$ contains only a portion of the full circle. In such a case, the method yields no information. But it can also occur that interior points of $\mathcal{C}$ lie exterior to $\mathcal{W}$; whenever that happens, any such point of $\mathcal{C}$ yields by Lemma 2 a suitable pair α_1, α_2 and excludes the possibility of any solution to the original problem. We summarize what we have found in the following theorem.

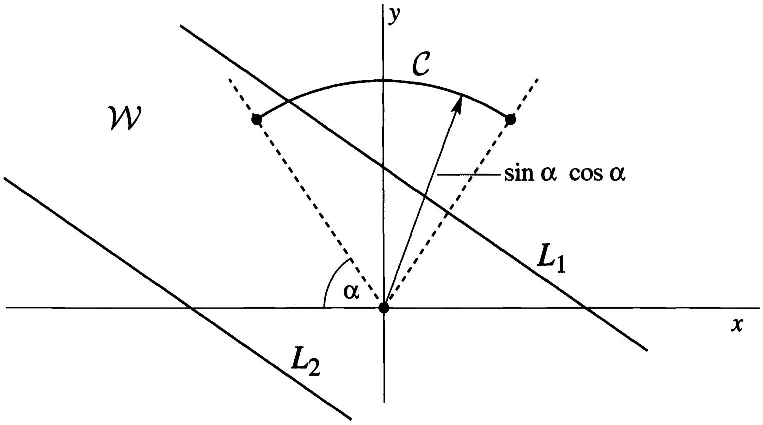


FIG. 8. *Normalized configuration* (*tentative*); *reference lines.*

THEOREM 2. *If C contains points exterior to the closed strip W determined by* (15), *then there is no solution of* (1)–(3) *in any neighborhood Ω^* of P in Ω, for any constant H; this result holds without growth condition at P. If C lies interior to W, then the method provides no information.*

The final statement of the theorem does not reflect a technical failure of the method, but rather arises from actual properties of the solutions. This will be apparent from the second of the following examples.

5. Four examples.

Example 1. Σ_1 and Σ_2 are linear, $\gamma_1 = \gamma_2 = \gamma$ (equivalently, $B_1 = B_2 = B$). Then (15) becomes the pair of horizontal lines $y = \pm(\sin^2\alpha\cos\alpha)/B$. The arc C is independent of B and is indicated in Fig. 9. If (B, B) is exterior to $\mathcal{E}$ then one of the lines crosses C as indicated and points of C will lie exterior to W; hence by Theorem 2 no solution can exist. If (B, B) is in the closure of $\mathcal{E}$ then C lies in the closed strip and Theorem 2 yields no information. However in this case we clearly have $|B_1 - B_2| < 2\cos^2\alpha$ and hence, by Theorem 1, a solution with continuous normal exists. Since (B, B) is exterior to $\mathcal{E}$ if and only if $\alpha < \left|\frac{\pi}{2} - \gamma\right|$ we retrieve exactly our earlier result for the constant angle case, from a more general point of view.

Example 2. Σ_1 and Σ_2 are linear, $\gamma_1 = \pi - \gamma_2 \neq \frac{\pi}{2}(B_1 = -B_2 = B \neq 0)$. Now C is as before, but (15) now yields the two vertical lines $Bx = \pm\sin\alpha\cos^2\alpha$. Since $|B| \leq 1$, C *always* lies interior to W, and thus Theorem 2 yields no information.

If in addition $B > \cos\alpha$, then $(B, -B)$ is exterior to $\mathcal{E}$, and according to Theorem 1 no solution with continuous normal can exist. Nevertheless, a solution to the original problem without growth hypotheses can exist, at least in a significant family of cases. Examples with $B = 1$ and any α are provided by the "moonies", whose existence is proved in [11]. These surfaces have $\gamma_1 = 0$ and $\gamma_2 = \pi$ on adjacent circular arcs of differing radius, see Fig. 10. Theorem 1 provides a new proof independent of the one given in [11], that these surfaces have discontinuous normals at P.

The existence proof in [11] can be modified without essential change to show that if the data $\gamma = 0$ and $\gamma = \pi$ are modified to γ and $\pi - \gamma$, with $0 \leq \gamma \leq \frac{\pi}{2}$, then a solution exists in the identical domain. Thus we obtain a solution of the problem just formulated for any B with $0 \leq B \leq 1$; these solutions have normal vectors discontinuous at P if $B > \cos\alpha$. It can be shown that, if $B < 1$, then the surface is bounded above and below in Ω; if $B = 1$, then $u(x, y) \to -\infty$ for any approach to the

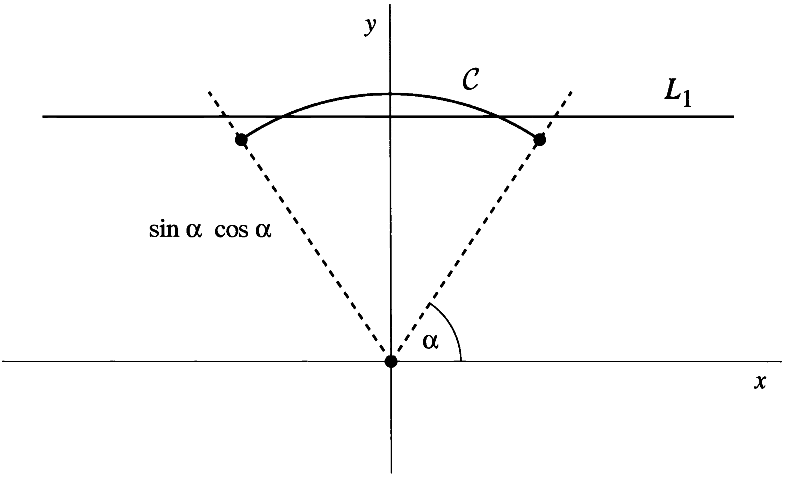


FIG. 9. *Configuration for Example* 1.

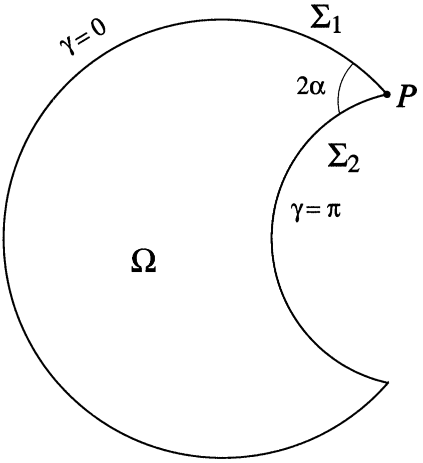


FIG. 10. *Domain for moonie.*

smaller circle, but remains bounded above and below on the larger one.

Thus, in a configuration with differing contact angles, solutions may appear whose behavior at P is very different from that which can occur in the equal angle case. These solutions could not have been obtained by the procedures used for Theorem 1. □

We observe that if $B_1 = -B_2$, then according to Theorem 1 solutions that are smooth up to P can be obtained with successively larger values of $|B|$ tending to unity, as the opening angle 2α closes down to zero. This is despite the discontinuity that occurs when $|B| > \cos\alpha$, and in contrast with the equal angle case, where the admissible B necessarily become small in magnitude with α.

Example 3. Σ_1 and Σ_2 are linear, $B_1 = B$, $B_2 = 0$, $\alpha < \frac{\pi}{4}$. We obtain once more the same $\mathcal{C}$, but (15) yields the sloping lines

$$\frac{B}{\cos\alpha}x + \frac{B}{\sin\alpha}y = \pm\sin 2\alpha.$$

These two lines will enclose $\mathcal{C}$ if and only if $|B| \le \sin 2\alpha$ (see Fig. 11). This is exactly the condition that $(B, 0)$ should be in the closure of $\mathcal{E}$ and also that $|B_1 - B_2| < 2\cos^2\alpha$

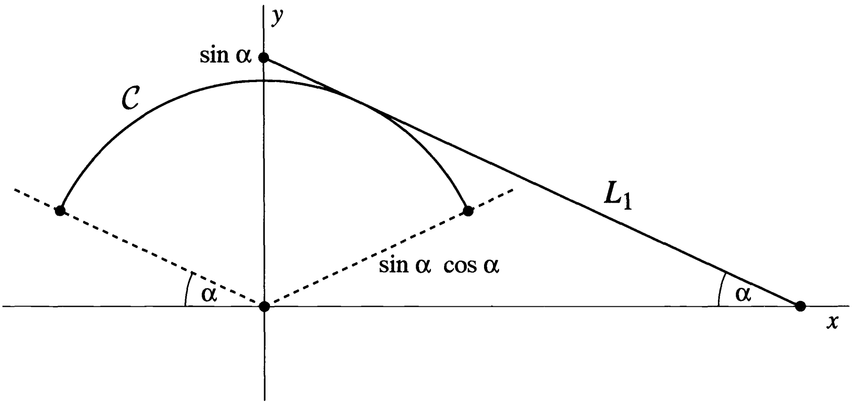


FIG. 11. *Configuration for Example* 3; $|B| = \sin 2\alpha$.

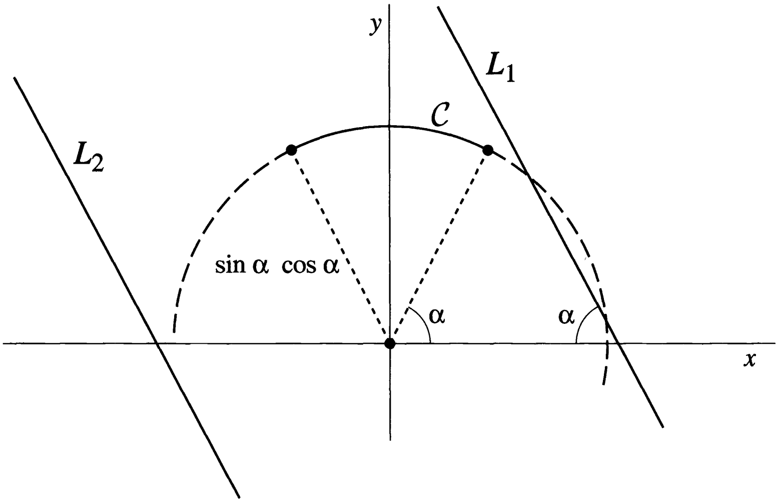


FIG. 12. *Configuration for Example* 4.

(in the given range of α). Thus, by Theorems 1 and 2, there is a solution if and only if $|B| \leq \sin 2\alpha$, and in this case, a solution can be found with continuous normal up to P.

This configuration connects naturally with the equal angle case, as any solution can be reflected across Σ_2 to obtain a solution with equal angles on the edges of a wedge of opening 4α, and conversely, every symmetric solution in such a wedge yields a solution for the given data $(B, 0)$ in the half angle. But the existence criterion for the equal angle problem is $2\alpha \geq \left|\frac{\pi}{2} - \gamma\right|$, which is equivalent to $|B| \leq \sin 2\alpha$. Thus, we obtain once more (as in Example 1) the existence theorem for the equal angle case from a more general point of view.

Example 4. Σ_1 and Σ_2 are linear, $B_1 = B$, $B_2 = 0$, $\frac{\pi}{4} < \alpha < \frac{\pi}{2}$. In this case, even though the lines L_1 and L_2 cut through the completed circle containing $\mathcal{C}$, the arc $\mathcal{C}$ itself lies between the lines (see Fig. 12), and thus Theorem 2 yields no information. If $|B| < \sin 2\alpha$, then Theorem 1 still guarantees the existence of a solution smooth to P. But in this case, $\sin 2\alpha > 2\cos^2 \alpha$, and hence we can no longer be assured of existence of solutions corresponding to data points on the boundary of $\mathcal{E}$.

If $B > \sin 2\alpha$, then Theorem 2 yields no information while Theorem 1 guarantees that no smooth solution can exist. As in Example 3, any solution would yield a solution of the constant angle problem in a wedge of opening 4α. For such a case, for which $4\alpha > \pi$, Korevaar [14] has shown the existence in a gravity field of solutions that are bounded but discontinuous at the vertex. Computations we have made [15] indicate that jump discontinuities can occur in the absence of gravity for values $\left|\frac{\pi}{2} - \gamma\right|$ that exceed a critical one, depending on α and on the global geometry. In a forthcoming work, Lancaster and Siegel [16] characterize the local qualitative behavior of the solutions at such points. □

It should be observed that not only can discontinuities appear at reentrant corners with constant data as above; discontinuities occasioned by abruptly changing data can occur for smooth boundaries, at least in the presence of gravity fields; an example is given on p. 15 of [17].

6. Other properties. We note that in all the above examples, the endpoints of $\mathcal{C}$ lie interior to the closed strip determined by L_1 and L_2. This behavior is in fact quite general.

LEMMA 3. *Let L_1 and L_2 be determined, respectively, by the plus and minus signs in* (15). *Then the right-hand endpoint of $\mathcal{C}$ lies on L_1 if and only if $B_1 = 1$; it lies on L_2 if and only if $B_1 = -1$. The left-hand endpoint of $\mathcal{C}$ lies on L_1 if and only if $B_2 = 1$; it lies on L_2 if and only if $B_2 = -1$. Both endpoints lie always interior to the closed strip $\mathcal{W}$.*

Proof. Setting

$$F(x,y) = \frac{B_1 - B_2}{\cos\alpha}x + \frac{B_1 + B_2}{\sin\alpha}y - \sin 2\alpha,$$

$$G(x,y) = \frac{B_1 - B_2}{\cos\alpha}x + \frac{B_1 + B_2}{\sin\alpha}y + \sin 2\alpha,$$

the lines L_1 and L_2 are characterized, respectively, by $F(x,y) = 0$ and by $G(x,y) = 0$. Choosing for x, y the coordinates of the right-hand endpoint of $\mathcal{C}$, we find $F(x,y) = (B_1 - 1)\sin 2\alpha$, $G(x,y) = (B_1 + 1)\sin 2\alpha$. This proves the assertions relating to B_1; those relating to B_2 are proved analogously. The same relations show that if $|B_1| < 1$, then the right-hand endpoint of $\mathcal{C}$ lies strictly between the two lines; similarly, the left-hand endpoint lies between the lines when $|B_2| < 1$. □

As a consequence of Lemma 3, we see that *the configuration indicated in* Fig. 8, *which was drawn to be indicative of a general situation, cannot occur as shown, as one of the endpoints of $\mathcal{C}$ lies exterior to the strip in that configuration.*

Referring to Fig. 2, we introduce $\mathcal{D}_1^+, \mathcal{D}_1^-, \mathcal{D}_2^+$, and $\mathcal{D}_2^-$ as in that figure. We adjoin to these domains all the boundary points that lie on the boundary of the square. On the line segment $B_1 = B_2 = B > 0$, L_1 takes the form $y = (\sin^2\alpha\cos\alpha)/B$, a horizontal line that is tangent to $\mathcal{C}$ at its midpoint when (B, B) is on the boundary of $\mathcal{E}$, and cuts through $\mathcal{C}$ when (B, B) lies exterior to $\mathcal{E}$, as in Fig. 9. Thus, according to Theorem 2, the wedge problem admits no solution corresponding to the segment $1 \geq B > \sin\alpha$ (cf., Example 1). We now allow (B_1, B_2) to move along the arc of $\partial\mathcal{E}$ between the two nearest contact points with the square. According to (16), the distance of L_1 to the origin remains unchanged, and thus we obtain a family of lines tangent to the circle on which $\mathcal{C}$ lies. Since $|B_1| < 1$ and $|B_2| < 1$ interior to the arc of $\partial\mathcal{E}$ considered, we find by Lemma 3 that all corresponding contact points with the circle actually lie interior to $\mathcal{C}$. Again by Lemma 3, as the points on $\partial\mathcal{E}$ move to the contact points with the square, L_1 becomes tangent to $\mathcal{C}$ at the respective endpoints; thus, all of $\mathcal{C}$ is covered. It is easy to see that the covering is 1−1.

Through each of the considered points of $\partial\mathcal{E}$, we construct the extended line segment from the origin. Repeating the reasoning given above for the $B_1 = B_2 > 0$ configuration, we find that, for all points of that line segment exterior to $\mathcal{E}$, there exists no solution to the wedge problem. Since these lines sweep out $\mathcal{D}_1^+$, there can be no solution for any point of $\mathcal{D}_1^+$. An identical reasoning excludes solutions for any point of $\mathcal{D}_1^-$.

We now consider the two remaining complementary domains $\mathcal{D}_2^+$ and $\mathcal{D}_2^-$, which have $(1,-1)$ and $(-1,1)$ as boundary points. In Example 2 above, we have shown the existence of solutions for every point on the line segment $-1 \le B_1 = -B_2 \le 1$; thus $\mathcal{C}$ lies between L_1 and L_2 for all points on that segment (see Fig. 12). These solution surfaces have discontinuous normals at P for all points exterior to $\mathcal{E}$. Essentially, a repetition of the reasoning directly above shows that $\mathcal{C}$ lies between L_1 and L_2 for all points of $\mathcal{D}_2^+$ and of $\mathcal{D}_2^-$. We have proved the following theorem.

THEOREM 3. *For any points (B_1, B_2) in the domains $\mathcal{D}_1^+$ and $\mathcal{D}_1^-$ defined above, there are points of $\mathcal{C}$ exterior to the strip $\mathcal{W}$, and hence there exists no solution to the wedge problem* (1)–(3) *in any neighborhood of the vertex P, for any constant H. In $\mathcal{D}_2^+$ and $\mathcal{D}_2^-$, $\mathcal{C}$ lies interior to $\mathcal{W}$; solutions do exist, at least on the symmetry line $B_1 = -B_2$ of those domains. For all points of that line exterior to $\mathcal{E}$, the unit normals to the solution surfaces are discontinuous at P.*

For all interior points of $\mathcal{E}$ and any H, solutions exist and all such solutions are smooth up to P. We conjecture that for any α in the range $0 < \alpha < \frac{\pi}{2}$ considered, there exist wedge domains for which solutions (with discontinuous normal) exist for all (B_1, B_2) lying in $\mathcal{D}_2^+$ or in $\mathcal{D}_2^-$.

We close with the following theorem.

THEOREM 4. *Whenever a bounded solution exists in a wedge domain, then every solution is bounded.*

This is an immediate formal consequence of the general comparison principle for capillary surfaces; see [17, §2] or [4, Chap. 5]. Unbounded solutions can occur, as in the "moonie" example above. In such a case, every solution is unbounded.

Acknowledgment. We wish to thank Erich Miersemann for helpful discussions.

REFERENCES

[1] P. CONCUS AND R. FINN, *On a comment by J. P. B. Vreeburg*, Microgravity Science and Technology, IV (1991), p. 60.

[2] ———, *On capillary-free surfaces in the absence of gravity*, Acta Math., 132 (1974), pp. 177–198.

[3] ———, *On the behavior of a capillary-free surface in a wedge*, Proc. Nat. Acad. Sci. U.S.A., 63 (1969), pp. 292–299.

[4] R. FINN, *Equilibrium Capillary Surfaces*, Springer-Verlag, New York, 1986.

[5] L.-F. TAM, *Regularity of capillary surfaces over domains with corners: Borderline case*, Pacific J. Math., 124 (1986), pp. 469–482.

[6] L. SIMON, *Regularity of capillary surfaces over domains with corners*, Pacific J. Math., 88 (1980), pp. 363–377.

[7] E. MIERSEMANN, *On capillary free surfaces without gravity*, Z. Anal. Anwendungen, 4 (1985), pp. 429–436.

[8] G. LIEBERMAN, *Hölder continuity of the gradient at a corner for the capillary problem and related results*, Pacific J. Math., 133 (1988), pp. 115–135.

[9] J. P. G. VREEBURG, *Comment on the paper "Parabolic flight experiments on fluid surfaces and wetting,"* Microgravity Science and Technology, III (1990), p. 125.

[10] J. B. KELLER, A. KING, AND G. MERCHANT, *Surface tension,* in Proc. Symposium for John

Miles, Scripps Institute Oceanography, University of California at San Diego, 1991, pp. 161–168.

[11] R. Finn, *Moon surfaces and boundary behavior of capillary surfaces for perfect wetting and nonwetting*, Proc. London Math. Soc., 57 (1988), pp. 542–576.

[12] P. Concus and R. Finn, *Capillary surfaces in a wedge; Differing contact angles*, Microgravity Science and Technology, VII (1994), pp. 152–155. Errata, Microgravity Science and Technology, VII (1994), p. 218.

[13] ———, *Comments on a paper of Keller, King, and Merchant*, Nonlinear Anal., to appear.

[14] N. J. Korevaar, *On the behavior of a capillary surface at a reentrant corner*, Pacific J. Math., 88 (1980), pp. 379–385.

[15] P. Concus, R. Finn, and F. Zabihi, *On canonical cylinder sections for accurate determination of contact angle in microgravity*, in Fluid Mechanics Phenomena in Microgravity, D. A. Siginer and M. M. Weislogel, eds., Applied Mechanics Division, vol. 154, American Society of Mechanical Engineers, New York, 1992, pp. 125–131.

[16] K. Lancaster and D. Siegel, *Radial limits of capillary surfaces*, Pacific J. Math., to appear.

[17] R. Finn, *Comparison principles in capillarity*, in Partial Differential Equations and Calculus of Variations, Springer Lecture Notes in Mathematics 1357, Springer-Verlag, New York, 1988, pp. 156–197.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 70–77, January 1996

GLOBAL EXISTENCE OF SOLUTIONS FOR THE SYSTEM OF COMPRESSIBLE ADIABATIC FLOW THROUGH POROUS MEDIA*

L. HSIAO[†] AND D. SERRE[‡]

Abstract. Consider the quasilinear hyperbolic system

$$(1)\qquad \begin{cases} v_t - u_x = 0, \\ u_t + p(v,s)_x = -\alpha u, \quad \alpha > 0, \quad p_v < 0 \text{ for } v > 0, \\ s_t = 0 \end{cases}$$

with initial data

$$(2)\qquad u(x,0) = u_0(x), \quad v(x,0) = \bar{v} + v_o(x), \quad s(x,0) = \bar{s} + s_0(x),$$

where $\bar{v} > 0, \bar{v}$ and $\bar{s}$ are constants, $(u_0(x), v_0(x)) \in C^1$ with a compact support, and $s_0(x) \in C^2$ with a compact support.

It is proved in this paper that there exists a globally defined classical solution for the Cauchy problem if the C^1-norm of $(u_0(x), v_0(x))$ and the C^2-norm of $s_0(x)$ are small.

Key words. global existence, the system of compressible adiabatic flow, damping dissipation, L_2-estimates, L_∞-esitmates.

AMS subject classifications. 35, 76

1. Introduction.

Consider the system

$$(1.1)\qquad \begin{cases} v_t - u_x = 0, \\ u_t + p(v,s)_x = -\alpha u, \\ s_t = 0, \end{cases}$$

which can be used to model the adiabatic gas flow through a porous medium, where v denotes the specific volume, u denotes the velocity, s denotes the entropy, and p denotes the pressure, with $p_v < 0$ for $v > 0$. This system is strictly hyperbolic, with eigenvalues $\lambda_1 = -\sqrt{-p_v}$, $\lambda_2 \equiv 0$, and $\lambda_3 = \sqrt{-p_v}$.

For the case $\alpha = 0$, namely

$$(1.2)\qquad \begin{cases} v_t - u_x = 0, \\ u_t + p(v,s)_x = 0, \\ s_t = 0, \end{cases}$$

there is no globally defined classical solution, in general. In fact, consider the initial data

$$(1.3)\qquad u(x,0) = \varepsilon u_0(x), \quad v(x,0) = \overline{v} + \varepsilon v_0(x), \quad s(x,0) = \varepsilon s_0(x)$$

such that $\overline{v}$ is a positive constant, and $(u_0(x), v_0(x), s_0(x)) \in C^1$ with a compact support. It is shown in [LI] and [LZ] that there exists a small $\varepsilon_0 > 0$ such that for any $\varepsilon : 0 < \varepsilon \leq \varepsilon_0$, the C^1 solution of this Cauchy problem blows up at a finite time and the life span is $\tilde{T}(\varepsilon) = O(\varepsilon^{-1})$.

* Received by the editors April 6, 1994; accepted for publication April 26, 1994.

[†] Institute of Mathematics, Academia Sinica, Beijing 100080, People's Republic of China. The research of this author was partially supported by the National Natural Science Foundation of China.

[‡] Centre National de la Recherche Scientifique 46, Ecole Normale Supérieure de Lyon, Allée d'Italie, 69364 Lyon cedex 07, France.

For system (1.1) with a damping term, however, the situation is different. Consider any initial data

$$u(x,0)=u_0(x), v(x,0)=\overline{v}+v_0(x), s(x,0)=\overline{s}+s_0(x), \tag{1.4}$$

where $\overline{v}>0, \overline{v}$ and $\overline{s}$ are constants, $(u_0(x),v_0(x))\in C^1$ with a compact support, and $s_0(x)\in C^2$ with a compact support. We prove that there exists a globally defined classical solution for the Cauchy problem if the C^1-norm of $(u_0(x),v_0(x))$ and the C^2-norm of $s_0(x)$ are small. This means that the damping dissipation is strong enough to preserve the smoothness of the initial data when it is small.

Another concern is the influence of the damping mechanism on the large-time behavior of solutions. For the case of isentropic flow, namely $s(x,t)\equiv$ constant, it has been proven [HL] that the solution of the Cauchy problem

$$\begin{cases} v_t-u_x=0, \\ u_t+p(v)_x=-\alpha u, \quad \alpha>0, \quad p'(v)<0 \quad \text{for } v>0, \\ v(x,0)=v_0(x), \qquad u(x,0)=u_0(x) \text{ with } \lim\limits_{x\to\mp\infty}(v_0(x),u_0(x))=(v^{\mp},u^{\mp}) \end{cases} \tag{1.5}$$

can be described by the solution of the problem

$$\begin{cases} v_t=-\frac{1}{\alpha}p(v)_{xx}, \\ u=-\frac{1}{\alpha}p(v)_x, \\ v(x,0)=\tilde{v}_0(x) \quad \text{with} \quad \lim\limits_{x\to\mp\infty}\tilde{v}_0(x)=v^{\mp} \end{cases} \tag{1.6}$$

time asymptotically. The system in (1.6) is obtained from (1.5) by approximating the momentum equation in $(1.5)_2$ with Darcy's law. Moreover, the L_2-norm and L_∞-norm of the difference between these two solutions tend to zero with a rate $t^{-\frac{1}{2}}$ as time t tends to infinity [HL]. This shows that certain nonlinear diffusive phenomena occur for the solution of (1.5), which is caused by the damping mechanism.

For system (1.1), the corresponding simplified system takes the form

$$\begin{cases} v_t=-\frac{1}{\alpha}p(v,s)_{xx}, \\ u=-\frac{1}{\alpha}p(v,s)_x, \\ s=s(x)=\overline{s}+s_0(x). \end{cases} \tag{1.7}$$

Consider the initial data

$$v(x,0)=\overline{v}+\tilde{v}_0(x), \qquad \tilde{v}_0\in C^1 \text{ with a compact support.} \tag{1.8}$$

It is expected that the large-time behavior of the solution for (1.1) and (1.4) can be described by the solution of (1.7) and (1.8). This is discussed in [HS].

The global existence result in this paper is established by the L_2-estimates for the solutions and L_∞-estimates on the first derivatives, which are given in the second section. The third section is devoted to the global existence theorem.

2. The a priori estimates. Denote $c=(-p_v(v,s))^{\frac{1}{2}}$. Let

$$\begin{cases} w=u-h(v,s), \\ z=u+h(v,s), \end{cases} \tag{2.1}$$

where $h(v,s)$ satisfies the equation

$$ch_v-p_v=0. \tag{2.2}$$

Then system (1.1) can be written as

$$(2.3)\qquad \begin{cases} w_t - cw_x = -\frac{\alpha}{2}(z+w) + (ch_s - p_s)s_x, \\ z_t + cz_x = -\frac{\alpha}{2}(z+w) + (ch_s - p_s)s_x, \\ s_t = 0, \end{cases}$$

which is equivalent to (1.1) for classical solutions.

For the system of gas dynamics,

$$(2.4)\qquad p(v,s) = (\gamma - 1)v^{-\gamma}e^s (1 < \gamma < 3),$$

and it is easy to find that

$$h(v,s) = 2\sqrt{\frac{\gamma}{\gamma-1}}v^{\frac{1-\gamma}{2}} \cdot e^{\frac{s}{2}}$$

and

$$ch_s - p_s = v^{-\gamma}e^s.$$

Thus, system (2.3) takes the form

$$(2.5)\qquad \begin{cases} D^- w = -\frac{\alpha}{2}(z+w) + f(z-w,s,s_x), \\ D^+ z = -\frac{\alpha}{2}(z+w) + f(z-w,s,s_x), \\ s(t,x) = s(0,x), \end{cases}$$

where

$$D^- = \partial_t - c\partial_x, \quad D^+ = \partial_t + c\partial_x,$$

$$f(z-w,s,s_x) = A(\gamma)(z-w)^{\frac{2\gamma}{\gamma-1}} \cdot e^{\frac{-s}{\gamma-1}} \cdot s_x,$$

$$c(z-w,s) = B(\gamma) \cdot e^{\frac{-s}{\gamma-1}} \cdot (z-w)^{\frac{\gamma+1}{\gamma-1}},$$

$$A(\gamma) = \left(\frac{\gamma-1}{16\gamma}\right)^{\frac{\gamma}{\gamma-1}},$$

$$B(\gamma) = \sqrt{\gamma(\gamma-1)} \cdot \left(\frac{\gamma-1}{16\gamma}\right)^{\frac{\gamma+1}{2(\gamma-1)}}.$$

For simplicity, we will only discuss system (1.1) for gas dynamics from now on.

Introduce $g(z-w,s,s_x) = g_1(z-w,s) + g_2(z-w,s)s_x$, in which g_1 and g_2 are chosen so that

$$g'_1 = -\frac{\alpha}{4}c^{-\frac{1}{2}},$$

$$g'_2 = -\frac{1}{4(\gamma-1)}c^{\frac{1}{2}},$$

where g'_i denotes $\frac{\partial g_i}{\partial(z-w)} (i = 1,2)$ (let us recall that c may be viewed as a function of $z - w$ only).

We may choose

$$(2.6)\qquad g_1(z-w,s) = \frac{\alpha(\gamma-1)}{2(3-\gamma)}[B(\gamma)]^{-\frac{1}{2}} e^{\frac{s}{2(\gamma-1)}} \cdot (z-w)^{\frac{\gamma-3}{2(\gamma-1)}},$$

$$(2.7)\qquad g_2(z-w,s) = \frac{-1}{2(3\gamma-1)}[B(\gamma)]^{-\frac{1}{2}} e^{\frac{-s}{2(\gamma-1)}} \cdot (z-w)^{\frac{3\gamma-1}{2(\gamma-1)}}.$$

Then it can be shown that

$$D^+[c^{\frac{1}{2}}z_x+g]=-l[c^{\frac{1}{2}}z_x+g]^2+m_F[c^{\frac{1}{2}}z_x+g]-G_F, \tag{2.8}$$

$$D^-[c^{\frac{1}{2}}w_x+g]=-l[c^{\frac{1}{2}}w_x+g]^2+m_B[c^{\frac{1}{2}}w_x+g]-G_B, \tag{2.9}$$

where

$$\begin{aligned}
&l(z-w,s)=c^{-\frac{1}{2}}\cdot c'\\
&m_F(z-w,s,s_x)=-\tfrac{\alpha}{2}+2g\cdot c^{-\frac{1}{2}}c'+\frac{cs_x}{\gamma-1},\\
&m_B(z-w,s,s_x)=-\tfrac{\alpha}{2}+2g\cdot c^{-\frac{1}{2}}c'-\frac{cs_x}{\gamma-1},\\
&G_F(z-w,s,s_x,s_{xx})=n(z-w,s)+Q_{1F}(z-w,s)s_x\\
&\qquad\qquad +Q_{2F}(z-w,s)s_x^2+Q_{3F}(z-w,s)s_{xx},\\
&G_B(z-w,s,s_x,s_{xx})=n(z-w,s)+Q_{1B}(z-w,s)s_x\\
&\qquad\qquad +Q_{2B}(z-w,s)s_x^2+Q_{3B}(z-w,s)s_{xx},\\
&n(z-w,s)=\frac{\alpha(\gamma-1)}{3-\gamma}g_1(z-w,s),\\
&Q_{1F}(z-w,s)=2g_1g_2\cdot c^{-\frac{1}{2}}c'-\tfrac{\alpha}{2}g_2+\frac{c}{2(\gamma-1)}g_1,\\
&Q_{1B}(z-w,s)=2g_1g_2\cdot c^{-\frac{1}{2}}c'-\tfrac{\alpha}{2}g_2-\frac{c}{2(\gamma-1)}g_1,\\
&Q_{2F}(z-w,s)=-g_{2s}c+\frac{(z-w)}{4\gamma(\gamma-1)}c^{\frac{3}{2}}+g_2\left[g_2\cdot c^{-\frac{1}{2}}c'+\frac{c}{\gamma-1}\right],\\
&Q_{2B}(z-w,s)=g_{2s}c+\frac{(z-w)}{4\gamma(\gamma-1)}c^{\frac{3}{2}}+g_2\left[g_2\cdot c^{-\frac{1}{2}}c'-\frac{c}{\gamma-1}\right],\\
&Q_{3F}(z-w,s)=-g_2c-\frac{(z-w)}{4\gamma}\cdot c^{\frac{3}{2}},\\
&Q_{3B}(z-w,s)=g_2c-\frac{(z-w)}{4\gamma}\cdot c^{\frac{3}{2}}.
\end{aligned}$$

Furthermore, it can be shown that

$$m_F^2-4lG_F=\frac{\alpha^2}{4}+c_{1F}(z-w,s)s_x+c_{2F}(z-w,s)s_x^2+c_{3F}(z-w,s)s_{xx}, \tag{2.10}$$

$$m_B^2-4lG_B=\frac{\alpha^2}{4}+c_{1B}(z-w,s)s_x+c_{2B}(z-w,s)s_x^2+c_{3B}(z-w,s)s_{xx}, \tag{2.11}$$

where

$$c_{1F}(z-w,s)=\frac{2\alpha}{3-\gamma}c-\frac{2\alpha(\gamma+1)}{(3-\gamma)g_1}\cdot Q_{1F},$$

$$c_{2F}(z-w,s)=\frac{4c^2}{(3\gamma-1)^2}-\frac{2\alpha(\gamma+1)}{(3-\gamma)g_1}\cdot Q_{2F},$$

$$c_{3F}(z-w,s)=-\frac{2\alpha(\gamma+1)}{(3-\gamma)g_1}\cdot Q_{3F},$$

$$c_{1B}(z-w,s) = -\frac{4\alpha\gamma}{(3-\gamma)(\gamma-1)}c - \frac{2\alpha(\gamma+1)}{(3-\gamma)g_1}Q_{1B},$$

$$c_{2B}(z-w,s) = \frac{16\gamma^2c^2}{(3\gamma-1)^2(\gamma-1)^2} - \frac{2\alpha(\gamma+1)}{(3-\gamma)g_1} \cdot Q_{2B},$$

$$c_{3B}(z-w,s) = -\frac{2\alpha(\gamma+1)}{(3-\gamma)g_1} \cdot Q_{3B}.$$

It is easy to see that $l(z-w,s) > 0$ for $z-w>0$. Moreover, $m_F^2 - 4lG_F > 0$ and $m_B^2 - 4lG_B > 0$ if $|s_x|$ and $|s_{xx}|$ are small.

Denote $c^{\frac{1}{2}}z_x + g = y$. (2.8) implies that

$$D^+[y] = -l(y-y_1)(y-y_2), \tag{2.12}$$

where

$$y_1 = \frac{(3-\gamma)}{\alpha(\gamma+1)}g_1\{m_F - \sqrt{m_F^2 - 4lG_F}\},$$
$$y_2 = \frac{(3-\gamma)}{\alpha(\gamma+1)}g_1\{m_F + \sqrt{m_F^2 - 4lG_F}\}.$$

Similarly, denote $c^{\frac{1}{2}}w_x + g = Y$. (2.9) implies that

$$D^-[Y] = -l(Y-Y_1)(Y-Y_2), \tag{2.13}$$

where

$$Y_1 = \frac{(3-\gamma)}{\alpha(\gamma+1)}g_1 \cdot \{m_B - \sqrt{m_B^2 - 4lG_B}\},$$
$$Y_2 = \frac{(3-\gamma)}{\alpha(\gamma+1)}g_1 \cdot \{m_B + \sqrt{m_B^2 - 4lG_B}\}.$$

Define

$$D = \{(z,w,s) : a_1 h(\bar{v},\bar{s}) < z-w < a_2 h(\bar{v},\bar{s}), |s-\bar{s}| < 1,$$
$$0 < a_1 < 2 < a_2 < +\infty,\ a_i \text{ is a constant, } i=1,2\},$$

where a_1 and a_2 are chosen so close to 2 such that for a given constant $b : \frac{1}{4} \le b < \frac{1}{2}$, it holds that

$$\sup_D g_1 < \frac{(1+b)\gamma + (1-3b)}{(2-b)\gamma - (2-3b)} \inf_D g_1. \tag{2.14}$$

Thus, it can be shown with a careful calculation that there exists a positive constant δ_1 depending only on D, b, α, and γ such that if, for any $x \in R$,

$$|s'_0(x)| \le \delta_1, \quad |s''_0(x)| \le \delta_1,$$

then it holds that

$$\sup_D y_1 < \inf_D y_2, \tag{2.15}$$

$$\sup_D Y_1 < \inf_D Y_2. \tag{2.16}$$

Furthermore, it can be claimed, with the help of (2.14) and the range of b, that there exist positive constants ε_0 and δ_2, depending only on D, b, α, and γ, such that if, for any $x \in R$,

$$|u'{}_0(x)| \le \varepsilon_0, \qquad |v'{}_0(x)| \le \varepsilon_0, \qquad |s'{}_0(x)| \le \delta_2, \qquad |s''{}_0(x)| \le \delta_2,$$

then it holds that, for any $x \in R$,

$$\begin{aligned} y(x,0) &> \sup_D y_1, \\ Y(x,0) &> \sup_D Y_1, \end{aligned} \tag{2.17}$$

provided $(z, w, s)(x, 0) \in D$.

In view of (2.12)–(2.17), it is not difficult to prove the following lemma.

LEMMA 2.1. *Suppose that the solution of* (1.1) *and* (1.4) *defined for* $0 \le t \le T$ *satisfies* $(z, w, s) \in D$. *Then the following estimates hold if*

$$\begin{aligned} &|u'{}_0(x)| \le \varepsilon_0, \quad |v'{}_0(x)| \le \varepsilon_0, \\ &|s'{}_0(x)| \le \delta_3, \quad |s''{}_0(x)| \le \delta_3, \end{aligned}$$

where $\delta_3 = \min\{\delta_1, \delta_2\}$.

For any $(t, x) \in [0, T] \times R$,

$$\begin{aligned} &\min\left\{ (c^{\frac{1}{2}} z_x + g)(0, x_3(0;t,x)), \inf_{\substack{x_3(\tau;t,x) \\ 0 \le \tau \le t}} y_2 \right\} \\ &\le \quad (c^{\frac{1}{2}} z_x + g)(t,x) \le \max\left\{ (c^{\frac{1}{2}} z_x + g)(0, x_3(0;t,x)), \sup_{\substack{x_3(\tau;t,x) \\ 0 \le \tau \le t}} y_2 \right\}, \\ &\min\left\{ (c^{\frac{1}{2}} w_x + g)(0, x_1(0;t,x)), \inf_{\substack{x_1(\tau;t,x) \\ 0 \le \tau \le t}} Y_2 \right\} \\ &\le \quad (c^{\frac{1}{2}} w_x + g)(t,x) \le \max\left\{ (c^{\frac{1}{2}} w_x + g)(0, x_1(0;t,x)), \sup_{\substack{x_1(\tau;t,x) \\ 0 \le \tau \le t}} Y_2 \right\}, \end{aligned} \tag{2.18}$$

where $x = x_3(\tau; t, x)$ denotes the forward characteristic passing through (t, x) at $\tau = t$ and meeting $(0, x_3(0; t, x))$ at $\tau = 0$ and $x = x_1(\tau; t, x)$ denotes the backward characteristic passing through (t, x) at $\tau = t$ and meeting $(0, x_1(0; t, x))$ at $\tau = 0$.

In order to get the L_2-norm of the solution, we use a different form of system (1.1), namely

$$\begin{cases} v_t - u_x = 0, \\ u_t + p(v,s)_x = -\alpha u, \alpha > 0, \\ [e(v,s) + \frac{u^2}{2}]_t + (pu)_x = -\alpha u^2, \end{cases} \tag{2.19}$$

where e denotes the specific internal energy for which $e_s \ne 0$ and $e_v + p = 0$ holds due to the second law of thermodynamics. In the present discussion, $e(v, s) = \frac{v}{\gamma - 1} p(v, s)$ and $p(v, s)$ is expressed by (2.4).

System (2.19) is equivalent to (1.1) for smooth solutions. $(2.19)_3$ can be written as

$$\left[\frac{1}{2}u^2 + e(v,s) - e(\overline{v},s) + \overline{p}v - \overline{pv}\right]_t + [p - p(\overline{v},\overline{s})u]_x = -\alpha u^2, \tag{2.20}$$

where $\overline{p} = p(\overline{v},\overline{s})$.

Integrate (2.20) over $[0,T] \times R$. By using the Cauchy inequality and the fact that $e_v(\overline{v},\overline{s}) + \overline{p} = 0$, Lemma 2.2 follows.

LEMMA 2.2. *Suppose that the solution of* (1.1) *and* (1.4) *is defined for* $0 \le t \le T$, *with* $(z,w,s) \in D$; *then it holds that*

$$\begin{aligned}&\frac{1}{2}\int_{-\infty}^{\infty} u^2(t,x)dx + \frac{a}{4}\int_{-\infty}^{\infty}(v-\overline{v})^2(t,x)dx\\ &\le \int_{-M}^{M}\left[\frac{1}{2}u_0^2(x) + \frac{3\hat{a}}{4}\cdot v_0^2(x) + \left(\frac{d^2}{a} + \frac{d^2}{\hat{a}}\right)s_0^2(x)\right]dx,\end{aligned} \tag{2.21}$$

where $a = \inf_{(z,w,s)\in D} e_{vv} > 0, \hat{a} = \sup_{(z,w,s)\in D} e_{vv} > 0, d = \sup_{(z,w,s)\in D} |e_v|$, *and* $[-M,M] \supset$ *the compact supports for* $(u_0(x), v_0(x))$ *and* $s_0(x)$, *respectively.*

3. The existence theorem.

THEOREM 3.1. *There exists a globally defined classical solution of* (1.1) *and* (1.4) *with* (2.4) *if the* C_1*-norm of* $(u_0(x), v_0(x))$ *is small and the* C_2*-norm of* $s_0(x)$ *is small.*

Proof. By a routine argument, it can be proved that there exists a local solution of (1.1) (1.4) for $0 \le t \le t_0$ which satisfies $(z,w,s) \in D$ if the C_0-norm of $(u_0(x), v_0(x), s_0(x))$ is so small that

(3.1)

$$\begin{aligned}(z(x,0), w(x,0), s(x,0)) \in D^* = \{&(z,w,s) : A_1 h(\overline{v},\overline{s}) < z - w < A_2 h(\overline{v},\overline{s}),\\ &|s-\overline{s}| < 1, \text{ where } a_1 < A_1 < 2, 2 < A_2 < a_2\}.\end{aligned}$$

Thus, Lemma 2.1 implies that

$$|(u_x, v_x)|_{L_\infty} \le M^* \tag{3.2}$$

if the C_1-norm of $(u_0(x), v(x))$ is small and the C_2-norm of $s_0(x)$ is small, where M^* only depends on D.

Therefore, Lemma 2.2 implies that

$$(z,w,s) \in D^* \tag{3.3}$$

if

$$\int_{-M}^{M}\left[\frac{1}{2}u_0^2(x) + \frac{3\hat{a}}{4}v_0^2(x) + \left(\frac{1}{a} + \frac{1}{\hat{a}}\right)d^2 \cdot s_0^2(x)\right]dx \le \frac{1}{3M^*}[h(\overline{v},\overline{s})]^3 \cdot (4\beta)^{-3}, \tag{3.4}$$

where $\beta = \sup_D |h_v|$.

Thus, there exists a positive τ_0 such that the solution of (1.1) (1.4) can be defined for $0 \le t \le t_0 + \tau_0$, where $(z,w,s) \in D$. Repeat the above argument, the Theorem 3.1 follows.

REMARK 3.2. *The assumption* $\gamma < 3$ *in* (2.4) *is crucial, since* (2.14) *would not be possible otherwise.*

REFERENCES

[HL] L. Hsiao and Tai-ping Liu, *Convergence to nonlinear diffusion waves for solutions of a system of hyperbolic conservation laws with damping*, Comm. Math. Phys., 143 (1992), pp. 599–605.

[HS] L. Hsiao and D. Serre, *Large-time behavior of solutions for the system of compressible adiabatic flow through porous media*, Chinese Ann. Math. Ser. B, 16B (1995), pp. 1–14.

[LI] Tai-ping Liu, *Development of singularities in the nonlinear waves for quasilinear hyperbolic partial differential equations*, J. Differential Equations., 33 (1979), pp. 92–111.

[LZ] T. T. Li, Y. Zhou, and D. X. Kong, *Weak linear degeneracy and global classical solutions for first order quasilinear hyperbolic systems*, in Global Classical Solutions for Quasilinear Hyperbolic Systems, Li Ta-tsien, ed., Wiley, New York, 1994, pp. 132–133.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 78–91, January 1996

A SPATIAL DECAY ESTIMATE FOR THE HYPERBOLIC HEAT EQUATION*

R. QUINTANILLA†

Abstract. In this paper we establish a spatial decay estimate for the hyperbolic heat equation similar to other known estimates for the parabolic equations. Alternatively, the results may be viewed as theorems of Phragmen–Lindelof type [J. B. Conway, *Functions of One Complex Variable*, Graduate Texts in Mathematics, 2nd ed., Springer-Verlag, New York, 1978], [R. Quintanilla, *Publ. Mat.*, 37 (1993), pp. 443–463], [C. O. Horgan and L. E. Payne, *Arch. Rational Mech. Anal.*, 122 (1993), pp. 123–144] for this kind of equation. We conclude the paper by extending the results for a type of semilinear wave equation.

Key words. Phragmen–Lindelof principle, hyperbolic heat equation, spatial decay estimate, semilinear damped hyperbolic equation

AMS subject classifications. 80A20, 30C80, 35B45, 35L70

1. Introduction.

In the last three decades many papers have studied spatial decay estimates for several types of partial differential equations and systems, but very little attention has been devoted to the study of the hyperbolic problems (see [1], [2]). We should mention the works of Flavin and Knops [3] and Flavin, Knops, and Payne [4], [5] as pioneering contributions for this kind of equation. We also recall the recent work [28] on elasticity with voids. Many authors (see [1]–[4]) seem to accept that for hyperbolic problems describing elastic wave propagation, one should not expect spatial decay estimates similar to those obtained for the elliptic and parabolic problems. In this paper we prove that for hyperbolic equations describing dynamical problems with damping effects, we have results similar to those which were obtained in the study of parabolic and pseudoparabolic equations [9]–[21]. In fact we derive a spatial decay estimate for a functional defined on the solutions of an initial-boundary value problem associated with the equation

$$\rho\ddot{u} + \eta\dot{u} = \Delta u, \tag{1.1}$$

where ρ and η are two positive numbers.

We recall that equation (1.1), among others, arises in the studies of heat propagation for an isotropic and homogeneous rigid body for the theories of Muller [6], [7] or Green and Laws [8]. Thus, it seems appropriate to mention some works dedicated to the study of spatial decay estimates for the parabolic linear heat conduction equation [9], [11]–[15], [21] and the nonlinear heat conduction equation [17]–[20]. Spatial decay estimates for the solutions of equation (1.1) for large time have been obtained in [3]. We also mention the papers of Rauch [22] and Cox and Zuazua [29], where the time exponential decay of solutions of equation (1.1) for bounded domains is proved. In [23], Lindsay and Straughan studied wave propagation for the nonlinear hyperbolic heat conduction equation.

Alternatively, the results may be viewed as theorems of Phragmen–Lindelof type [24], [26], [27] for equation (1.1).

* Received by the editors October 13, 1993; accepted for publication (in revised form) March 25, 1994.

† Departamento de Matemática Aplicada II, Universidad Politécnica de Catalunya, Colom 11, Terrassa, 08222 Barcelona, Spain.

We would like to emphasize that the energy decay estimate known for the Laplace equation (see [1]) is quicker than the one obtained here for our hyperbolic equation when the cross-section is small. Thus, the estimate for the parabolic heat equation is also quicker than the estimate for the hyperbolic heat equation.

Though the equation we study in this paper agrees with the one studied in [3], our results apply to the transient function, while the object of [3] was the study of the low frequency range effects.

The method is based on the study of a nonhomogeneous first-order differential inequality. In our case the unknown of the differential inequality is related to a cross-sectional integral involving the temperature scalar field, its time derivative, and its gradient.

In §2, we state the notation and basic definitions. We also prove two lemmas. These lemmas will be used in §3, where the main result of the paper is stated and proved. Thus a spatial decay estimate is obtained. In §4 we obtain an upper bound for a quantity arising in the estimates we have found in §3. This bound is given in terms of the boundary and initial conditions of the problem. We finish the paper in §5 by describing some possible extensions of the methods and results for a class of semilinear wave equations.

2. Preliminaries. In what follows letters in boldface stand for vectors. We shall employ the usual summation and differentiation conventions: Latin subscripts are understood to range over the integers (1,2,3), while Greek subscripts take only the values (2,3); summation over repeated subscripts is implied and subscripts preceded by a comma denote partial differentiation with respect to the corresponding Cartesian coordinate. We use an overdot to denote partial differentiation with respect to the time. As usual ∇ denotes the three-dimensional gradient operator and by $\nabla.$ we denote the three-dimensional divergence operator.

Let B be the interior of a semiinfinite cylinder. We choose the Cartesian coordinates in such a way that the origin lies in the finite end of the cylinder. We can express B as the union of its simply connected two-dimensional cross-sections $D(x_1)$ for all $x_1 \geq 0$ and such that $D(0)$ is contained in the $x_2 0 x_3$ plane. We suppose that the boundary of the cross-sections $\partial D(x_1)$ allows the use of the divergence theorem. By $B(z, z')$ we denote the subset of points $(x_1, x_2, x_3) \in B$ such that $z \leq x_1 \leq z'$.

We shall be concerned with the problem defined by equation (1.1), the boundary conditions

$$u = 0 \text{ on } \partial D(x_1) \times [0, T), \qquad x_1 \geq 0, \qquad T > 0,$$

$$u = f(x_2, x_3, t) \text{ on } D(0) \times [0, T), \tag{2.1}$$

where $f(x_2, x_3, t)$ is a prescribed function, and the initial conditions

$$u(\mathbf{x}, 0) = u^0(\mathbf{x}), \qquad \dot{u}(\mathbf{x}, 0) = v^0(\mathbf{x}), \quad \mathbf{x} \in B. \tag{2.2}$$

We restrict our attention to initial conditions which satisfy

$$\int_B R^0(\mathbf{x}) dv < \infty, \tag{2.3}$$

where

$$R^0(\mathbf{x}) = \frac{\eta}{(1+\rho)} \left(\rho u^0 v^0 + \frac{\eta}{2} u^0 u^0 \right) + \frac{1}{2} (u^0{}_{,i} u^0{}_{,i} + \rho v^0 v^0). \tag{2.4}$$

Now, we define a function on the solutions of the initial-boundary value problem (1.1), (2.1), (2.2). This function is a measure on the solutions. The knowledge of its asymptotic behavior is the objective of §3.

First we state a couple of equalities. Let u be a solution to the equation (1.1); we have

$$\nabla.(u\nabla u) = u\Delta u + u_{,i}u_{,i} = \rho u\ddot{u} + \eta u\dot{u} + u_{,i}u_{,i} = \rho\left\{\frac{d}{dt}(u\dot{u}) - \dot{u}^2\right\} + \frac{\eta}{2}\frac{d}{dt}u^2 + u_{,i}u_{,i}.$$

After an integration from 0 to t_0, we obtain

$$(2.5)\quad \int_0^{t_0} \nabla.(u\nabla u)dt = \int_0^{t_0} (u_{,i}u_{,i} - \rho\dot{u}^2)dt + \rho u\dot{u} + \frac{\eta}{2}u^2 - \rho u^0 v^0 - \frac{\eta}{2}u^0u^0 \text{ for all } t_0 \geq 0.$$

By a similar method we also obtain

$$(2.6)$$
$$\int_0^{t_0} \nabla.(\dot{u}\nabla u)dt = \eta\int_0^{t_0} \dot{u}^2 dt + \frac{1}{2}(u_{,i}u_{,i} + \rho\dot{u}^2) - \frac{1}{2}(u^0{}_{,i}u^0{}_{,i} + \rho v^0 v^0) \text{ for all } t_0 \geq 0.$$

Let $\lambda_0 = \frac{\eta}{1+\rho}$; we may consider

$$(2.7)\qquad F(\mathbf{x}, t_0) = \int_0^{t_0} \nabla.[(\dot{u} + \lambda_0 u)\nabla u]dt \text{ for } \mathbf{x} \in B \text{ and } 0 \leq t_0 < T.$$

LEMMA 2.1. *The function F satisfies the inequality*

$$F(\mathbf{x}, t_0) \geq \lambda_0 \int_0^{t_0} (u_{,i}u_{,i} + \rho\dot{u}^2)dt + \frac{\rho}{2(1+\rho)}\dot{u}^2 + \frac{1}{2}u_{,i}u_{,i} - R^0(\mathbf{x}) \textit{ for all } 0 \leq t_0 < T,$$

where $R^0(\mathbf{x})$ is defined at (2.4).

Proof. For all positive constants ϵ we have the equality

$$\frac{1}{2}\rho\dot{u}^2 + \rho\lambda_0 u\dot{u} + \frac{\lambda_0\eta}{2}u^2$$
$$= \frac{1}{2}\left(\rho\epsilon\dot{u} + \frac{\eta}{(1+\rho)\epsilon}u\right)^2 + \frac{1}{2}(\rho - \rho^2\epsilon^2)\dot{u}^2 + \frac{\eta^2}{2(1+\rho)}(1 - (\epsilon^2(1+\rho))^{-1})u^2.$$

Now, by taking $\epsilon^2 = (1+\rho)^{-1}$ we obtain

$$(2.8)\qquad \frac{1}{2}\rho\dot{u}^2 + \rho\lambda_0 u\dot{u} + \frac{\lambda_0\eta}{2}\dot{u}^2 \geq \frac{\rho}{2(1+\rho)}\dot{u}^2.$$

On the other hand, we also have the equalities

$$(2.9)\quad \lambda_0(u_{,i}u_{,i} - \rho\dot{u}^2) + \eta\dot{u}^2 = \lambda_0 u_{,i}u_{,i} + \frac{1}{1+\rho}(\eta(1+\rho) - \eta\rho)\dot{u}^2 = \lambda_0(u_{,i}u_{,i} + \dot{u}^2).$$

Lemma 2.1 is a direct consequence of equalities (2.5), (2.6), and (2.9) and inequality (2.8).

We now state a lemma which gives some information on the asymptotic behavior of the solutions of a nonhomogeneous first-order ordinary differential inequality. This abstract lemma will be applied to our problem in the next section to conclude spatial decay estimates for the solutions of (1.1).

LEMMA 2.2. *Let ω and R be two positive functions defined on the nonnegative real line such that*

$$\lim_{z\to\infty}\int_z^\infty R(t)\exp\Big[-\int_0^t \omega^{-1}(\chi)d\chi\Big]dt = 0. \tag{2.10}$$

Let $G(t)$ be a C^1 function satisfying

$$G \le \omega(t)\left(\frac{dG}{dt}+R(t)\right),$$

$$\lim_{t\to\infty} G(t)\Big[-\int_0^t \omega^{-1}(\chi)d\chi\Big] = 0. \tag{2.11}$$

Then

$$\limsup_{t\to\infty} G(t) \le 0.$$

Proof. First we consider the function

$$\begin{aligned} R_1(t) = & R_1(t_0)\exp\Big[\int_{t_0}^t \omega^{-1}(\chi)d\chi\Big] \\ & + \exp\Big[\int_0^t \omega^{-1}(\chi)d\chi\Big]\int_{t_0}^t R(\chi)\exp\Big[-\int_0^\chi \omega^{-1}(\xi)d\xi\Big]\,d\chi. \end{aligned}$$

After some calculations we get

$$\omega R = \omega\frac{d}{dt}R_1 - R_1.$$

Thus we conclude

$$G + R_1 \le \omega(t)\frac{d}{dt}(G+R_1) \text{ for all } t \ge t_0 \ge 0.$$

Let us suppose that there exists t_0 such that $G(t_0) > 0$; we obtain

$$(G+R_1)(t) \ge (G+R_1)(t_0)\exp\Big[\int_{t_0}^t \omega^{-1}(\chi)d\chi\Big] \text{ for all } t \ge t_0.$$

Thus we deduce

$$G(t) \ge \exp\Big[\int_{t_0}^t \omega^{-1}(\chi)d\chi\Big]\left\{G(t_0) - \int_{t_0}^t R(\chi)\exp\Big[-\int_0^\chi \omega^{-1}(\xi)d\xi\Big]\,d\chi\right\} \text{ for all } t \ge t_0.$$

Now if there exists t_0 such that

$$G(t_0) > \int_{t_0}^\infty R(\chi)\exp\Big[-\int_0^\chi \omega^{-1}(\xi)d\xi\Big]\,d\chi,$$

the asymptotic condition (2.11) on G would not be satisfied. Thus we obtain

$$G(t) \le \int_t^\infty R(\chi)\exp\Big[-\int_0^\chi \omega^{-1}(\xi)d\xi\Big]\,d\chi \text{ for all } t \ge 0.$$

In view of (2.10) Lemma 2.2 is proved.

3. A decay theorem. In this section we state a spatial decay estimate for the solutions of the problem defined by (1.1), (2.1), (2.2). We define a function on the cross-sections

$$G(z,t_0)=\int_0^{t_0}\int_{D(z)} u_{,1}(\dot{u}+\lambda_0 u)\,dadt \text{ for all } z\geq 0 \text{ and } t_0>0. \tag{3.1}$$

Using the divergence theorem we obtain

$$G(z+h,t_0)=G(z,t_0)+\int_{B(z,z+h)} F(\mathbf{x},t_0)\,dv \text{ for all } h>0 \text{ and } t_0>0. \tag{3.2}$$

Direct differentiation gives

$$\frac{\partial G}{\partial z}(z,t_0)=\int_{D(z)} F(\mathbf{x},t_0)\,da.$$

Using the Schwarz inequality we find

$$|G(z,t_0)|\leq\int_0^{t_0}\left[\left(\int_{D(z)} u_{,1}u_{,1}da\right)^{\frac{1}{2}}\left\{\left(\int_{D(z)}\dot{u}^2da\right)^{\frac{1}{2}}+\lambda_0\left(\int_{D(z)}u^2da\right)^{\frac{1}{2}}\right\}\right]dt.$$

From Poincaré and the arithmetic-geometric mean inequalities we deduce

$$\begin{aligned}|G(z,t_0)|&\leq\frac{1}{2}[(\epsilon_1)^{-1}+(\epsilon_2)^{-1}]\int_0^{t_0}\int_{D(z)}u_{,1}u_{,1}dadt\\&+\frac{\epsilon_1}{2}\int_0^{t_0}\int_{D(z)}\dot{u}^2dadt+\lambda_0^2\alpha(z)^{-1}\frac{\epsilon_2}{2}\int_0^{t_0}\int_{D(z)}u_{,\alpha}u_{,\alpha}dadt,\end{aligned}$$

where $\alpha(z)$ is the first nonzero eigenvalue of the Laplacian operator in $D(z)$ with homogeneous Dirichlet boundary condition. Now, if we take $\epsilon_1(z)=\epsilon_2(z)\rho\lambda_0^2\alpha(z)^{-1}$ and $\epsilon_2(z)=\lambda_0^{-1/2}\alpha(z)^{1/2}(1+\alpha(z)\rho^{-1}\lambda_0^{-2})^{1/2}$, we obtain

$$\begin{aligned}|G(z,t_0)|&\leq\frac{1}{2}[\lambda_0^{-1}\rho^{-1}+\lambda_0\alpha(z)^{-1}]^{1/2}\frac{\partial G}{\partial z}(z,t_0)\\&+\frac{1}{2}\int_{D(z)}[\lambda_0^{-1}\rho^{-1}+\lambda_0\alpha(z)^{-1}]^{1/2}R^0(\mathbf{x})da.\end{aligned}\tag{3.3}$$

Now we recall the Faber–Krahn inequality

$$\alpha(z)\geq\pi j_0^2A^{-1}(z),\tag{3.4}$$

where j_0 is the smallest positive zero of the Bessel function J_0 and $A(z)$ is the area of the section $D(z)$. Inequalities (3.3) and (3.4) lead to the estimates

$$G(z,t_0)\leq\omega(z)^{1/2}\frac{\partial G}{\partial z}(z,t_0)+\omega(z)^{1/2}\int_{D(z)}R^0(\mathbf{x})da,$$

$$-G(z,t_0)\leq\omega(z)^{1/2}\frac{\partial G}{\partial z}(z,t_0)+\omega(z)^{1/2}\int_{D(z)}R^0(\mathbf{x})da,\tag{3.5}$$

where $\omega(z)=\frac{1}{4}[\lambda_0^{-1}\rho^{-1}+\lambda_0A(z)/\pi j_0^2]$.

If we suppose that the initial condition satisfies (2.3) as well as the following asymptotic condition on the solutions

$$\lim_{z\to\infty} G(z,t_0)\exp\left[-\int_0^z \omega(\tau)^{-1/2}d\tau\right]=0, \tag{3.6}$$

then Lemma 2.2 and the first inequality of (3.5) imply

$$\limsup_{z\to\infty} G(z,t_0)\le 0. \tag{3.7}$$

From the second inequality in (3.5) we get

$$\begin{aligned}&-G(z,t_0)\le -G(z_0,t_0)\exp\left[-\int_{z_0}^z \omega(\tau)^{-1/2}d\tau\right]\\ &\quad+\exp\left[-\int_0^z \omega(\tau)^{-1/2}d\tau\right]\int_{z_0}^z \exp\left[\int_0^{x_1}\omega(\tau)^{-1/2}d\tau\right]\Big(\int_{D(x_1)}R^0(\mathbf{x})da\Big)dx_1\end{aligned} \tag{3.8}$$

for all $z\ge z_0\ge 0$.

Equalities (3.2) and (3.4) imply

$$\begin{aligned}&\int_{B(z,z^*)}F(\mathbf{x},t_0)dv\le G(z^*,t_0)\\ &\quad+\left[\int_{B(z_0,z^*)}F(\mathbf{x},t_0)dv-G(z^*,t_0)\right]\exp\left[-\int_{z_0}^z \omega(\tau)^{-1/2}d\tau\right]\\ &\quad+\exp\left[-\int_0^z \omega(\tau)^{-1/2}d\tau\right]\int_{z_0}^z \exp\left[\int_0^{x_1}\omega(\tau)^{-1/2}d\tau\right]\Big(\int_{D(x_1)}R^0(\mathbf{x})da\Big)dx_1\end{aligned} \tag{3.9}$$

for all $z^*\ge z\ge z_0\ge 0$.

Thus we have proved the following result.

PROPOSITION 3.1. *Let u be a solution of the initial-boundary value problem* (1.1), (2.1), (2.2) *that satisfies the asymptotic condition* (3.6). *We also suppose that the initial conditions satisfy* (2.3). *Then the estimate* (3.9) *is satisfied for all $z^*\ge z\ge z_0\ge 0$.*

In case there exists $z^*>0$ such that $G(z^*,t_0)=0$, the following estimate holds for all $z^*\ge z\ge z_0\ge 0$:

$$\begin{aligned}&\int_{B(z,z^*)}F(\mathbf{x},t_0)dv\le\left[\int_{B(z_0,z^*)}F(\mathbf{x},t_0)dv\right]\exp\left[-\int_{z_0}^z \omega(\tau)^{-1/2}d\tau\right]\\ &\quad+\exp\left[-\int_0^z \omega(\tau)^{-1/2}d\tau\right]\int_{z_0}^z \exp\left[\int_0^{x_1}\omega(\tau)^{-1/2}(\tau)d\tau\right]\Big(\int_{D(x_1)}R^0(\mathbf{x})da\Big)dx_1.\end{aligned} \tag{3.10}$$

Remark. If we relax condition (2.3) on the initial conditions to the asymptotic condition

$$\lim_{z\to\infty}\int_z^\infty\Big(\int_{D(x_1)}R^0(\mathbf{x})da\Big)\exp\left[-\int_0^{x_1}(\omega(\tau))^{-1/2}d\tau\right]dx_1=0, \tag{3.11}$$

then inequalities (3.9) and (3.10) also hold.

Now if we suppose

$$\int_0^{\infty} \omega(\tau)^{-1/2} d\tau = \infty \tag{3.12}$$

and

$$\lim_{z\to\infty} \omega(z) \int_{D(z)} R^0(\mathbf{x}) da = 0, \tag{3.13}$$

then after l'Hôpital's rule we have

$$\lim_{z\to\infty} \exp\Big[-\int_0^z \omega(\tau)^{-1/2} d\tau\Big] \int_{z_0}^z \exp\Big[\int_0^{x_1} \omega(\tau)^{-1/2} d\tau\Big] \Big(\int_{D(x_1)} R^0(\mathbf{x}) da\Big) dx_1 = 0.$$

Inequalities (3.7) and (3.8) imply that $G(\infty, t_0) = 0$. Thus, we may state the following result.

THEOREM 3.1. *Let u be a solution of the initial-boundary value problem* (1.1), (2.1), (2.2) *that satisfies the asymptotic condition* (3.6). *If the cylinder satisfies* (3.12) *and the initial conditions satisfy* (2.3) *and* (3.13), *then the following estimate holds:*

(3.14)

$$\begin{aligned}&\int_{B(z,\infty)} \Big[\eta \int_0^{t_0} (u_{,i}u_{,i} + \rho \dot{u}^2) d\tau + \frac{\rho}{2}\dot{u}^2 + \frac{1+\rho}{2} u_{,i}u_{,i}\Big] dv \\ &\leq (1+\rho)\Big[\int_{B(z,\infty)} R^0(\mathbf{x}) dv + \Big[\int_{B(z_0,\infty)} F(\mathbf{x}, t_0) dv\Big] \exp\Big[-\int_{z_0}^z \omega(\tau)^{-1/2} d\tau\Big] \\ &+ \exp\Big[-\int_0^z \omega(\tau)^{-1/2} d\tau\Big] \int_{z_0}^z \exp\Big[\int_0^{x_1} \omega(\tau)^{-1/2}(\tau) d\tau\Big] \Big(\int_{D(x_1)} R^0(\mathbf{x}) da\Big) dx_1\Big]\end{aligned}$$

for all $z_0 \leq z \leq \infty$.

The methods used to prove Theorem 3.1 can be adapted for bounded cylinders whenever we adjoin the condition to the boundary conditions (2.1):

$$u = 0 \text{ on } D(z^*) \times [0, T),$$

where $D(z^*)$ is the other plane face of the cylinder.

If the area of the cross-sections satisfies the inequality

$$A(z) \leq Cz^2 \text{ for all } z \geq 0,$$

condition (3.12) holds. Now condition (3.13) will be satisfied whenever

$$\lim_{z\to\infty} z \int_{D(z)} R^0(\mathbf{x}) da = 0.$$

Many other sufficient conditions can be given to satisfy (3.12), (3.13).

In the case in which the initial conditions for the temperature and the time derivative of the temperature are zero, inequality (3.14) implies

$$\begin{aligned}&\int_{B(z,z^*)} \Big[\eta \int_0^{t_0} (u_{,i}u_{,i} + \rho \dot{u}^2) d\tau + \frac{\rho}{2}\dot{u}^2 + \frac{1+\rho}{2} u_{,i}u_{,i}\Big] dv \\ &\leq (1+\rho) \left[\int_{B(z_0,z^*)} F(\mathbf{x}, t_0) dv\right] \exp\Big[-\int_{z_0}^z \omega(\tau)^{-1/2} d\tau\Big]\end{aligned}$$

for all $z^* \geq z \geq z_0$.

4. A bound for $\int_B F(\mathbf{x},t)dv$. Estimate (3.11) is not of practical use unless the quantity

$$\mathcal{Q}(t_0) = \int_B F(\mathbf{x},t)dv$$

is bounded in terms of the prescribed data. The object of this section is to obtain an estimate for $\mathcal{Q}(t_0)$ in terms of boundary conditions (2.1) and initial conditions (2.2). For simplicity, we assume that the cylinder is semiinfinite and prismatic. We follow the methods used in [25] (see also [26]).

Let ξ be a function defined on $B \times [0,t_0]$, chosen to satisfy boundary conditions (2.1). The divergence theorem shows

$$\mathcal{Q}(t_0) = \int_0^{t_0} \int_{D(0)} u_{,1}(\dot{\xi}+\lambda_0\xi)dadt = \int_0^{t_0} \int_B \{u_{,i}(\dot{\xi}_{,i}+\lambda_0\xi_{,i})+(\rho\ddot{u}+\eta\dot{u})(\dot{\xi}+\lambda_0\xi)\}dvdt.$$

Use of the Schwarz inequality and integration by parts allows us to obtain the next inequality

$$\begin{aligned}\mathcal{Q}(t_0) \leq & \Big(\int_0^{t_0} \int_B u_{,i}u_{,i}dvdt\Big)^{\frac{1}{2}} \times \Big(\int_0^{t_0} \int_B (\dot{\xi}_{,i} + \lambda_0\xi_{,i})(\dot{\xi}_{,i} + \lambda_0\xi_{,i})dvdt\Big)^{\frac{1}{2}} \\ & + \int_B \rho[\dot{u}(t_0)\dot{\xi}(t_0) + \lambda_0\dot{u}(t_0)\xi(t_0) - (\dot{u}(0)\dot{\xi}(0) + \lambda\dot{u}(0)\xi(0))]dv \\ & + \int_0^{t_0} \int_B \left[\dot{u}\left(\lambda_0\eta\xi + \frac{\eta}{1+\rho}\dot{\xi} - \rho\ddot{\xi}\right)\right] dvdt.\end{aligned} \tag{4.1}$$

Using the Schwarz inequality again we obtain

$$\int_B \rho[\dot{u}\dot{\xi} + \lambda_0\dot{u}\xi]dv \leq \Big(\int_B \rho\dot{u}\dot{u}\Big)^{\frac{1}{2}} \times \Big(\int_B \rho(\dot{\xi} + \lambda_0\xi)^2 dv\Big)^{\frac{1}{2}}$$

and

$$\begin{aligned}&\int_0^{t_0} \int_B \left[\dot{u}\left(\lambda_0\eta\xi + \frac{\eta}{\rho}\dot{\xi} - \rho\ddot{\xi}\right)\right] dvdt \\ &\leq \Big(\int_0^{t_0} \int_B \dot{u}\dot{u}dvdt\Big)^{\frac{1}{2}} \times \Big(\int_0^{t_0} \int_B \Big(\lambda_0\eta\xi + \frac{\eta}{1+\rho}\dot{\xi} - \rho\ddot{\xi}\Big)^2 dvdt\Big)^{\frac{1}{2}}.\end{aligned}$$

Thus we have

$$\begin{aligned}\mathcal{Q}(t_0) \leq & \Big[\int_0^{t_0} \int_B (u_{,i}u_{,i} + \dot{u}\dot{u})dvdt + \int_B \rho\dot{u}\dot{u}dv\Big]^{\frac{1}{2}} \\ & \times \Big[\Big(\int_0^{t_0} \int_B (\dot{\xi}_{,i} + \lambda_0\xi_{,i})(\dot{\xi}_{,i} + \lambda_0\xi_{,i})dvdt\Big)^{\frac{1}{2}} \\ & \quad + \Big(\int_B \rho(\dot{\xi} + \lambda_0\xi)^2 dv\Big)^{\frac{1}{2}} + \Big(\int_0^{t_0} \int_B \Big(\lambda_0\eta\xi + \frac{\eta}{1+\rho}\dot{\xi} - \rho\ddot{\xi}\Big)^2 dvdt\Big)^{\frac{1}{2}}\Big] \\ & \quad - \int_B \Big[\rho(\dot{u}(0)\dot{\xi}(0) + \lambda_0\dot{u}(0)\xi(0))\Big]dv.\end{aligned}$$

Using Lemma 2.1 we deduce

(4.2)

$$\begin{aligned}\mathcal{Q}(t_0) \leq & \kappa\Big[Q(t_0) + \int_B R^0(\mathbf{x})dv\Big]^{\frac{1}{2}} \\ & \times \Big[\Big(\int_0^{t_0}\int_B (\dot{\xi}_{,i} + \lambda_0\xi_{,i})(\dot{\xi}_{,i} + \lambda_0\xi_{,i})dvdt\Big)^{\frac{1}{2}} \\ & + \Big(\int_B \rho(\dot{\xi} + \lambda_0\xi)^2 dv\Big)^{\frac{1}{2}} + \Big(\int_0^{t_0}\int_B \Big(\lambda_0\eta\xi + \frac{\eta}{1+\rho}\dot{\xi} - \rho\ddot{\xi}\Big)^2 dvdt\Big)^{\frac{1}{2}}\Big] \\ & - \int_B \Big[\rho(\dot{u}(0)\dot{\xi}(0) + \lambda_0\dot{u}(0)\xi(0))\Big]dv,\end{aligned}$$

where $\kappa = 3\{\max(\frac{1+\rho}{\eta}, 2(1+\rho), \frac{1+\rho}{\eta\rho})\}^{1/2}$ and $R^0(\mathbf{x})$ is defined in (2.4). Because of the positivity of $R^0(\mathbf{x})$, we obtain the inequality

$$\begin{aligned}& \Big[\Big\{\mathcal{Q}(t_0)) + \int_B R^0(\mathbf{x})dv\Big\}^{\frac{1}{2}} - \frac{\kappa}{2}\Big(\Big(\int_0^{t_0}\int_B (\dot{\xi}_{,i} + \lambda_0\xi_{,i})(\dot{\xi}_{,i} + \lambda_0\xi_{,i})dvdt\Big)^{\frac{1}{2}} \\ & \quad + \Big(\int_B \rho(\dot{\xi} + \lambda_0\xi)^2 dv\Big)^{\frac{1}{2}} + \Big(\int_0^{t_0}\int_B \Big(\lambda_0\eta\xi + \frac{\eta}{1+\rho}\dot{\xi} - \rho\ddot{\xi}\Big)^2 dvdt\Big)^{\frac{1}{2}}\Big)\Big]^2 \\ & \leq \Big|\int_B \rho(\dot{u}(0)\dot{\xi}(0) + \lambda_0\dot{u}(0)\xi(0))]dv\Big| + \int_B R^0(\mathbf{x})dv \\ & \quad + \frac{\kappa^2}{4}\Big(\Big(\int_0^{t_0}\int_B (\dot{\xi}_{,i} + \lambda_0\xi_{,i})(\dot{\xi}_{,i} + \lambda_0\xi_{,i})dvdt\Big)^{\frac{1}{2}} + \Big(\int_B \rho(\dot{\xi} + \lambda_0\xi)^2 dv\Big)^{\frac{1}{2}} \\ & \quad + \Big(\int_0^{t_0}\int_B \Big(\lambda_0\eta\xi + \frac{\eta}{1+\rho}\dot{\xi} - \rho\ddot{\xi}\Big)^2 dvdt\Big)^{\frac{1}{2}}\Big)^2\end{aligned}$$

Now we may deduce the estimate

(4.3)

$$\begin{aligned}\mathcal{Q}(t_0) \leq & -\int_B R^0(\mathbf{x})dv + \Big[\Big(\Big|\int_B \rho(\dot{u}(0)\dot{\xi}(0) + \lambda_0\dot{u}(0)\xi(0))]dv\Big| + \int_B R^0(\mathbf{x})dv \\ & + \frac{\kappa^2}{4}\Big(\Big(\int_0^{t_0}\int_B (\dot{\xi}_{,i} + \lambda_0\xi_{,i})(\dot{\xi}_{,i} + \lambda_0\xi_{,i})dvdt\Big)^{\frac{1}{2}} + \Big(\int_B \rho(\dot{\xi} + \lambda_0\xi)^2 dv\Big)^{\frac{1}{2}} \\ & + \Big(\int_0^{t_0}\int_B \Big(\lambda_0\eta\xi + \frac{\eta}{1+\rho}\dot{\xi} - \rho\ddot{\xi}\Big)^2 dvdt\Big)^{\frac{1}{2}}\Big)^2\Big)^{\frac{1}{2}} \\ & + \frac{\kappa}{2}\Big(\Big(\int_0^{t_0}\int_B (\dot{\xi}_{,i} + \lambda_0\xi_{,i})(\dot{\xi}_{,i} + \lambda_0\xi_{,i})dvdt\Big)^{\frac{1}{2}} + \Big(\int_B \rho(\dot{\xi} + \lambda_0\xi)^2 dv\Big)^{\frac{1}{2}} \\ & + \Big(\int_0^{t_0}\int_B \Big(\lambda_0\eta\xi + \frac{\eta}{1+\rho}\dot{\xi} - \rho\ddot{\xi}\Big)^2 dvdt\Big)^{\frac{1}{2}}\Big)\Big]^2.\end{aligned}$$

We make the following choice:

$$\xi(\mathbf{x}, t) = f(x_\beta, t)\exp(-\alpha x_1),$$

where α is a positive constant and $f(x_\beta, t)$ is defined in (2.1).

Easy calculations lead to

$$
\begin{aligned}
(4.4)\qquad &\dot\xi(\mathbf{x},t) = \dot f(x_\beta,t)\exp(-\alpha x_1), \quad \ddot\xi(\mathbf{x},t) = \ddot f(x_\beta,t)\exp(-\alpha x_1),\\
&\xi_{,\omega}(\mathbf{x},t) = f_{,\omega}(x_\beta,t)\exp(-\alpha x_1), \quad \xi_{,1}(\mathbf{x},t) = -\alpha f(x_\beta,t)\exp(-\alpha x_1),\\
&\dot\xi_{,\omega}(\mathbf{x},t) = \dot f_{,\omega}(x_\beta,t)\exp(-\alpha x_1), \quad \text{and } \dot\xi_{,1}(\mathbf{x},t) = \alpha\dot f(x_\beta,t)\exp(-\alpha x_1).
\end{aligned}
$$

By using equalities (4.4) we may obtain the following equalities for the right-hand side terms of (4.3):

$$
\int_B(\dot u(0)\dot\xi(0) + \lambda_0\dot u(0)\xi(0))dv = \int_B v^0(\mathbf{x})(\dot f(x_\beta,0) + \lambda_0 f(x_\beta,0))\exp(-\alpha x_1)dv,
$$

$$
\begin{aligned}
&\int_0^{t_0}\int_B(\dot\xi_{,i} + \lambda_0\xi_{,i})(\dot\xi_{,i} + \lambda_0\xi_{,i})dvdt\\
&= \frac{1}{2\alpha}\int_0^{t_0}\int_D(\dot f_{,\omega}(x_\beta,t) + \lambda_0 f_{,\omega}(x_\beta,t))(\dot f_{,\omega}(x_\beta,t) + \lambda_0 f_{,\omega}(x_\beta,t))dadt\\
&\quad + \frac{\alpha}{2}\int_0^{t_0}\int_D(\dot f(x_\beta,t) + \lambda_0 f(x_\beta,t))^2 dadt,
\end{aligned}
$$

$$
\int_B\rho(\dot\xi + \lambda_0\xi)^2dv = \frac{1}{2\alpha}\int_D(\dot f(x_\beta,t) + \lambda_0 f(x_\beta,t))^2 da,
$$

and

$$
\begin{aligned}
&\int_0^{t_0}\int_B\Big(\lambda_0\eta\xi + \frac{\eta}{1+\rho}\dot\xi - \rho\ddot\xi\Big)^2 dvdt\\
(4.5)\qquad &= \frac{1}{2\alpha}\int_0^{t_0}\int_D\Big(\lambda_0\eta f(x_\beta,t) + \frac{\eta}{1+\rho}\dot f(x_\beta,t) - \rho\ddot f(x_\beta,t)\Big)^2 dvdt.
\end{aligned}
$$

Inequality (4.3) and equalities (4.5) give the desired estimate for $\mathcal{Q}(t_0)$ in terms of boundary conditions (2.1) and initial conditions (2.2). Thus we have

(4.6)

$$
\begin{aligned}
&\mathcal{Q}(t_0)\\
&\leq -\int_B R^0(\mathbf{x})dv + \Big[\Big(\Big|\int_B\rho v^0(\mathbf{x})(\dot f(x_\beta,0) + \lambda_0 f(x_\beta,0))\exp(-\alpha x_1))dv\Big| + \int_B R^0(\mathbf{x})dv\\
&\quad + \frac{\kappa^2}{4}\Big(\Big(\int_0^{t_0}\int_D\Big[\frac{1}{2\alpha}(\dot f_{,\omega}(x_\beta,t) + \lambda_0 f_{,\omega}(x_\beta,t))(\dot f_{,\omega}(x_\beta,t) + \lambda_0 f_{,\omega}(x_\beta,t))\\
&\quad + \frac{\alpha}{2}(\dot f(x_\beta,t) + \lambda_0 f(x_\beta,t))^2\Big]dadt\Big)^{\frac{1}{2}} + \Big(\frac{1}{2\alpha}\int_D(\dot f(x_\beta,t) + \lambda_0 f(x_\beta,t))^2 da\Big)^{\frac{1}{2}}\\
&\quad + \Big(\frac{1}{2\alpha}\int_0^{t_0}\int_D\Big(\lambda_0\eta f(x_\beta,t) + \frac{\eta}{1+\rho}\dot f(x_\beta,t) - \rho\ddot f(x_\beta,t)\Big)^2 dvdt\Big)^{\frac{1}{2}}\Big)^2\Big)^{\frac{1}{2}}\\
&\quad + \frac{\kappa}{2}\Big(\Big(\int_0^{t_0}\int_D\Big[\frac{1}{2\alpha}(\dot f_{,\omega}(x_\beta,t) + \lambda_0 f_{,\omega}(x_\beta,t))(\dot f_{,\omega}(x_\beta,t) + \lambda_0 f_{,\omega}(x_\beta,t))\\
&\quad + \frac{\alpha}{2}(\dot f(x_\beta,t) + \lambda_0 f(x_\beta,t))^2\Big]dadt\Big)^{\frac{1}{2}} + \Big(\frac{1}{2\alpha}\int_D(\dot f(x_\beta,t) + \lambda_0 f(x_\beta,t))^2 da\Big)^{\frac{1}{2}}\\
&\quad + \Big(\frac{1}{2\alpha}\int_0^{t_0}\int_D\Big(\lambda_0\eta f(x_\beta,t) + \frac{\eta}{1+\rho}\dot f(x_\beta,t) - \rho\ddot f(x_\beta,t)\Big)^2 dvdt\Big)^{\frac{1}{2}}\Big)\Big]^2.
\end{aligned}
$$

Remark. To obtain inequality (4.3), we have not made use of the prismatic form of the cylinder. Estimates like (4.6) could be deduced by changing the choice of the function ξ in a suitable way. In the case in which the section $D(x_1)$ can be obtained from $D(0)$ by a homothetic motion of ratio $1 + cx_1$ (where $c > 0$), we may take

$$\xi(\mathbf{x}, t) = f\left(\frac{x_2}{1+cx_1}, \frac{x_3}{1+cx_1}, t\right) \exp(-\alpha x_1),$$

where α is a positive constant and f is given in (2.1). The reader can conclude the analysis in a way similar to that in the appendix of [26].

5. An extension for semilinear wave equations. In this section we discuss some possible extensions of the previous methods and results to nonlinear damped hyperbolic equations. Let us consider the semilinear wave equation of the form

$$\ddot{u} + g(\dot{u}) = \Delta u, \tag{5.1}$$

where g is a nonlinear function.

Some recent papers devoted to the study of temporal decay estimates for semilinear wave equations are [30]–[34] (see also the references therein). Now we suppose that g satisfies

$$g(s)s \geq c_1|s|^2 \text{ and } |g(s)| \leq c_2|s|. \tag{5.2}$$

To avoid cumbersome calculations we will make a couple of simplifications in the statement of the problem. We consider the problem determined by equation (5.1), boundary conditions (2.1), and the initial conditions

$$u(\mathbf{x}, 0) = 0, \quad \dot{u}(\mathbf{x}, 0) = 0, \quad \mathbf{x} \in B. \tag{5.3}$$

We suppose that the cross-section is constant for all $x_1 \geq 0$ and we denote that cross-section by D. Similar calculations to those used in §2 lead to the equalities

$$\int_0^{t_0} \nabla.(u\nabla u)dt = u\dot{u} + \int_0^{t_0} (u_{,i}u_{,i} + ug(\dot{u}) - \dot{u}\dot{u})dt,$$

$$\int_0^{t_0} \nabla.(\dot{u}\nabla u)dt = \frac{1}{2}(\dot{u}\dot{u} + u_{,i}u_{,i}) + \int_0^{t_0} g(\dot{u})\dot{u}dt.$$

In a way similar to the definition of the function G in §3, we define the function

$$H(z, t_0) = \int_0^{t_0} \int_{D(z)} u_{,1}(\dot{u} + \lambda_1 u)dadt \text{ for all } z \geq 0 \text{ and } t_0 > 0,$$

where λ_1 is a positive constant to be specified later. Also, the use of the divergence theorem leads to the equality

(5.4)
$$H(z+h, t_0) = H(z, t_0) + \int_{B(z,z+h)} \int_0^{t_0} \nabla.[(\dot{u} + \lambda_1 u)\nabla u]dtdv \text{ for all } h > 0 \text{ and } t_0 > 0.$$

Direct differentiation gives

$$\frac{\partial H}{\partial z}(z, t_0) = \int_D \int_0^{t_0} \nabla.[(\dot{u} + \lambda_1 u)\nabla u]dtda.$$

Now we take λ_1, a positive number which is less than $\alpha(0)^{\frac{1}{2}}$. As in §3, $\alpha(0)$ is the first nonzero eigenvalue of the Laplacian operator in D with homogeneous boundary conditions. We also recall estimate (3.4), which relates $\alpha(0)$ with the area of D. Thus we obtain

$$\frac{\partial}{\partial z}H(z,t) \geq \int_D \int_0^{t_0} \Big(\lambda_1(u_{,i}u_{,i} + ug(\dot{u}) - \dot{u}\dot{u}) + g(\dot{u})\dot{u}\Big)dtds. \tag{5.5}$$

Using the Schwarz inequality and the second inequality in (5.2) we have

$$\int_D ug(\dot{u})ds \leq c_2\Big(\int_D u^2ds\Big)^{\frac{1}{2}}\Big(\int_D \dot{u}^2ds\Big)^{\frac{1}{2}}.$$

After using the arithmetic-geometric mean and Poincaré inequalities we conclude

$$\int_D ug(\dot{u})ds \leq \frac{c_2}{2\epsilon\alpha(0)}\int_D u_{,i}u_{,i}ds + \frac{c_2\epsilon}{2}\int_D \dot{u}^2ds \tag{5.6}$$

for all positive ϵ. From (5.5) and (5.6) we deduce

$$\frac{\partial}{\partial z}H(z,t) \geq \int_D \int_0^{t_0} \Big(\lambda_1\Big(1 - \frac{c_2}{2\epsilon\alpha(0)}\Big)u_{,i}u_{,i} + \Big(c_1 - \lambda_1\Big(1 + \frac{c_2\epsilon}{2}\Big)\Big)\dot{u}^2\Big)dtds.$$

We may always take $\epsilon > 0$ so large that

$$1 - \frac{c_2}{2\epsilon\alpha(0)} \geq \beta_1 > 0$$

and $\lambda_1 > 0$ so small that

$$c_1 - \lambda_1\Big(1 + \frac{c_2\epsilon}{2}\Big) \geq \beta_2 > 0.$$

Then we obtain

$$\frac{\partial}{\partial z}H(z,t) \geq \omega\int_D \int_0^{t_0} \Big(u_{,i}u_{,i} + \dot{u}^2\Big)dtds \text{ for all } z \geq 0 \text{ and } 0 \leq t_0 < T, \tag{5.7}$$

where $\omega = \min(\lambda_1\beta_1, \beta_2)$. Now we may proceed in a way similar to §3 to deduce spatial exponential decay for the solutions to the problem determined by equation (5.1), boundary conditions (2.1), and initial conditions (5.3).

Remark. The methods used in the previous sections could be extended to this kind of equation. We could consider other initial conditions and more general geometries. Of course, new initial conditions and new geometries imply new decay estimates. We also may obtain bounds for

$$\int_{B(0,\infty)} \int_0^{t_0} \nabla.[(\dot{u} + \lambda_1 u)\nabla u]dtdv$$

in terms of the initial and boundary conditions using the methods of §4.

We apologize for the fact that the results presented in this section cannot be applied to the case $g(s) = s^{2n+1}$, where n is a positive natural number. This kind of problem and some natural extensions will be the object of a future paper.

REFERENCES

[1] C. O. Horgan and J.K. Knowles, *Recent developments concerning Saint-Venant's Principle*, in Advances in Applied Mechanics 23, J. W. Hutchinson and T. Y. Wu, eds., Academic Press, New York, 1983, pp. 179–269.

[2] C. O. Horgan, *Recent developments concerning Saint-Venant's Principle: An update*, Appl. Mech. Rev., 42 (1989), pp. 295–303.

[3] J. N. Flavin and R. J. Knops., *Some spatial decay estimates in continuum dynamics*, J. Elasticity, 17 (1987), pp. 249–264.

[4] J. N. Flavin, R. J. Knops, and L. E. Payne, *Energy bounds in dynamical problems for a semi-infinite elastic beam*, in Elasticity, Mathematical Methods and Applications, Ellos–Horwood, Chichester, 1989, pp. 101–111.

[5] ———, *Some decay estimates for the constrained elastic cylinder of variable cross section*, Quart. Appl. Math., XLVII (1989), pp. 325–350.

[6] I. Muller, *On the entropy inequality*, Arch. Rational Mech. Anal., 26 (1967), pp. 118–141.

[7] ———, *The coldness: A universal function in thermoelastic bodies*, Arch. Rational Mech. Anal., 41 (1971), pp. 319–332.

[8] A. E. Green and N. Laws, *On the entropy production inequality*, Arch. Rational Mech. Anal., 45 (1972), pp. 47–53.

[9] W. S. Edelstein, *A spatial decay estimate for the heat equation*, J. Appl. Math. Phys., 20 (1969), pp. 900–905.

[10] ———, *Further study of spatial decay estimates for semilinear parabolic equations*, J. Math. Anal. Appl., 35 (1971), pp. 577–590.

[11] J. K. Knowles, *On the spatial decay of solutions of the heat equation*, J. Appl. Math. Phys., 22 (1971), pp. 1050–1056.

[12] C. O. Horgan and L.T. Wheeler, *Spatial decay estimates for the heat equation via the maximum principle*, J. Appl. Math. Phys., 27 (1976), pp. 371–376.

[13] ———, *On maximum principles and spatial decay estimates for heat equations*, in Proc. 12th. Annual Meeting Soc. Engrg. Sci., Austin, TX, 1975, pp. 331–339.

[14] V. G. Sigilito, *On the spatial decay of solutions of parabolic equations*, J. Appl. Math. Phys., 21 (1970), pp. 1078–1081.

[15] C. O. Horgan, L. E. Payne, and L. T. Wheeler, *Spatial decay estimates in transient heat conduction*, Quart. Appl. Mathematics., XLII (1984), pp. 119–127.

[16] C. O. Horgan and L. T. Wheeler, *Spatial decay estimates for pseudoparabolic equations*, Lett. Appl. Engrg. Sci., 3 (1975), pp. 237–243.

[17] J. W. Nunziato, *On the spatial decay of solutions in the nonlinear theory of heat conduction*, J. Math. Anal. Appl., 48 (1974), pp. 687–698.

[18] C. O. Horgan, *Integrals bounds for solutions of nonlinear reaction-diffusion equations*, J. Appl. Math. Phys., 28 (1977), pp. 197–205.

[19] W. E. Fitzgibbon, J. J. Morgan, and L. T. Wheeler, *Spatial decay estimates for reaction diffusion systems*, Quart. Appl. Math., XLVII (1989), pp. 529–538.

[20] S. Breuer and J. J. Roseman, *Spatial decay estimates for nonlinear equations in semi- infinite cylinder*, J. Appl. Math. Phys., 41 (1990), pp. 524–536.

[21] R. Russo, *On generalized Saint Venant' s principle and time asymptotic behavior of solution to a partial differential system of parabolic type in unbounded domains*, Boll. Un. Mat. Ital. B (7), 2 (1988), pp. 729–746.

[22] J. Rauch, *Qualitative behavior of dissipative wave equations on bounded domains*, Arch. Rational. Mech. Anal., 62 (1976), pp. 77–85.

[23] K. A. Lindsay and B. Straughan, *Temperature waves in a rigid heat conductor*, J. Appl. Math. Phys., 27 (1976), pp. 653–661.

[24] J. B. Conway, *Functions of one Complex Variable,* 2nd ed., Graduate Texts in Mathematics, Springer-Verlag, New York, 1978.

[25] C. O. Horgan and L. E. Payne, *Decay estimates for second order quasilinear partial differential equations transient heat conduction*, Adv. Appl. Math., 5 (1984), pp. 309–332.

[26] R. Quintanilla, *Some theorems of Phragmen–Lindeloff type for nonlinear partial differential equations*, Publ. Mat., 37 (1993), pp. 443–463.

[27] C. O. Horgan and L. E. Payne, *Phragmen–Lindeloff type results for harmonic functions with nonlinear boundary conditions*, Arch. Rational Mech. Anal., 122 (1993), pp. 123–144.

[28] D. Iesan and R. Quintanilla, *Decay estimates and energy bounds for porous elastic cylinders*,

J. Appl. Math. Phys., 46 (1995), pp. 268–281.

[29] S. Cox and E. Zuazua, *Estimations sur le taux de decroissance exponentielle de l'energie dans des equations des ondes dissipatives lineaires*, C. R. Acad. Sci. Paris Ser. I. Math., 317 (1993), pp. 249–254.

[30] E. Zuazua, *Exponential decay for the semilinear wave equation with locally distributed damping*, Comm. Partial Differential Equations, 15 (1990), pp. 205–235..

[31] ———, *Exponential decay for the semilinear wave equation with localized damping in unbounded domains*, J. Math. Pures Appl., 70 (1991), pp. 513–529..

[32] E. Feireisl and E. Zuazua, *Global attractores for semilinear equations with locally distributed nonlinear damping and critical exponent*, Comm. Partial Differential Equations, to appear.

[33] E. Zuazua, *Stability and decay for a class of nonlinear hyperbolic problems*, Asymptotic Anal., 1 (1988), pp. 161–185.

[34] E. Feireisl, *Strong decay for wave equations with nonlinear nonmonotone damping*, Nonlinear Anal., 21 (1993), pp. 49–63.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 92–109, January 1996

CONVERGENCE OF THE CHILD–LANGMUIR ASYMPTOTICS OF THE BOLTZMANN EQUATION OF SEMICONDUCTORS*

NAOUFEL BEN ABDALLAH†

Abstract. In a previous paper, the Child–Langmuir asymptotics of the Boltzmann equation of semiconductors were presented, a limit problem was then derived, and its well posedness was analyzed. In this paper, the convergence of the perturbed problem to the limit one is proved. The proof is done in three steps. In the first step, we prove uniform bounds by using supersolution and support estimates. The second step involves combining the supersolution technique and semiexplicit formulas derived from the phase portrait of the Boltzmann equation to improve the regularity results. Finally, the limit equation is integrated and the convergence is proved in the third step.

Key words. Child–Langmuir, Boltzmann, singular perturbation, convergence, support, supersolution, bounded measure

AMS subject classifications. 35B25, 35B45, 35Q95, 35J05, 78A35

1. Introduction and main result.

Various models are used to describe electron transport in a semiconductor. The simplest one is the drift–diffusion model consisting of the charge conservation equation and a phenomenological relation expressing the current as the sum of a drift term due to the electrostatic field and a term related to the density diffusion. Many mathematical results are available for this model (see [13], [14], and references therein) but its validity is not very clear, especially in short devices. Another kind of model is the hydrodynamic model, which includes an equation on the energy and is valid when the mean free path is very small compared to the characteristic length of the device. The main difficulty related to this model is that the physical coefficients appearing in the equations are not well known and are often experiment dependent. In short devices, a kinetic description is necessary to model correctly the transport of particles: the population of charged particles is described not only by the position variable, but also by the velocity. A wide variety of problems is treated with these models [17], [15] and numerical solutions are computed with Monte Carlo techniques [11], [20] or deterministic particle methods [16], [9], [6]. The kinetic models are in fact generalizations of the drift–diffusion and hydrodynamic models; indeed one can derive the drift–diffusion and the hydrodynamic models as asymptotic approximations of the kinetic ones when the mean free path tends to zero (see, for example, [18], [19]).

In the situation we are interested in here, the mean free path is not so small that a complete kinetic model need be used. This presents a serious drawback when numerical resolution is considered since kinetic models are very costy (the dimension is multipied by two) and present oscillations in the interface between lowly doped and highly doped regions. This singular behavior is often due to the existence of different scales in the physical parameters and lead mathematically speaking to singular perturbation problems.

In this paper, we deal with a special singular perturbation problem which arises when the velocity of injected particles is small compared to the velocity they reach

[*] Received by the editors December 9, 1993; accepted for publication (in revised form) May 31, 1994.

[†] Mathèmatiques pour l'Industrie et la Physique, Centre National de la Recherche Scientifique, Unitė Mixte de Recherche 9974, UFR MIG, Universitė Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse cedex, France. This research was performed while the author was a member of Centre de Mathėmatiques Appliquėes, Ecole Polytechnique, France.

after being accelerated by the electric field. These asymptotics are called the Child–Langmuir asymptotics and were first introduced to analyze a vacuum diode under high voltages (see [10], [7], and [8] for mathematical treatment and [12] for physical background). The asymptotics lead to a reduced problem much simpler than the complete one and which contains most of the physics. Our aim here is to extend the analysis to semiconductors where specific issues, namely the collisions, enter into account and complicate seriously the analysis of the problem.

In a previous paper [3], P. Degond and the author introduced a model for electron transport in a unipolar $N^+ - N^- - N^+$ junction at low temperatures taking into account the collisions with the lattice. This model consists of the linear Boltzmann equation of semiconductors (with nondegeneracy assumption), in the parabolic band, and in relaxation time approximations, coupled with the Poisson equation. We introduced a scaling of the equations describing the model and got a singular perturbation problem for the Boltzmann–Poisson equation (the so called Child–Langmuir asymptotics). These asymptotics are intended to model the device behavior when the lattice temperature is *small* or equivalently when the applied voltage is *large.* A limit problem was formally derived [3] and its well posedness was analyzed. However, we did not prove any convergence result of the perturbed system to the limit one. In this paper we fill this gap and prove the convergence. Let us first recall the scaled one-dimensional stationary Vlasov–Poisson–Boltzmann system

$$v\frac{\partial f^\varepsilon}{\partial x} + \frac{1}{2}\frac{d\varphi^\varepsilon}{dx}\frac{\partial f^\varepsilon}{\partial v} = -\frac{1}{\tau}\left(f^\varepsilon - n^\varepsilon \frac{1}{\varepsilon} M_0\left(\frac{v}{\varepsilon}\right)\right), \tag{1.1}$$

$$\frac{d^2\varphi^\varepsilon}{dx^2} = n^\varepsilon(x), \quad x \in [0,1], \tag{1.2}$$

$$n^\varepsilon(x) = \int_{-\infty}^{+\infty} f^\varepsilon(x,v)\,dv, \quad x \in [0,1], \tag{1.3}$$

$$f^\varepsilon(0,v) = g^\varepsilon(v) = \frac{1}{\varepsilon^2} g\left(\frac{v}{\varepsilon}\right), \quad v > 0, \tag{1.4}$$

$$f^\varepsilon(1,v) = 0, \quad v < 0, \tag{1.5}$$

$$\varphi^\varepsilon(0) = 0, \quad \varphi^\varepsilon(1) = 1, \tag{1.6}$$

where

$$M_0(v) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right) \tag{1.7}$$

and

$$g(v) \le C_1 \exp\left(-\frac{v^2}{2}\right). \tag{1.8}$$

In these equations f^ε is the distribution function and φ^ε is the electric potential. We refer to [3] or [5] for the derivation and the scaling of the equations. Here, ε is a small parameter and stands for the temperature of the crystal lattice. Notice that the singularity of this system appears not only in the boundary condition as in the collisionless case, but also in the equations themselves.

Remark 1.1. We notice that in the Poisson equation (1.2), the doping density was neglected. This hypothesis is highly questionable from the phyiscal point of view since the scaled doping density is of the order of 0.5 in the test case we considered (see [1]). The reason we neglect it is mathematical since this assumption guarantees the monotonicity of the limiting potential as we shall see it later. However, we can conjecture that the monotonicity still holds for the small doping profile (this is in fact the case when there is no collision [2]). For physical applications, a reduced model including the doping density was implemented and the results so obtained are in very good agreement with experimental measures (see [4]). But in this case we do not yet have any mathematical result.

Let us now come back to our problem and begin with the following existence theorem proved by F. Poupaud [17].

THEOREM 1.1 (see [17]). *The system* (1.1)–(1.8) *has a solution* $(\varphi^\varepsilon, f^\varepsilon) \in W^{2,\infty}(0,1) \times [L^1 \cap L^\infty(0,1) \times \mathbb{R}]$ *such that*

$$f^\varepsilon(x,v) \leq \frac{C_1}{\varepsilon^2} \exp \frac{-v^2 + \varphi^\varepsilon(x)}{2\varepsilon^2}. \tag{1.9}$$

The main topic of this paper is the analysis of the behavior of the solutions $(f^\varepsilon, \varphi^\varepsilon)$ as ε tends to zero. The limit problem was derived formally in [3] where the limit distribution function f was written in the following form:

$$f(x,v) = n_1(x)\,\delta(v - \sqrt{\varphi(x)}) + \int_0^x \overline{n}_2(x,y)\delta(v - \sqrt{\varphi(x) - \varphi(y)})\,dy. \tag{1.10}$$

The first part of the distribution function represents the so-called ballistic electrons which simply drift along the principal characteristics $v = \sqrt{\varphi(x)}$, whereas the second part stands for electrons which collide with the crystal lattice. The integral representing them can be interpreted as a change of variable. Indeed, if we set $u = \sqrt{\varphi(x) - \varphi(y)}$, then

$$\begin{aligned}\int_0^x \overline{n}_2(x,y)\delta(v - \sqrt{\varphi(x) - \varphi(y)})\,dy &= \int_0^{\sqrt{\varphi(x)}} \overline{n}_2(x,y)\delta(v-u)\,\frac{2udu}{\varphi'(y)} \\ &= \overline{n}_2(x,x^*)\,\frac{2v}{\varphi'(x^*)},\end{aligned}$$

where x^* is given by the equation $\varphi(x^*) = \varphi(x) - v^2$. Now if we replace f by the expression (1.10) in the formal limit of the Boltzmann equation (see [3] for the formal computations), we end up with the following semilinear (and nonlocal) elliptic equation for the electric potential:

$$\frac{d^2\varphi}{dx^2} = n(x), \tag{1.11}$$

$$n(x) = n_1(x) + n_2(x), \tag{1.12}$$

$$n_1(x) = \frac{j}{\sqrt{\varphi(x)}}\, g_\varphi(x,0), \tag{1.13}$$

$$n_2(x) - \int_0^x \frac{g_\varphi(x,y)}{\tau\sqrt{\varphi(x)-\varphi(y)}} n_2(y)\,dy = \int_0^x \frac{g_\varphi(x,y)}{\tau\sqrt{\varphi(x)-\varphi(y)}} n_1(y)\,dy, \tag{1.14}$$

$$\varphi(0) = 0, \qquad \varphi(1) = 1, \tag{1.15}$$

where

$$g_\varphi(x,y) = \exp\left(-\frac{1}{\tau}\int_y^x \frac{dz}{\sqrt{\varphi(z)-\varphi(y)}}\right), \tag{1.16}$$

and

$$n_2(x) = \int_0^x \overline{n}_2(x,y)dy, \quad \overline{n}_2(x,y) = \frac{n(y)}{\tau}\frac{g_\varphi(x,y)}{\sqrt{\varphi(x)-\varphi(y)}}. \tag{1.17}$$

Notice that (1.10) can be rewritten using the above expression of $\overline{n}_2$

$$f(x,v) = n_1(x)\,\delta(v-\sqrt{\varphi(x)}) + \frac{2n(x^*)}{\varphi'(x^*)} g_\varphi(x,x^*), \quad \varphi(x^*) = \varphi(x) - v^2. \tag{1.18}$$

We proved [3] the following results for the limit problem. They state that the limit problem has a unique Child–Langmuir current (corresponding to a vanishing electric field at $x = 0$) and no solution if the current j is large.

THEOREM 1.2 (see [3]). *There exist $\tau_1 \geq 7/9$ and $\tau_2 \leq 4/5$ such that for every $\tau \in (0,\tau_1] \cup [\tau_2,\infty)$, there exists a unique value $j = j_{CL}(\tau)$ such that the problem* (1.11)–(1.16) *has a unique solution φ satisfying $d\varphi/dx(0) = 0$. Moreover, $j_{CL}(\tau) \sim 4/9$ when τ tends to ∞ and $j_{CL}(\tau) \sim \tau 9/16$ when τ tends to zero.*

THEOREM 1.3 (see [3]). *There exists a value $j_{\max}(\tau)$ such that the system* (1.11)–(1.16) *has no solution for $j > j_{\max}(\tau)$. This value satisfies the following estimate:*

$$j_{CL}(\tau) \leq j_{\max}(\tau) < \min\left(\frac{4}{9}, \frac{9}{16}\tau\right).$$

Since the derivation of the limit problem was purely formal, we did not prove any convergence theorem in [3]. The aim of this paper is then to prove the convergence of the perturbed problem toward the above limit problem when ε tends to zero. Before stating the main theorem of this paper, we introduce some notation which will be used throughout the paper. We define, respectively, the injected current, the total current, the kinetic energy, and the total energy by

$$j_g = \int_0^{+\infty} vg(v)dv, \tag{1.19}$$

$$j^\varepsilon(x) = \int_{-\infty}^{+\infty} vf^\varepsilon(x,v)dv, \tag{1.20}$$

$$k^\varepsilon(x) = \int_{-\infty}^{+\infty} v^2 f^\varepsilon(x,v)dv, \tag{1.21}$$

$$h^\varepsilon(x) = k^\varepsilon(x) - \frac{1}{4}\left(\frac{d\varphi^\varepsilon}{dx}(x)\right)^2 + \frac{j^\varepsilon x}{\tau}. \tag{1.22}$$

Now we are able to give the main result of this paper.

THEOREM 1.4. *Let τ be in $(0,\tau_1] \cup [\tau_2,+\infty)$.*

(i) *Suppose that $j_g > j_{\max}(\tau)$. Then, the solutions $(f^\varepsilon,\varphi^\varepsilon)$ defined in Theorem 1.1 converge as ε goes to zero to (f,φ) where f is given by* (1.10) *and φ is the unique solution of* (1.11)–(1.16) *corresponding to $j = j_{CL}(\tau)$.*

(ii) *Suppose that $j_g \leq j_{CL}(\tau)$, then there exists a subsequence $(f^\varepsilon,\varphi^\varepsilon)$ converging as ε goes to zero to (f,φ) where f is given by* (1.10) *and φ is a solution of* (1.11)–(1.16) *corresponding to $j = j_g$.*

(iii) *In the case $j_g \in (j_{CL}, j_{\max})$, the convergence holds for a subsequence but the limit current is not fully determined. However, we have $j \in \{j_g, j_{CL}\}$.*

The convergence holds in the $C^1([0,1])$ strong topology for φ^ε and in the weak $*$ topology of bounded measures for f^ε.

The scheme of the proof follows closely, at the beginning, the previous works on collisionless Child–Langmuir asymptotics [10], [7], [8], [2]. Using the supersolution technique and the maximum principle, we first prove some energy identities and deduce from these identities and support estimates the boundedness of the L^1 norm of the density. This implies the weak convergence of the distribution function f^ε in the bounded measure topology.

Besides, the nonlinear term in the weak formulation of the Boltzmann equation reads

$$\frac{1}{2}\int \frac{d\varphi^\varepsilon}{dx} f^\varepsilon \psi \, dx dv,$$

where ψ is a test function. Since f^ε tends weakly toward a bounded measure, the electric potential has to converge strongly in the C^1 topology, otherwise the nonlinear term does not pass to the limit. Hence, the L^1 bound on the density is not sufficient since the maximum regularity of the electric potential allowed by the L^1 estimate on the density is $W^{1,\infty}$.

In the collisionless case, this difficulty was solved by using the energy invariance which implies that the electric field φ' can be expressed by means of the kinetic energy k. Indeed, in the collisionless case, the limiting distribution function is monokinetic which allows us to express in a simple way the kinetic energy by means of the current. With this expression, one can prove the uniform convergence of the kinetic energy which in turn implies the C^1 strong convergence of the electric potential. Unfortunately, in the collisional case, the distribution function is no longer monokinetic because of the collisions. Therefore there is no hope for the kinetic energy to be expressed in a simple way and the energy invariance does not help to prove the C^1 convergence of the potential.

An alternative approach lies in the use of the Poisson equation. We first prove that the density is bounded in $L^\infty[\alpha,1]$ for every $\alpha > 0$. We deduce from this (using the Poisson equation), that the potential converges in $C^1[\alpha,1]$ for every $\alpha > 0$ and using the convexity of the potential we prove that the C^1 convergence actually holds on the whole interval $[0,1]$. The main difficulty in proving the L^∞ bound on the density is that the supersolution introduced in Theorem 1.1 is almost of no use since it is much larger than the solution in the regions where we intended to use it. The solution we propose is to combine the supersolution bound with semiexplicit formulas derived from the knowledge of the phase portrait of the Boltzmann equation.

Afterward, we pass to the limit in the Boltzmann equation and determine the structure of the limit problem by integrating the equations. The final step in the proof is the knowledge of the current and is achieved by using the results of [3].

The outline of the paper is as follows. In §2, we first establish uniform bounds on the solution of the perturbed problem using energy invariants and supersolution bounds. Then we give some qualitative properties satisfied by the limit solution. In a third part we prove, using these properties, the C^1 strong convergence of the potential. In §3, we pass to the limit in the equations and prove the main theorem. In §4, we summarize our results.

2. Regularity and support estimates. This section is divided into three parts. In the first part, we prove some preliminary estimates following essentially from the supersolution exhibited in Theorem 1.1. From these estimates, we deduce the convergence of the solution by a compactness argument. In the second part, we give some information on the structure of the limiting problem: support of the distribution function, positivity of the current and of the limit potential. Finally, in the third part we prove the C^1 convergence of the potential.

2.1. Preliminary estimates. We begin with an immediate result obtained by taking the 0th and the first moments of the Boltzmann equation with respect to v.

LEMMA 2.1. (i) *The current* $j^\varepsilon(x)$ *does not depend on* x *and* $0 \le j^\varepsilon \le j_g$.

(ii) *The total energy* h^ε *does not depend on* x.

LEMMA 2.2. *For* ε *small enough, we have the following:*

(i) $8\varepsilon^2 \ln\varepsilon \le \varphi^\varepsilon(x) \le 1$.

(ii) $(j^\varepsilon - C\varepsilon^5)/\sqrt{\varphi^\varepsilon(x) - 10\varepsilon^2\ln\varepsilon} \le n^\varepsilon(x) \le \frac{C}{\varepsilon}\exp(\frac{\varphi^\varepsilon(x)}{2\varepsilon^2})$.

(iii) $-C\sqrt{-\varepsilon\ln\varepsilon} \le {\varphi^\varepsilon}'(0) \le 1$.

(iv) *There exits* $C > 0$ *such that the following implications hold:*

$$\begin{aligned} \varphi^\varepsilon(x) \ge -10\varepsilon^2\ln\varepsilon \quad &\Longrightarrow \quad ({\varphi^\varepsilon}'(x))^2 \ge C(j^\varepsilon - C\varepsilon^5)\sqrt{\varphi^\varepsilon(x)}, \\ \varphi^\varepsilon(x) \ge \varepsilon^2 \quad &\Longrightarrow \quad ({\varphi^\varepsilon}'(x))^2 \ge C(j^\varepsilon - C\varepsilon^5)\sqrt{\frac{\varphi^\varepsilon(x)}{-\ln\varepsilon}}. \end{aligned}$$

Proof. (i) The integration of (1.9) with respect to v gives

$$0 \le n^\varepsilon(x) \le \frac{C}{\varepsilon}\exp\left(\frac{\varphi^\varepsilon(x)}{2\varepsilon^2}\right).$$

Therefore φ^ε is a supersolution of the following semilinear elliptic equation

$$\left\{\begin{array}{l} -\psi'' + \dfrac{C}{\varepsilon}\exp\left(\dfrac{\psi}{2\varepsilon^2}\right) = 0, \\ \psi(0) = 0, \quad \psi(1) = 1. \end{array}\right. \tag{2.23}$$

It is easy to check that the function

$$\psi^\varepsilon(x) = 8\varepsilon^2\ln\varepsilon + \varepsilon^2x^2$$

is a subsolution of this equation and thus $\varphi^\varepsilon(x) \ge \psi^\varepsilon(x)$; the upper bound on φ^ε is a direct consequence of the positivity of n^ε.

(ii) We have

$$\begin{aligned} n^\varepsilon(x) &\ge \frac{1}{\sqrt{\varphi^\varepsilon(x) - 10\varepsilon^2\ln\varepsilon}} \int_0^{\sqrt{\varphi^\varepsilon(x) - 10\varepsilon^2\ln\varepsilon}} vf^\varepsilon(x,v)dv \\ &\ge \frac{1}{\sqrt{\varphi^\varepsilon(x) - 10\varepsilon^2\ln\varepsilon}}\left[j^\varepsilon - \int_{\sqrt{\varphi^\varepsilon(x)-10\varepsilon^2\ln\varepsilon}}^{+\infty} vf^\varepsilon(x,v)dv\right]. \end{aligned}$$

We deduce from (1.9) that

$$\int_{\sqrt{\varphi^\varepsilon(x)-10\varepsilon^2\ln\varepsilon}}^{+\infty} vf^\varepsilon(x,v)dv \leq C\varepsilon^5,$$

which ends the proof of (ii).

(iii) The second inequality is easily deduced from the convexity of φ^ε. For the first inequality, we suppose that ${\varphi^\varepsilon}'(0) < 0$. Then, there exists $r^\varepsilon > 0$ such that φ^ε is negative on $[0, r^\varepsilon]$. Therefore $n^\varepsilon \leq \frac{C}{\varepsilon}$ on $[0, r^\varepsilon]$. This implies that

$${\varphi^\varepsilon}'(x) \leq {\varphi^\varepsilon}'(0) + \frac{C}{\varepsilon}x$$

and

$$\varphi^\varepsilon(x) \leq {\varphi^\varepsilon}'(0)x + \frac{C}{2\varepsilon}x^2.$$

Let $x_\varepsilon = -\frac{\varepsilon}{C}{\varphi^\varepsilon}'(0)$, then $x^\varepsilon < r^\varepsilon$ and

$$\varphi^\varepsilon(x_\varepsilon) \leq -\frac{\varepsilon}{2C}({\varphi^\varepsilon}'(0))^2.$$

The result follows easily since $\varphi^\varepsilon \geq 10\varepsilon^2\ln\varepsilon$.

(iv) We first remark that ${\varphi^\varepsilon}'$ is positive on the set $\varphi^\varepsilon > 0$ (because φ^ε is convex and is equal to zero at $x = 0$). Let x_0 be the largest point where φ^ε vanishes. Then the set $\{\varphi^\varepsilon > 0\}$ is equal to $(x_0, 1]$. We consider now $x > x_0$, we multiply the first inequality of (ii) by ${\varphi^\varepsilon}'$, and integrate it between x_0 and x. We end up with the following inequality:

$$({\varphi^\varepsilon}')^2(x) \geq 4(j^\varepsilon - C\varepsilon^5)(\sqrt{\varphi^\varepsilon(x) - 10\varepsilon^2\ln\varepsilon} - \sqrt{-10\varepsilon^2\ln\varepsilon}), \quad x > x_0.$$

This yields the results of (iv). □

PROPOSITION 2.1. *We have the following uniform estimate*

$$|h^\varepsilon| + ||k^\varepsilon||_{L^\infty} + ||\varphi^\varepsilon||_{W^{1,\infty}\cap W^{2,1}} + ||n^\varepsilon||_{L^1} \leq C.$$

Proof. First, we have

$$h^\varepsilon = k^\varepsilon(0) - \frac{1}{4}\left(\frac{d\varphi^\varepsilon}{dx}(0)\right)^2;$$

then since the integration of (1.9) gives

$$k^\varepsilon(x) \leq C\varepsilon\exp\left(\frac{\varphi^\varepsilon(x)}{2\varepsilon^2}\right), \tag{2.24}$$

we deduce from Lemma 2.2(iii) that h^ε is bounded. Assume for the moment that $k^\varepsilon(1)$ is bounded. Then we deduce from the definition of h^ε that ${\varphi^\varepsilon}'(1)$ is bounded. Now since φ^ε is convex, this leads to the boundedness of $||{\varphi^\varepsilon}'||_{L^\infty}$. Using h^ε once more, we deduce that $||k^\varepsilon||_{L^\infty}$ is bounded. Additionally, since $||n^\varepsilon||_{L^1} = {\varphi^\varepsilon}'(1) - {\varphi^\varepsilon}'(0)$, we conclude that $||n^\varepsilon||_{L^1}$ is bounded.

Now we prove that $k^\varepsilon(1)$ is bounded. Since $f^\varepsilon(1,v)$ vanishes for negative v's, then

$$\begin{aligned} k^\varepsilon(1) &= \int_0^{+\infty} v^2 f^\varepsilon(1,v)dv \\ &= \int_0^2 v^2 f^\varepsilon(1,v)dv + \int_2^{+\infty} v^2 f^\varepsilon(1,v)dv \\ &\leq 2j^\varepsilon + \frac{C_1}{\varepsilon^2}\int_2^{+\infty} v^2 \exp\left(\frac{-v^2+1}{2\varepsilon^2}\right) dv. \end{aligned}$$

Hence

$$k^\varepsilon(1) \leq 2j^\varepsilon + O(\varepsilon), \tag{2.25}$$

and the result holds since j^ε is bounded. □

At this stage we can conclude that for a subsequence we have

$$\begin{array}{ll} f^\varepsilon \rightharpoonup f & \mathcal{M}_b([0,1]\times \mathbb{R}) \text{ weak } * \\ n^\varepsilon \rightharpoonup n & \mathcal{M}_b([0,1]) \text{ weak } * \\ \varphi^\varepsilon \rightarrow \varphi & W^{1,p}([0,1]) \text{ strong } \forall 1<p<\infty \\ \varphi^\varepsilon \rightharpoonup \varphi & W^{1,\infty}([0,1]) \text{ weak } * \\ k^\varepsilon \rightharpoonup k & L^\infty([0,1]) \text{ weak } * \\ j^\varepsilon \rightarrow j & \\ h^\varepsilon \rightarrow h & \end{array}$$

and we have obviously

$$n = \langle f, 1\rangle_{\mathcal{M}_b,C_0}, \quad j = \langle f, v\rangle_{\mathcal{M}_b,C_0}, \quad k = \langle f, v^2\rangle_{\mathcal{M}_b,C_0}$$

and

$$h = k(x) - \frac{1}{4}\left(\frac{d\varphi}{dx}(x)\right)^2 + \frac{jx}{\tau}, \quad \varphi \geq 0,$$

where $\langle\cdot,\cdot\rangle_{\mathcal{M}_b,C_0}$ denotes the duality product between the bounded measures and the continuous compactly supported functions (the duality is taken with respect to the velocity variable only). Of course, the functions $1, v, v^2$ are not compactly supported, but the duality product still holds because thanks to (1.9) the support of f is compact ($f = 0$ for $v^2 > \varphi(x)$).

2.2. The limiting solution: Qualitative structure. In this section we prove that the limiting current cannot vanish. We deduce from this that the limiting electric potential vanishes only at $x = 0$. Finally, using this result we characterize completely the support of the limiting distribution function. We begin with the following lemma.

LEMMA 2.3. *Let us assume that φ^ε converges uniformly to φ in the $C^1([0,1])$ strong topology and that $\varphi'(0) > 0$. Then the limit current j is equal to the injected current j_g.*

Proof. Since the convergence holds in $C^1([0,1])$, then for ε small enough, the function $\varphi^\varepsilon(x) \geq \alpha x$, with $\alpha > 0$. Since φ^ε is positive, then the characteristics passing through a point (x,v) with $v < 0$ are issued from the boundary $x = 1$. Hence, writing the Boltzmann equation

$$v\frac{\partial f^\varepsilon}{\partial x} + \frac{1}{2}\frac{d\varphi^\varepsilon}{dx}\frac{\partial f^\varepsilon}{\partial v} + \frac{1}{\tau}f^\varepsilon = n^\varepsilon \frac{1}{\tau\varepsilon}M_0\left(\frac{v}{\varepsilon}\right),$$

we can express $f^\varepsilon(x,v)$ explicitly in terms of the density

$$f^\varepsilon(x,v)=\int_x^1 \frac{n^\varepsilon(y)}{\varepsilon\tau\sqrt{2\pi}}\frac{\exp(-\frac{v^2+\varphi^\varepsilon(y)-\varphi^\varepsilon(x)}{2\varepsilon^2})}{\sqrt{v^2+\varphi^\varepsilon(y)-\varphi^\varepsilon(x)}}\exp\left(\frac{-1}{\tau}\int_x^y\frac{dz}{\sqrt{v^2+\varphi^\varepsilon(z)-\varphi^\varepsilon(x)}}\right)dy$$

for v negative. Hence

$$\begin{aligned}|j^\varepsilon-j_g|&=\left|\int_{-\infty}^0 vf^\varepsilon(0,v)dv\right|\\ &\le C\int_0^{+\infty}\int_0^1 n^\varepsilon(y)\frac{v}{\varepsilon}\frac{\exp(-\frac{v^2+\varphi^\varepsilon(y)}{2\varepsilon^2})}{\sqrt{v^2+\varphi^\varepsilon(y)}}dy\,dv\\ &\le C\varepsilon\int_0^1 n^\varepsilon(y)\frac{\exp(-\frac{\varphi^\varepsilon(y)}{2\varepsilon^2})}{\sqrt{\varphi^\varepsilon(y)}}\,dy. \end{aligned}\tag{2.26}$$

We split the integral of the right-hand side into two integrals, the first from zero to ε and the second from ε to one. To estimate the first integral, we use Lemma 2.2(ii); this yields

$$C\varepsilon\int_0^\varepsilon n^\varepsilon(y)\frac{\exp(-\frac{\varphi^\varepsilon(y)}{2\varepsilon^2})}{\sqrt{\varphi^\varepsilon(y)}}\,dy\le C\int_0^\varepsilon\frac{dy}{\sqrt{\varphi^\varepsilon(y)}},$$

and using the estimate $\varphi^\varepsilon(y)\ge\alpha y$, we obtain

$$C\varepsilon\int_0^\varepsilon n^\varepsilon(y)\frac{\exp(-\frac{\varphi^\varepsilon(y)}{2\varepsilon^2})}{\sqrt{\varphi^\varepsilon(y)}}\,dy\le C\sqrt{\varepsilon}.\tag{2.27}$$

For the second integral, we recall that φ^ε is increasing so that

$$C\varepsilon\int_\varepsilon^1 n^\varepsilon(y)\frac{\exp(-\frac{\varphi^\varepsilon(y)}{2\varepsilon^2})}{\sqrt{\varphi^\varepsilon(y)}}\,dy\le C\varepsilon\frac{\exp(-\frac{\varphi^\varepsilon(\varepsilon)}{2\varepsilon^2})}{\sqrt{\varphi^\varepsilon(\varepsilon)}}||n^\varepsilon||_{L^1}$$

and taking once more into account the estimate $\varphi^\varepsilon(y)\ge\alpha y$, we have

$$C\varepsilon\int_\varepsilon^1 n^\varepsilon(y)\frac{\exp(-\frac{\varphi^\varepsilon(y)}{2\varepsilon^2})}{\sqrt{\varphi^\varepsilon(y)}}\,dy\le C\sqrt{\varepsilon}\exp\left(-\frac{\alpha}{2\varepsilon}\right).\tag{2.28}$$

In view of (2.26), inequalities (2.27) and (2.28) lead to $j=j_g$. □

The following proposition is a direct consequence of this lemma.

PROPOSITION 2.2. (i) *The limit current j is strictly positive.*

(ii) *The limit potential φ is strictly positive on* $(0,1]$.

Proof. (i) Assume that the current j^ε tends to zero. Then we deduce from (2.24) and (2.25) that $\lim k^\varepsilon(0)=\lim k^\varepsilon(1)=0$. Therefore, (1.22) implies

$$\lim\left(\frac{d\varphi^\varepsilon}{dx}(0)\right)^2=\lim\left(\frac{d\varphi^\varepsilon}{dx}(1)\right)^2=-4h.$$

Since φ^ε is convex, we deduce that $\lim(\varphi^{\varepsilon\prime})^2(1)=-4h\ge 1$. Hence, Lemma 2.2(iii) implies that $\varphi^{\varepsilon\prime}(0)\ge 0$ for ε small enough. Therefore

$$\lim\frac{d\varphi^\varepsilon}{dx}(0)=\lim\frac{d\varphi^\varepsilon}{dx}(1),$$

which implies, since φ^ε is convex, that $d\varphi^\varepsilon/dx$ converges uniformly to a constant on $[0,1]$. We deduce finally that $\varphi^\varepsilon(x)$ converges to x in $C^1([0,1])$, therefore the preceding lemma applies and gives $j = j_g$, which obviously contradicts the above hypothesis.

(ii) Assume that there exists $r > 0$ such that $\varphi(r) = 0$. Then, since φ is convex and nonnegative, it vanishes on the whole interval $[0,r]$. Using (i) and Lemma 2.2 (ii), we claim that the following estimate holds for ε small enough:

$$n^\varepsilon(x) \geq \frac{j}{2\sqrt{\varphi^\varepsilon(x) - 10\varepsilon^2 \ln \varepsilon}}.$$

Since φ^ε converges uniformly toward φ, we deduce from the last estimate that

$$\lim_{\varepsilon\to 0} \inf_{[0,r]} n^\varepsilon = +\infty,$$

which contradicts the boundedness of $||n^\varepsilon||_{L^1}$. □

The following lemma characterizes completely the support of the limit distribution function and shows that the first moment of f^ε is bounded.

LEMMA 2.4. (i) *The sequence* $\int_{-\infty}^{+\infty} |v| f^\varepsilon(x,v)dv$ *is bounded in* $L^\infty([0,1])$.

(ii) *The limit distribution function f is supported by the set*

$$\left\{(x,v),\ 0 \leq v \leq \sqrt{\varphi(x)}\right\}.$$

(iii) *There exists a constant C independent of ε such that*

$$k^\varepsilon \leq C\sqrt{|\varphi^\varepsilon|} + o(1).$$

Proof. (i) Let $\theta_\alpha(v)$ be a decreasing C^1 function such that $\theta_\alpha(v) = 0$ for $v \geq 0$ and $\theta_\alpha(v) = 1$ for $v \leq -\alpha$ with $\alpha > 0$ and $0 \leq \theta_\alpha \leq 1$. Then, since

$$v\frac{\partial f^\varepsilon}{\partial x} + \frac{1}{2}\frac{d\varphi^\varepsilon}{dx}\frac{\partial f^\varepsilon}{\partial v} + \frac{f^\varepsilon}{\tau} = \frac{n^\varepsilon}{\tau\varepsilon}M_0\left(\frac{v}{\varepsilon}\right),$$

we have

$$\frac{d}{dx}\int v\theta_\alpha(v)f^\varepsilon dv - \frac{1}{2}\frac{d\varphi^\varepsilon}{dx}\int \theta'_\alpha(v)f^\varepsilon(v)dv + \frac{\int \theta_\alpha(v)f^\varepsilon(v)dv}{\tau} = \frac{n^\varepsilon}{\tau}\int M_0(v)\theta_\alpha(\varepsilon v)dv.$$

Let $x \geq r^\varepsilon$, where r^ε is defined by

$$\varphi^\varepsilon(r^\varepsilon) = \varepsilon^2. \tag{2.29}$$

Then since ${\varphi^\varepsilon}'(x) > 0$ on $[r^\varepsilon, 1]$, we deduce from the above formula that

$$\frac{d}{dx}\int v\theta_\alpha(v)f^\varepsilon dv \leq \frac{n^\varepsilon}{\tau}\int M_0(v)\theta_\alpha(\varepsilon v)dv, \quad x \geq r_\varepsilon.$$

Integrating this inequality between x and 1, we obtain

$$\frac{-||n^\varepsilon||_{L^1}}{\tau}\int M_0(v)\theta_\alpha(\varepsilon v)dv \leq \int v\theta_\alpha(v)f^\varepsilon(x,v)dv \leq 0, \quad x \in [r^\varepsilon, 1]. \tag{2.30}$$

Letting α tend to zero implies that $\sup_{x\in[r^\varepsilon,1]} |\int_{-\infty}^0 vf^\varepsilon(x,v)dv|$ is bounded and since the total current is bounded, this leads to the boundedness of

$$\sup_{x\in[r^\varepsilon,1]}\int_{-\infty}^{+\infty} |v| f^\varepsilon(x,v)dv.$$

On $[0, r^\varepsilon]$ the boundedness can be proved by the use of (1.9), since $\varphi^\varepsilon \le \varepsilon^2$ on $[0, r^\varepsilon]$.

(ii) It is readily seen that the estimate (1.9) implies

$$\operatorname{supp} f \subset \left\{(x,v),\ v^2 \le \varphi(x)\right\}.$$

Let us now show that $f(x,v) = 0$ for every $v < 0$. For $x = 0$ there is nothing to show. For $x > 0$, the estimate (2.30) holds for every fixed α and for ε small enough. Since, for $\alpha > 0$ the integral

$$\int M_0(v)\theta_\alpha(\varepsilon v)dv$$

is exponentially small, then we pass to the limit (with a fixed α) and get

$$\langle f, v\theta_\alpha(v)\rangle_{\mathcal{M}_b, C_0} = 0, \quad \forall \alpha > 0,$$

which implies that $f(x,v) = 0$ if $v < 0$.

The proof of (iii) is a consequence of (1.9) and (i). Indeed,

$$k_\varepsilon(x) = \int_{|v| \le \sqrt{\varphi(x) - 10\varepsilon^2 \ln \varepsilon}} v^2 f^\varepsilon dv + \int_{|v| \ge \sqrt{\varphi(x) - 10\varepsilon^2 \ln \varepsilon}} v^2 f^\varepsilon dv.$$

The first integral is bounded by

$$\sqrt{\varphi(x) - 10\varepsilon^2 \ln \varepsilon} \int |v| f^\varepsilon dv$$

and the second one is small in view of (1.9). □

2.3. Showing the C^1 strong convergence. Up to now, the results we obtained in §2.1 needed essentially the estimate (1.9) and the procedure used to obtain them is an adaptation of [7] (the collisionless case). In the collisionless case, the support of the limit distribution function is the set $v = \sqrt{\varphi(x)}$. This allows us to express the kinetic energy by means of the current and to obtain the C^1 regularity result for the potential. In fact, this regularity result could be proven (still in the collisionless case) by expressing the density by means of the current which yields that n^ε is bounded in L^∞ far from the point where φ vanishes. This is not straightforward when the collisions are taken into account, because the Maxwellian M_ε tends to a delta measure which implies that the distribution function does not vanish on the axis $v = 0$. Besides, the estimate (1.9) is very bad when $v^2 < \varphi(x)$ (the support of f is a part of this region) and this is a serious limitation of the supersolution method which, alone, does not allow us to prove a sharper estimate on the density. We shall show this estimate by using semiexplicit formulas for the distribution function. This is stated in the following proposition, for which the proof is somewhat lengthy and could be skipped in a first reading.

PROPOSITION 2.3. (i) *The function* $n^\varepsilon\sqrt{|\varphi^\varepsilon|}$ *is bounded in* $L^\infty([0,1])$.

(ii) *The distribution function* f^ε *is bounded in* $L^\infty_{loc}([0,1] \times \mathrm{IR} - \{v = \sqrt{\varphi(x)}\})$.

Proof. (i) Let r^ε be defined in (2.29). Then since $\varphi^\varepsilon \le \varepsilon^2$ on $[0, r^\varepsilon]$, the estimate (1.9) gives

$$\sup_{[0,r^\varepsilon]} n^\varepsilon \sqrt{|\varphi^\varepsilon|} \le C.$$

Hence, we focus on the interval $[r^\varepsilon, 1]$. We first rewrite the Boltzmann equation in the following form:

$$v\frac{\partial f^\varepsilon}{\partial x} + \frac{1}{2}\frac{d\varphi^\varepsilon}{dx}\frac{\partial f^\varepsilon}{\partial v} + \frac{f^\varepsilon}{\tau} = \frac{n^\varepsilon(x)}{\tau\,\varepsilon}M_0\left(\frac{v}{\varepsilon}\right),$$

where the right-hand side is considered as a source term. The integration of this equation over the characteristics gives, for $(x,v) \in [r^\varepsilon, 1] \times \mathbb{R}_-$,

(2.31)

$$f^\varepsilon(x,v) = \int_x^1 \frac{n^\varepsilon(y)}{\varepsilon\tau\sqrt{2\pi}}\frac{\exp(-\frac{v^2+\varphi^\varepsilon(y)-\varphi^\varepsilon(x)}{2\varepsilon^2})}{\sqrt{v^2+\varphi^\varepsilon(y)-\varphi^\varepsilon(x)}}\exp\left(\frac{-1}{\tau}\int_x^y \frac{dz}{\sqrt{v^2+\varphi^\varepsilon(z)-\varphi^\varepsilon(x)}}\right)dy$$

and, for $(x,v) \in [r^\varepsilon, 1] \times [0, \sqrt{\varphi^\varepsilon})$,

(2.32)

$$f^\varepsilon(x,v) = f(x^*,0)\exp\left(\frac{-1}{\tau}\int_{x^*}^x \frac{dz}{\sqrt{v^2+\varphi^\varepsilon(z)-\varphi^\varepsilon(x)}}\right)$$
$$+\int_{x^*}^x \frac{n^\varepsilon(y)}{\varepsilon\tau\sqrt{2\pi}}\frac{\exp(-\frac{\varphi^\varepsilon(y)-\varphi^\varepsilon(x^*)}{2\varepsilon^2})}{\sqrt{\varphi^\varepsilon(y)-\varphi^\varepsilon(x^*)}}\exp\left(\frac{-1}{\tau}\int_y^x \frac{dz}{\sqrt{\varphi^\varepsilon(z)-\varphi^\varepsilon(x^*)}}\right)dy,$$

where x^* is uniquely defined by

$$\varphi^\varepsilon(x^*) = \varphi^\varepsilon(x) - v^2. \tag{2.33}$$

These relations follow from the integration of the Boltzmann equation over the characteristics and can be understood by viewing Fig. 1 which represents the phase portrait.

Estimating the distribution function. Let us now estimate (2.31). For v negative, we first remark that

$$\int_x^1 \frac{\varphi^{\varepsilon\prime}(y)\exp(-\frac{\varphi^\varepsilon(y)-\varphi^\varepsilon(x)}{2\varepsilon^2})}{\varepsilon\sqrt{\varphi^\varepsilon(y)-\varphi^\varepsilon(x)}}dy = \int_0^{\frac{\sqrt{1-\varphi^\varepsilon(x)}}{\varepsilon}}\exp\left(-\frac{u^2}{2}\right)du \le C. \tag{2.34}$$

We then remark that $\sqrt{\varphi^\varepsilon}\varphi^{\varepsilon\prime}$ is increasing on $(r^\varepsilon, 1]$ so that (2.31) leads to

$$f^\varepsilon(x,v) \le \frac{Ce^{-\frac{v^2}{2\varepsilon^2}}}{\varepsilon\sqrt{\varphi^\varepsilon(x)}\varphi^{\varepsilon\prime}(x)}\int_x^1 \frac{n^\varepsilon(y)\sqrt{\varphi^\varepsilon(y)}\varphi^{\varepsilon\prime}(y)\exp(-\frac{\varphi^\varepsilon(y)-\varphi^\varepsilon(x)}{2\varepsilon^2})}{\sqrt{\varphi^\varepsilon(y)-\varphi^\varepsilon(x)}}dy.$$

The term $\sqrt{\varphi^\varepsilon}\varphi^{\varepsilon\prime}$ is introduced in the integral to make the quantity $||n^\varepsilon\sqrt{|\varphi^\varepsilon|}||_{L^\infty}$ appear and to use the estimate (2.34). This leads to the following estimate:

$$f^\varepsilon(x,v) \le \frac{Ce^{-\frac{v^2}{2\varepsilon^2}}||n^\varepsilon\sqrt{|\varphi^\varepsilon|}||_{L^\infty}}{\sqrt{\varphi^\varepsilon(x)}\varphi^{\varepsilon\prime}(x)}.$$

Moreover, since $j > 0$, we deduce from Lemma 2.2(iv) that for ε small enough, we have

$$f^\varepsilon(x,v) \le \frac{Ce^{-\frac{v^2}{2\varepsilon^2}}||n^\varepsilon\sqrt{|\varphi^\varepsilon|}||_{L^\infty}}{\sqrt{\varphi^\varepsilon(x)}(\varphi^\varepsilon(x))^{1/4}}, \quad \varphi^\varepsilon(x) \ge -10\varepsilon^2\ln\varepsilon, v \le 0 \tag{2.35}$$

$$\le \frac{Ce^{-\frac{v^2}{2\varepsilon^2}}||n^\varepsilon\sqrt{|\varphi^\varepsilon|}||_{L^\infty}}{\sqrt{\varphi^\varepsilon(x)}(\varphi^\varepsilon(x))^{1/4}}(-\ln\varepsilon)^{1/4}, \quad \varepsilon^2 \le \varphi^\varepsilon(x), v \le 0. \tag{2.36}$$

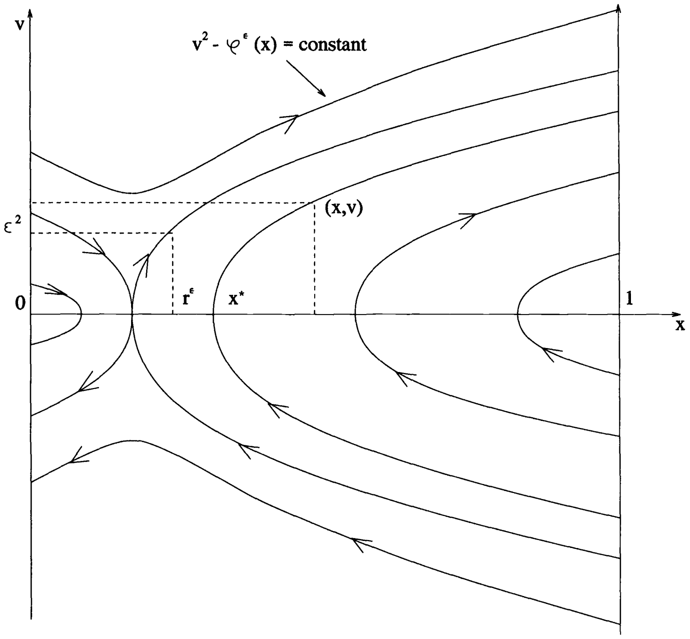


FIG. 1. *Phase portrait.*

For v positive, we obtain, in view of (2.32), the same estimates as above with v replaced by zero and x by x^*. Hence

$$f^\varepsilon(x,v) \le \frac{C||n^\varepsilon\sqrt{|\varphi^\varepsilon|}||_{L^\infty}}{\sqrt{\varphi^\varepsilon(x^*)}(\varphi^\varepsilon(x^*))^{1/4}}, \quad \varphi^\varepsilon(x^*) \ge -10\varepsilon^2 \ln\varepsilon, v \in [0, \sqrt{\varphi^\varepsilon(x)}) \tag{2.37}$$

$$\le \frac{C||n^\varepsilon\sqrt{|\varphi^\varepsilon|}||_{L^\infty}}{\sqrt{\varphi^\varepsilon(x^*)}(\varphi^\varepsilon(x^*))^{1/4}}(-\ln\varepsilon)^{1/4}, \quad \varepsilon^2 \le \varphi^\varepsilon(x), v \in [0, \sqrt{\varphi^\varepsilon(x)}). \tag{2.38}$$

Estimating the density. Let us now estimate $n_-^\varepsilon(x) = \int_{-\infty}^0 f^\varepsilon(x,v)dv$. To this end, we do not need the sharp estimate (2.35). Estimate (2.36) is sufficient and gives after integration on $\mathbb{R}_-$

$$\sqrt{\varphi^\varepsilon(x)}n_-^\varepsilon(x) \le C\frac{\varepsilon(-\ln\varepsilon)^{1/4}}{(\varphi^\varepsilon(x))^{1/4}}||n^\varepsilon\sqrt{|\varphi^\varepsilon|}||_{L^\infty} \le C\varepsilon^{1/2}(-\ln\varepsilon)^{1/4}||n^\varepsilon\sqrt{|\varphi^\varepsilon|}||_{L^\infty}. \tag{2.39}$$

For positive v's, more care should be taken since $\varphi(x^*) = 0$ if $v = \sqrt{\varphi(x)}$. Then let $k \ge 2$ be a constant (to be fixed later) and $0 \le v \le \sqrt{\varphi^\varepsilon(x)}/k$. Then $\varphi^\varepsilon(x^*) \ge \frac{3}{4}\varphi^\varepsilon(x)$ and therefore, we can replace x^* by x in (2.38) and (2.37) (modulo a change in C). By integration this leads to

$$\sqrt{\varphi^\varepsilon(x)}\int_0^{\frac{\sqrt{\varphi^\varepsilon(x)}}{k}} f^\varepsilon \le \frac{C}{k}||n^\varepsilon\sqrt{|\varphi^\varepsilon|}||_{L^\infty}(\varphi^\varepsilon(x))^{1/4}, \quad \varphi^\varepsilon(x) \ge -10\varepsilon^2\ln\varepsilon$$

$$\le \frac{C}{k}||n^\varepsilon\sqrt{|\varphi^\varepsilon|}||_{L^\infty}(-\varphi^\varepsilon(x)\ln\varepsilon)^{1/4}, \quad \varepsilon^2 \le \varphi^\varepsilon(x) \le -10\varepsilon^2\ln\varepsilon.$$

This implies that

$$\sqrt{\varphi^\varepsilon(x)}\int_0^{\frac{\sqrt{\varphi^\varepsilon(x)}}{k}} f^\varepsilon dv \leq \frac{C}{k}||n^\varepsilon\sqrt{|\varphi^\varepsilon|}||_{L^\infty}, \quad x \in [r^\varepsilon, 1],$$

which together with (2.39) leads to

$$(2.40) \qquad \sqrt{\varphi^\varepsilon(x)}\int_{-\infty}^{\frac{\sqrt{\varphi^\varepsilon(x)}}{k}} f^\varepsilon(x,v)\,dv \leq \left(\frac{C}{k}+o(1)\right)||n^\varepsilon\sqrt{|\varphi^\varepsilon|}||_{L^\infty}.$$

Additionally, we have

$$\sqrt{\varphi^\varepsilon(x)}\int_{\frac{\sqrt{\varphi^\varepsilon(x)}}{k}}^{+\infty} f^\varepsilon(x,v)\,dv \leq k\int_{\frac{\sqrt{\varphi^\varepsilon(x)}}{k}}^{+\infty} vf^\varepsilon(x,v)\,dv,$$

and by Lemma 2.4(i), we deduce that

$$\sqrt{\varphi^\varepsilon(x)}\int_{\frac{\sqrt{\varphi^\varepsilon(x)}}{k}}^{+\infty} f^\varepsilon(x,v)\,dv \leq Ck, \quad x \geq r^\varepsilon.$$

This, combined with (2.40), implies that

$$||n^\varepsilon\sqrt{|\varphi^\varepsilon|}||_{L^\infty} \leq Ck + \left(\frac{C}{k}+o(1)\right)||n^\varepsilon\sqrt{|\varphi^\varepsilon|}||_{L^\infty}.$$

Now, to prove the L^∞ bound, we just have to choose $k > C$ and the result follows immediately.

(ii) follows from estimate (2.35) and (2.37) which give a uniform bound for f^ε as long as $\varphi(x) > 0$ (i.e. $x \neq 0$) and $v < \sqrt{\varphi(x)}$. □

PROPOSITION 2.4. *The convergence of φ^ε toward φ holds in the $C^1([0,1])$ strong topology and $n \in L^\infty_{loc}((0,1]) \cap L^1([0,1])$.*

Proof. (i) First since φ^ε converges uniformly to φ, then $n^\varepsilon\sqrt{|\varphi^\varepsilon|}$ converges in the bounded measure weak topology to $n\sqrt{\varphi}$. We deduce from the preceding lemma that $n\sqrt{\varphi}$ is in $L^\infty(0,1)$ and by a diagonal extraction we deduce that n^ε converges to n in the weak $*$ L^∞ topology on any interval $[\alpha, 1]$ with $\alpha > 0$. This implies that φ^ε converges in $C^1[\alpha,1]$ for $\alpha > 0$. Since $\varphi' \geq 0$ and increasing, then it has a limit $\varphi'(0)$ when x tends to zero. Suppose first that $\varphi'(0) = \lim \varphi^{\varepsilon\prime}(0)$ has already been shown. Then by the use of Dini's theorem, we deduce that the convergence of $\varphi^{\varepsilon\prime}$ to φ' holds in $C^0([0,1])$. Let us show now that $\lim \varphi^{\varepsilon\prime}(0) = \varphi'(0)$. It is sufficient, in view of Lemma 2.2, to prove the equality for absolute values. We have

$$h^\varepsilon + \frac{1}{4}\left(\frac{d\varphi^\varepsilon}{dx}(x)\right)^2 = k^\varepsilon(x) + \frac{j^\varepsilon x}{\tau}.$$

Therefore, passing to the limit for $x > 0$ and taking into account point (iii) of Lemma 2.4, we get

$$\left|h + \frac{1}{4}\left(\frac{d\varphi}{dx}(x)\right)^2\right| \leq C\sqrt{\varphi(x)} + \frac{jx}{\tau}.$$

Afterward, we let x tend to zero and obtain

$$h + \frac{1}{4}\left(\frac{d\varphi}{dx}(0)\right)^2 = 0.$$

Furthermore, using (2.24), we have

$$\left| h^\varepsilon + \frac{1}{4}\left(\frac{d\varphi^\varepsilon}{dx}(0)\right)^2 \right| = k^\varepsilon(0) \le C\varepsilon$$

which leads to

$$h + \lim \frac{1}{4}\left(\frac{d\varphi^\varepsilon}{dx}(0)\right)^2 = 0.$$

Therefore $\lim |\frac{d\varphi^\varepsilon}{dx}(0)| = |\frac{d\varphi}{dx}(0)|$, which is the expected result.

(ii) Since $n\sqrt{\varphi} \in L^\infty([0,1])$ and φ vanishes only at $x = 0$, we can deduce that

$$n(x) = n_{reg}(x) + n_0\delta(x),$$

where n_{reg} is an $L^\infty_{loc}((0,1]) \cap L^1([0,1])$ function and n_0 is a nonnegative constant. Since by the weak convergence of n^ε to n we have

$$\varphi'(x) = \varphi'(0) + n_0 + \int_0^x n_{reg},$$

we deduce by letting x tend to zero that

$$\varphi'(0) = n_0 + \varphi'(0),$$

which implies that $n_0 = 0$ and ends the proof. □

3. Characterization of the limit problem. We can pass to the limit in the Boltzmann equation since the nonlinear term passes to the limit. Indeed, for every test function $\theta(x,v) \in C^1([0,1] \times \mathbb{R})$ such that $\theta = 0$ is in the neighborhood of $\{0\} \times \mathbb{R} \cup \{1\} \times \mathbb{R}_+$, we have

$$\int_{x,v} f^\varepsilon \left[v\frac{\partial \theta}{\partial x} + \frac{1}{2}\frac{d\varphi^\varepsilon}{dx}\frac{\partial \theta}{\partial v} - \frac{\theta}{\tau} \right] dxdv + \frac{1}{\tau}\int_x \left(\int_v M_0(v)\theta(x,\varepsilon v)dv \right) n^\varepsilon(x)dx = 0.$$

Since $\varphi^{\varepsilon\prime}$ converges uniformly and f^ε converges weakly, the nonlinear term

$$f^\varepsilon \frac{d\varphi^\varepsilon}{dx}\frac{\partial \theta}{\partial v}$$

passes to the limit and we obtain finally

$$\left\langle\left\langle f, v\frac{\partial \theta}{\partial x} + \frac{1}{2}\frac{d\varphi}{dx}\frac{\partial \theta}{\partial v} - \frac{\theta}{\tau} \right\rangle\right\rangle_{\mathcal{M}_b, C_0} + \frac{1}{\tau}\int_x \theta(x,0)n(x)dx = 0. \tag{3.41}$$

This equation means that for $(x,v) \in (0,1] \times \mathbb{R}_+$ f is a solution of the equation

$$v\frac{\partial f}{\partial x} + \frac{1}{2}\frac{d\varphi}{dx}\frac{\partial f}{\partial v} + \frac{f}{\tau} = 0$$

with the boundary condition

$$f(x,0) = \frac{2n(x)}{\tau\varphi'(x)}.$$

Notice that φ is in $W^{2,\infty}_{loc}(0,1]$ and that the characteristics passing through a point (x,v) such that $0 < x$ and $0 \leq v < \sqrt{\varphi(x)}$ are issued from the axis $(0,1]$. Therefore, $f(x,v)$ has a unique expression in terms of the potential and the density (obtained after an integration of the equation over the characteristics)

(3.42)
$$f(x,v) = f_2(x,v) = \frac{2n(x^*)}{\tau\varphi'(x^*)} \exp\left(-\frac{1}{\tau}\int_{x^*}^{x} \frac{dz}{\sqrt{\varphi(z)-\varphi(x^*)}}\right), \ v \in [0, \sqrt{\varphi(x)}),$$

where x^* is uniquely defined by

$$\varphi(x^*) = \varphi(x) - v^2.$$

Since f is a bounded measure supported in the set $\{0 \leq v \leq \sqrt{\varphi(x)}\}$ and since $f = f_2$ almost everywhere in the set $\{0 \leq v < \sqrt{\varphi(x)}\}$, then we can conclude that

$$f = f_1 + f_2,$$

where f_1 is a bounded measure supported in the set $\{v = \sqrt{\varphi(x)}\}$. We now determine the measure f_1 by noticing that the current is constant. Since f_1 is supported in the set $\{v = \sqrt{\varphi(x)}\}$, then it reads

$$f_1 = n_1(x)\delta(v - \sqrt{\varphi(x)}),$$

where n_1 is a bounded measure on $[0,1]$. We have obviously $n_1 = \langle f_1, 1\rangle_v$. Let j_1 be the current associated with f_1

$$j_1 = \langle f_1, v\rangle_v.$$

We have obviously

$$j_1(x) = n_1(x)\sqrt{\varphi(x)}.$$

Let n_2 and j_2 be the density and the current corresponding to f_2 defined above. We have

$$f(x,v) = f_1(x,v) + f_2(x,v),$$

$$n(x) = n_1(x) + n_2(x),$$

$$j = j_1(x) + j_2(x).$$

Moreover, we have

$$n_2 = \langle f_2, 1\rangle_v = \int_0^{\sqrt{\varphi(x)}} f_2(x,v)dv.$$

The change of variable $y = x^*(x, v)$ gives

$$n_2(x) = \int_0^x \frac{n(y)}{\tau\sqrt{\varphi(x) - \varphi(y)}} \exp\left(-\frac{1}{\tau}\int_y^x \frac{dz}{\sqrt{\varphi(z) - \varphi(y)}}\right) dy. \tag{3.43}$$

Analogously, we get

$$j_2(x) = \int_0^x \frac{n(y)}{\tau} \exp\left(-\frac{1}{\tau}\int_y^x \frac{dz}{\sqrt{\varphi(z) - \varphi(y)}}\right) dy. \tag{3.44}$$

Therefore

$$j_2(x) \leq \frac{1}{\tau}\int_0^x n(y)dy$$

and since n is an L^1 function, then

$$\lim_{x \to 0} j_2(x) = 0.$$

Thus

$$j_1(0) = j. \tag{3.45}$$

Notice that this identity was an *assumption* in the derivation of the limit problem in [3]. Here it is *proved.* The differentiation of (3.44) using (3.43) implies that

$$\frac{dj_2}{dx} = \frac{n}{\tau} - \frac{n_2}{\tau} = \frac{n_1}{\tau} = \frac{j_1}{\tau\sqrt{\varphi}}$$

and since $j = j_1 + j_2$ is constant, we deduce that

$$\frac{dj_1}{dx} + \frac{j_1}{\tau\sqrt{\varphi}} = 0,$$

which together with the condition (3.45) gives

$$j_1(x) = j \exp\left(-\frac{1}{\tau}\int_0^x \frac{dz}{\sqrt{\varphi(z)}}\right) \tag{3.46}$$

and

$$n_1(x) = \frac{j}{\sqrt{\varphi(x)}} \exp\left(-\frac{1}{\tau}\int_0^x \frac{dz}{\sqrt{\varphi(z)}}\right). \tag{3.47}$$

Proof of Theorem 1.4. At this stage, we have proven that (1.11)–(1.18) are satisfied. To prove Theorem 1.4, it remains now to determine the current. Point (iii) is straightforward since $j = j_g$ if $\varphi'(0) > 0$ (Lemma 2.3) and $j = j_{CL}$ if $\varphi'(0) = 0$ (Theorem 1.2). To prove point (i), we suppose that $j \neq j_{CL}$. Then $j = j_g$ and this is not possible in view of Theorem 1.3. Point (ii) follows easily from the fact that $j \leq j_g$, which implies that $j < j_{CL}$. □

4. Conclusion. In this paper we gave a mathematical justification of a formal model derived in a previous paper [3] from a singular perturbation problem. The convergence result is not complete because the limit problem is not fully solved; namely, the uniqueness of the solution of the limit problem is not shown and the possible difference between the Child–Langmuir current j_{CL} and the maximal current $j_{\max}$ is not known. Such problems were already noted in [7] and [8] where the authors show that the maximal current can be different from the Child–Langmuir one in some particular situations. In that case, statement (iii) of Theorem 1.4 cannot be improved.

Acknowledgments. The author would like to thank Pierre Degond and Frédéric Poupaud for many encouraging discussions and enlightening remarks.

REFERENCES

[1] H. U. Baranger, *Ballistic electrons in a submicron semiconducting structure: A Boltzmann equation approach*, Ph.D. thesis, Cornell University, Ithaca, NY, 1986.

[2] N. Ben Abdallah, *The Child–Langmuir regime for electron transport in a plasma including a background of positive ions*, Math. Models Methods Appl. Sci., 4 (1994), pp. 409–438.

[3] N. Ben Abdallah and P. Degond, *The Child–Langmuir law for the Boltzmann equation of semiconductors*, SIAM J. Math. Anal., 26 (1995), pp. 364–398.

[4] ———, *The Child–Langmuir law in the kinetic theory of charged particles, part* 3: *Semiconductors*, to appear.

[5] ———, *On the Child-Langmuir law for semiconductors*, in Mathematics and Its Applications, vol. 59, Springer-Velag, New York, 1994.

[6] P. Degond and F. Guyot-Delaurens, *Particle simulations of the semiconductor Boltzmann equation for one-dimensional inhomogeneous structure*, J. Comput. Phys. 90 (1990), pp. 65–97.

[7] P. Degond, S. Jaffard, F. Poupaud, and P. A. Raviart, *The Child–Langmuir asymptotics of the Vlasov–Poisson equation for cylindrically or spherically symmetric diodes, part* I: *Statement of the problem and basic estimates*, Math. Methods. Appl. Sci., to appear.

[8] ———, *The Child–Langmuir asymptotics of the Vlasov–Poisson equation for cylindrically or spherically symmetric diodes, Part* II, *Analysis of the reduced problem and determination of the Child–Langmuir current*, Math. Methods Appl. Sci., to appear.

[9] P. Degond and F. J. Mustieles, *A deterministic particle method for the kinetic model of semiconductors: The homogeneous field model*, Solid-State Electronics, 34 (1991), pp. 1335–1345.

[10] P. Degond and P. A. Raviart, *An asymptotic analysis of the one-dimensional Vlasov-Poisson system: The Child-Langmuir law*, Asymptotic Anal., 4 (1991), pp. 187–214.

[11] P. Hesto, *Simulation Monte-Carlo du transport non stationnaire dans les dispositifs submicroniques: Importance du phénemène balistique dans* GaAs *a* 77K, Thèse d'Etat, Université Paris-Sud, 1994.

[12] I. Langmuir and K. T. Compton, *Electrical discharges in gases, part* II: *Fundamental phenomena in electrical discharges*, Rev. Modern Phys., 3 (1931), pp. 191–257.

[13] P. A. Markowich, *The stationary semiconductor device equations*, Springer-Verlag, Vienna, New York, 1986.

[14] M. S. Mock, *Analysis of Mathematical Models of Semiconductor Devices*, Boole Press, Dublin, 1983.

[15] F. J. Mustieles, *Global exitence of weak solutions for a system of non-linear Boltzmann equations in semiconductor physics*, Math. Methods Appl. Sci., 14 (1991), pp. 139–153.

[16] B. Niclot, P. Degond, and F. Poupaud, *Deterministic particle simulations of the Boltzmann transport equation of semiconductors*, J. Comput. Phys., 78 (1988), pp. 313–349.

[17] F. Poupaud, *Boundary value problems for the stationary Vlasov–Maxwell systems*, Forum Math., 4 (1992), pp. 499–527.

[18] ———, *Derivation of hydrodynamic system hierarchy for semiconductors from the Boltzmann equation*, Appl. Math. Lett., 4 (1991), pp. 75–79.

[19] ———, *Diffusion approximation of the linear semiconductor Boltzmann equation: Analysis of boundary layers*, Asymptotic Anal., 4 (1991), pp. 293–317.

[20] L. Reggiani, ed., *Hot-Electron Transport in Semiconductors*, Springer, Berlin, 1985.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 110–134, January 1996

008

RIGOROUS WKB FOR FINITE-ORDER LINEAR RECURRENCE RELATIONS WITH SMOOTH COEFFICIENTS*

OVIDIU COSTIN[†] AND RODICA COSTIN[†]

Abstract. The $\epsilon \to 0$ behavior of recurrence relations of the type $\sum_{j=0}^{l} a_j(k\epsilon, \epsilon) y_{k+j} = 0$, $k \in \mathbf{Z}$ (l fixed) is studied. The a_j are C^∞ functions in each variable on $I \times [0, \epsilon_0]$ for a bounded interval I and $\epsilon_0 > 0$. Under certain regularity assumptions we find the asymptotic behavior of the solutions of such recurrences. In typical cases, there exists a fundamental set of solutions in the form $\{\exp(\epsilon^{-1} F_m(k\epsilon, \epsilon))\}_{m=1,\ldots,l}$, where the functions F_m are C^∞ in each variable on the same domain as the a_j, showing in particular that the formal perturbation-series solutions are asymptotic to true solutions of these recurrences. Some applications are also briefly discussed.

Key words. recurrence relations, asymptotic behavior

AMS subject classifications. 41A60, 65M06, 65M12

1. Introduction.

In this paper, we study the asymptotic behavior to all orders in $\epsilon \to 0$ of the solutions of one-dimensional recurrence relations of the form

$$\sum_{j=0}^{l} a_j(k\epsilon, \epsilon) y_{k+j} = 0, \tag{1}$$

which we may interpret as follows: for each fixed k, y_{k+l} is determined from its predecessors $y_k, \ldots, y_{k+l-1}$ (this is assumed possible—see condition (6) below).

Under some further regularity assumptions, we prove that the general solution of the recurrence can be piecewise represented as a sum

$$y_{k,\epsilon} = \sum_{m=1}^{l} C_{m,p} \exp[\epsilon^{-1} F_m(k\epsilon, \epsilon)], \tag{2}$$

where the functions F_m are everywhere smooth with the exception of a small neighborhood of the points where two characteristic roots (24) cross and where the representation is different (Proposition 2.2) . In particular, a fundamental system of solutions can be chosen such that each of them has, for small ϵ and between crossings, a Wentzel–Kramer–Brillouin (WKB)-like expansion

$$y_k \sim \exp\left(\epsilon\phi(k\epsilon)\right)\left(A_0(k\epsilon) + \epsilon A_1(k\epsilon) + \cdots\right), \tag{3}$$

where ϕ is the root of the eikonal equation

$$\exp(\phi'(x)) = \lambda(x),$$

λ is one of the l roots of the characteristic equation (24), and the successive amplitudes A_i can be determined by perturbation–expansion.

For technical reasons, we prefer the less familiar notation (2). It is essential for our arguments that a continuous branch of ln A_0 can be chosen.

* Received by the editors April 26, 1993; accepted for publication (in revised form) May 24, 1994. This work was supported, in part, by Air Force Office of Scientific Research grant AF-0115.

[†] Mathematics Department, Hill Center, Rutgers University, New Brunswick, NJ 08903. Current address: Department of Mathematics, University of Chicago, 5734 University Avenue, Chicago, IL 60637-1538 (costin@math.uchicago.edu).

One of the applications of the rigorous WKB approach for discrete schemes is in determining the spectrum of large matrices with slowly varying entries. Such problems appear for instance in the continuum limit of the Toda lattice. This system, described by the Hamiltonian $H = \frac{1}{2}\sum_{k=1}^{n} p_k^2 + \sum_{k=1}^{n-1} \exp(x_k - x_{k-1})$, is completely integrable; this can be expressed in terms of the constancy of the spectrum of the matrix

$$\begin{pmatrix} a_1 & b_1 & 0 & \cdots & 0 \\ b_1 & a_2 & b_2 & \cdots & 0 \\ 0 & b_2 & a_3 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}. \tag{4}$$

The spectral problem in the case where the coefficients are of the form $a_k = a(k\epsilon)$, $b_k = b(k\epsilon)$, with a and b smooth and satisfying some regularity conditions, leads to a recurrence of the type of (1) that is solved asymptotically by the methods described below [1].

In (1) the number l of steps of the recurrence is fixed, $\epsilon > 0$ is a small parameter, and $k \in \mathbb{Z}$ is such that $k\epsilon \in I$ where $I \subset \mathbb{R}$ is a compact interval. Some initial or boundary conditions are assumed.

The coefficients $a_0(x,\epsilon), \ldots, a_l(x,\epsilon) : I \times [0,\epsilon_0] \to \mathbb{C}$ are assumed C^∞ in x and in ϵ in some domain $I \times [0,\epsilon_0]$. We also require the existence of a uniformly asymptotic series for a_j: for any $t \in \mathbb{N}$,

$$\left| a_j(x,\epsilon) - \sum_{s=0}^{t} a_{j,s}(x)\epsilon^s \right| < M_{j,s}\epsilon^{t+1}, \tag{5}$$

where the functions $a_{j,s}$ are C^∞ in x (for instance, $a_j \in C^\infty(I \times [0,\epsilon_0])$).

We are also imposing the nonsingularity condition

$$\inf_{x\in I}\{|a_0(x,0)|\} > 0 \text{ and } \inf_{x\in I}\{|a_l(x,0)|\} > 0. \tag{6}$$

We begin by giving some simple examples and deriving heuristically their small-ϵ behavior. The contents of the paper will subsequently make the given solutions rigorous.

Example a. Consider the one-step recurrence relation

$$y_{k+1} = e^{f(k\epsilon)} y_k, \tag{7}$$

where $y_0 = 1$, $0 \le k \le \epsilon^{-1}$, and f is a C^∞ function on $[0,1]$. It has the explicit solution

$$y_k = \exp\left(\sum_{j=0}^{k-1} f(j\epsilon)\right). \tag{8}$$

When $f = f_0$ is constant, $y_k = \exp(\epsilon^{-1} f_0 k\epsilon)$ and this should also be the order of magnitude of y_k for a general smooth f when ϵ is very small; we then try a formal asymptotic solution

$$y_k \sim \exp(\epsilon^{-1}\Phi_0(k\epsilon) + \Phi_1(k\epsilon) + \epsilon\,\Phi_2(k\epsilon) + \cdots) \tag{9}$$

in (7) . To be consistent with (7) we must have

$$\exp\left[\sum_{m=0}^{\infty}\epsilon^{m-1}\Phi_m(k\epsilon+\epsilon)\right]\sim\exp\left[f(k\epsilon)+\sum_{m=0}^{\infty}\epsilon^{m-1}\Phi_m(k\epsilon)\right],\ \epsilon\to 0. \tag{10}$$

Expanding the exponent in the left-hand side of (10) in a Taylor series around $k\epsilon$ and then identifying the corresponding powers of ϵ in (10) we obtain

$$\Phi_0'=f,\quad \Phi_1'=-\frac{1}{2}f',\ldots,\Phi_j'=C_j\ f^{(j)},\ldots, \tag{11}$$

where the constants C_j are seen to be f-independent and can thus be determined by choosing a particular f for which the sum in (8) can be explicitly evaluated. Taking, e.g., $f(x)=\exp(x)$, this sum is

$$(\exp(x)-1)(\exp(\epsilon)-1)^{-1}\sim\epsilon^{-1}(\exp(x)-1)\sum_{j=0}^{\infty}\frac{B_{2j}}{(2j)!}\epsilon^{2j},\quad \epsilon\to 0, \tag{12}$$

where B_n are the Bernoulli numbers. It follows that $C_j=B_j$ ($B_{2j+1}=0$) and thus, for any smooth f,

$$\sum_{j=0}^{k-1}f(j\epsilon)\sim\frac{1}{\epsilon}\sum_{n=0}^{\infty}\frac{B_{2n}}{(2n)!}\epsilon^{2n}\int_0^x f^{(n)}(t)dt,\quad \epsilon\to 0, \tag{13}$$

which is, of course, the Euler–Maclaurin summation formula. Proposition 2.1 below justifies the derivation of (13).

Example b. Analogously, we can easily obtain the asymptotic behavior of special functions from their generating recurrence relation. For instance, the recurrence relation

$$y_{k+1}+y_{k-1}=2(1+k\epsilon)y_k \tag{14}$$

has for a fundamental set of solutions the Bessel functions $J_{k+\epsilon^{-1}}(\epsilon^{-1})$ and $Y_{k+\epsilon^{-1}}(\epsilon^{-1})$.

Let $x=k\epsilon$, $\nu=k+\epsilon^{-1}$, $a=\epsilon^{-1}\nu^{-1}$. We fix $\rho<\frac{2}{3}$ and for $x>\epsilon^{\rho}$ we take y of the form (9) determining the successive terms by substituting the formal series in (14). To leading order,

$$e^{\Phi_0'(x)}+e^{-\Phi_0'(x)}=2(1+x) \tag{15}$$

and we can choose two independent solutions

$$\epsilon^{-1}\Phi_{0;\pm}=\pm\nu(\alpha-\tanh(\alpha))\quad(\alpha\equiv(\cosh)^{-1}(1/a)). \tag{16}$$

For the next order $\Phi_{1;\pm}$, we obtain

$$\Phi_{1;\pm}'/\Phi_{1;\pm}=-\frac{1}{2}\Phi_{0;\pm}''\coth(\Phi'), \tag{17}$$

which can again be integrated explicitly to give $\exp(\Phi_{1;\pm})=\text{const}\sinh(\alpha)^{-1/2}$. That is, the asymptotic behavior is

$$(\nu\tanh(\alpha))^{-1/2}(A\exp[\nu(\alpha-\tanh(\alpha))+\cdots]+B\exp[-\nu(\alpha-\tanh(\alpha))+\cdots]) \tag{18}$$

and we get the familiar expressions in the theory of Bessel functions (see [2], [3]). All the successive orders are easily obtained in the same way. For $x < -\epsilon^{\rho}$ we obtain from the same equations (15), (17) similar expressions but with trigonometric instead of hyperbolic functions. For small values of $|x|$, $|x| \lesssim \epsilon^{2/3}$, the above asymptotic series becomes singular. The appropriate WKB-like series in this region is in powers of $\epsilon^{1/3}$ and the coefficients will be smooth functions of $\epsilon^{1/3}k$, but otherwise the calculation can be done explicitly in the same way; there is a region of overlap where both asymptotic series are valid, namely $\epsilon^{2/3} < |x| < \epsilon^{1/2}$, and where the series can be matched.

Proposition 2.2 can be used to make the above approach rigorous.

Example c. Finally, let q be a smooth function and consider the Cauchy problem

$$y'' = q(x)y, \quad y(0) = 0, \quad y'(0) = 1. \tag{19}$$

Assume for simplicity that $q : [0, 1] \to \mathbb{R}^+$ and consider the associated Euler scheme

$$y_{k+\epsilon;\epsilon} + y_{k-\epsilon;\epsilon} = (2 + \epsilon^2 q(k\epsilon))y_{k;\epsilon}, \quad y_{0;\epsilon} = 0, \quad y_{1;\epsilon} = \epsilon. \tag{20}$$

To characterize $y_{k;\epsilon}$ for small ϵ we substitute

$$y_{k;\epsilon} \sim \sum_{m=0}^{\infty} \epsilon^m \Xi_m(k\epsilon) \tag{21}$$

in (20). Note that if $\Xi_0 \neq 0$ we obtain this type of series from (9), when $\Phi_0 = 0$, by expanding the exponential.

The substitution leads to the equations

$$\Xi_m'' = q(x)\Xi_m - 2 \sum_{1 \le s; 2s \le m} \frac{\Xi_{m-2s}^{(2s)}}{(2s)!} \tag{22}$$

with the initial conditions

$$\Xi_m(0) = 0, \quad \Xi_m'(0) = -\sum_{k=1}^{m} \frac{\Xi_{m-k}^{(k+1)}(0)}{(k+1)!}. \tag{23}$$

In particular, it follows that the scheme converges to the solution of the given Cauchy problem (as it should) and because, as it is easily seen, $\Xi_1 \equiv 0$, the error $|y_{k+\epsilon} - \Xi_0(k\epsilon)|$ is $O(\epsilon^2)$. Proposition 2.3 applies to this example.

2. Main results. In this section, we consider the problem (1) and the assumptions following it and give conditions under which the exact solutions of the recurrence have asymptotic series to all orders in ϵ.

PROPOSITION 2.1. *Let $\lambda_1(x), \dots, \lambda_l(x)$ be the roots of the characteristic polynomial*

$$\sum_{j=0}^{l} a_j(x, 0)\lambda^j = 0 \tag{24}$$

and assume that they are simple throughout I

$$\inf_{x \in I}\{| \lambda_m(x) - \lambda_n(x) |\} > 0 \quad \textit{if } m \neq n. \tag{25}$$

(consequently, we will choose λ_m to be C^∞ functions). Suppose also that the interval I is a finite union of (closed) intervals such that in each one of them the ordering of the moduli of the roots does not change, i.e.,

$$\begin{gathered} I = \bigcup_{p=1}^{P} I_p \quad \textit{and } \forall p \leq P \; \exists (i_1, \ldots, i_l) \textit{ such that} \\ |\lambda_{i_1}(x)| \leq |\lambda_{i_2}(x)| \leq \cdots \leq |\lambda_{i_l}(x)| \quad \textit{for all } x \in I_p, \end{gathered} \tag{26}$$

where $(i_1, \ldots, i_l)$ is a permutation of $(1, \ldots, l)$. Then there exists $\epsilon' \leq \epsilon_0$ and l functions $F_1(x,\epsilon), \ldots, F_l(x,\epsilon)$ on $I \times [0,\epsilon']$, C^∞ in each variable, such that for each ϵ, $\{\exp(\epsilon^{-1} F_m(k\epsilon,\epsilon))\}_{m=1,\ldots,l}$ form a fundamental set of solutions of the recurrence relation (1), in the sense that for any solution $y_{k,\epsilon}$ of (1), there exist constants $C_1^{(p)}, \ldots, C_l^{(p)}$ such that in each I_p,

$$y_{k,\epsilon} = \sum_{m=1}^{l} C_m^{(p)} \exp[\epsilon^{-1} F_m(k\epsilon,\epsilon)]. \tag{27}$$

REMARK 1. *In particular, this means that for small ϵ,*

$$F_m(x,\epsilon) \sim \sum_{s=0}^{\infty} \Phi_{m,s}(x) \epsilon^s, \tag{28}$$

where the $\Phi_{m,s}$ are smooth; they can be computed from (1) by usual perturbation expansions in ϵ. For example, the first term is obtained from the eikonal equation,

$$\exp(\Phi'_{m,0}(x)) = \lambda_m(x), \tag{29}$$

giving the connection between the functions F_m and the roots λ_m of the characteristic polynomial. The second term is obtained from

$$\Phi'_{n,1}(x) = -\frac{\sum_{j=0}^{l} [j^2 \Phi_{n,0}(x)''/2 A_{j,0}(x) + \exp(\Phi'_{j,0}(x)) A_{j,1}(x)]}{\sum_{j=0}^{l} j A_{j,0}(x) \exp(\Phi'_{j,0}(x))} \tag{30}$$

and so on.

Note also that if not all the functions F_m have the same real part, then there exist solutions of the recurrence which are exponentially small relative to the dominant ones and which will therefore be unstable in the sense that a small "generic" perturbation of the initial condition will completely change the behavior of these solutions; a perturbation series not involving terms beyond all orders will only see the component of the solution along the dominant directions. However, the relative size of the solutions could change with x and then all the coefficients in the expansion (27) could be important for matching different regions.

We now address the question of the asymptotic behavior of the solutions of the recurrence when two characteristic roots cross. Proposition 2.1 is generalized below to the case when two of the roots of the characteristic polynomial become equal at some point in I provided the roots do not coalesce too quickly (condition (31)). In this case a fundamental set of solutions has a more complicated structure. Not too close to the crossing point, a solution is still a linear combination of the form (27) but the coefficients $C_m^{(p)}$ can change at the crossing (Stokes phenomenon). Very close to the crossing, the coalescing roots bring in the asymptotic expression of the solution series in noninteger powers of ϵ. This is in some sense the discrete counterpart of the turning-point behavior of the solution of a differential equation depending on a small parameter.

Consider a subinterval I_0 of I such that two roots of the characteristic polynomial cross once in I_0 (say, at $x = 0$) and except for this, the ordering of the moduli of the characteristic roots is constant in I_0. The crossing is assumed to be generic:

$$|\lambda_m(x) - \lambda_n(x)| > \mathrm{const}\sqrt{|x|} \quad (\text{for } m \neq n). \tag{31}$$

To avoid excessive branching of the discussion and formuli, we also assume that the coalescing roots are *complex conjugate* for negative x and *real valued* for x positive. The general case is treated in a very similar way.

PROPOSITION 2.2. *Fix two constants* $\frac{1}{3} < \beta < \alpha < \frac{1}{2}$. *Then,*

(i) *For* $|k| > \epsilon^{-\beta}$ *the general solution of the recurrence can be written in the form*

$$y_k = \sum_{j=1}^{l} C_j \exp(\epsilon^{-1} F_j(k\epsilon, \epsilon)) \tag{32}$$

with F_j *as in Lemma* 3.2. *The constants* C_j *depend, in general, on the sign of* k.

(ii) *For* $|k| < \epsilon^{-\alpha}$ *a fundamental set of solutions can be chosen such that* $l-2$ *solutions are of the form* $\exp(\epsilon^{-1}F_j(k\epsilon,\epsilon))$ *and two special solutions are of the form*

$$\exp(F_{\pm}(k\epsilon^{1/3}, \epsilon^{1/3})),$$

where the functions F_j *and* $F_{\pm}$ *are smooth and* $\exp(F_{\pm}(x,0)) = \mathrm{Ai}(\Theta\, x) \pm \mathrm{Bi}(\Theta\, x)$, *where* Ai *and* Bi *are the Airy functions* (*for the value of* Θ, *see equation* (89)).

Moreover there is a particular solution of the recurrence which has the behavior

$$y_k \sim \mathrm{Ai}(\Theta\, k\epsilon^{1/3})(1 + \epsilon^{1/3} A_1(\Theta\, k\epsilon^{1/3}) + \cdots)$$

for large $k < \epsilon^{-\alpha}$.

The representations in (i) *and* (ii) *are simultaneously valid in the region* $\epsilon^{-\beta} < |k| < \epsilon^{-\alpha}$, *where the asymptotic series can be matched.*

Note. A similar result can be proven if condition (31) is replaced by

$$|\lambda_m(x) - \lambda_n(x)| > \mathrm{const}\, x^{\mathrm{const}'} \tag{33}$$

for some const > 0. Another case which is interesting for schemes that converge to differential equations for small ϵ is covered by the following proposition, in which we assume that the roots of the *complete* characteristic polynomial

$$\sum_{j=0}^{l} a_j(x,\epsilon)\lambda^j, \qquad m = 1, \dots, l \tag{34}$$

are nondegenerate in a *higher* order in ϵ.

PROPOSITION 2.3. *Assume that the roots* $\lambda_1(x,\epsilon), \dots, \lambda_l(x,\epsilon)$ *of* (34) *satisfy the estimates*

$$|\lambda_m(x,\epsilon) - 1| = \epsilon^q (Q_m(x) + o(\epsilon)), \quad m \leq l \tag{35}$$

for some $q \in \mathbb{N}$, where the smooth functions Q_i verify

$$\inf_{x\in I} Q_m(x) > 0, \ \inf_{x\in I} |Q_m(x) - Q_n(x)| > 0 \qquad (m \neq n)\ m,n \leq l.$$

Then the conclusion of Proposition 2.1 *holds and, furthermore, for any formal series solution of the recurrence relation* (1), *there exists a true solution which is asymptotic to it.*

In this particular case, since, as we shall see, $F_m(x,0) = 0$, it is more natural to represent the formal solutions as power series

$$\Gamma = \sum_{i=0}^{\infty} \Psi_i(x)\epsilon^i, \tag{36}$$

where the Ψ_i are smooth and subject to the condition

$$\sum_{j=0}^{l} a_j(x,\epsilon)\Gamma_s(x+j\epsilon;\epsilon) = o(\epsilon^s) \qquad \forall s,\ x \in I, \tag{37}$$

where $\Gamma_s(x) = \sum_{i=0}^{s} \Psi_i(x)\epsilon^i$. The series (36) and those appearing in the previous WKB expansions are usually divergent and one could imagine that by iterating the recurrence the small error appearing in the local condition (37) could quickly reach O(1). Under the given restrictions, however, Proposition 2.3 guarantees that Γ_s is indeed an $o(\epsilon^s)$ approximation to a genuine solution.

It can now be checked without difficulty that Propositions 2.1–2.3 apply to Examples a, b, and c, respectively.

The layout of the remainder of this paper is as follows: in §3 we prove our main results and in §4 we discuss some further applications of these results.

3. Proof of the results. To prove Proposition 2.1, we show (Lemma 3.1) the existence of l formal series solutions of the form (9) to the recurrence.

We then show (Lemma 3.2) that the proposition is true if $P = 1$ (cf. (26)). The proof is by induction on the order l of the recurrence. First, we choose the particular formal solution corresponding (cf. (39)) to the root with maximum modulus, which gives the "stable" direction, and show that there is a true solution with this asymptotic behavior. Next, we use this particular solution to decrease the order of the recurrence by 1.

3.1. The functions $\Phi_{m,s}$ of the following lemma will turn out to be the functions giving the asymptotic expansions (28) and can be obtained by requiring that (1) is satisfied in all orders in ϵ by the formal solution $y_m = \exp(\epsilon^{-1}\sum_s \Phi_{m,s}(x)\epsilon^s)$.

LEMMA 3.1. *For each $m = 1,\dots,l$ there exists a sequence $\{\Phi_{m,s}\}_{s\in\mathbb{N}}$ of functions in $C^\infty(I)$ such that*

$$\exp(-\epsilon^{-1}\Phi_{m,0}(k\epsilon))\sum_{j=0}^{l} a_j(k\epsilon,\epsilon)\exp\left(\epsilon^{-1}\sum_{t=0}^{s}\epsilon^t\Phi_{m,t}((k+j)\epsilon)\right) = O(\epsilon^{s+1}) \tag{38}$$

and

$$\exp(\Phi'_{m,0}(x)) = \lambda_m(x) \tag{39}$$

for $\epsilon \leq \epsilon_0$, $s \in \mathbb{N}$, and $k\epsilon \in I$.

The proof of Lemma 1 is by induction on the expansion order s. In view of (5) and (6), we can define for each characteristic root $\lambda_m(x)$ (i.e., root of (24)), a function $\Phi_{m,0} \in C^\infty(I)$ such that (39) holds.

It is then straightforward to show that $\exp(\epsilon^{-1}\Phi_{m,0})$ verifies the recurrence (1) to $O(\epsilon)\exp(\epsilon^{-1}\Phi_{m,0}(x))$ so that (38) holds for $s = 0$. Assuming now that $\Phi_{m,0}, \Phi_{m,1}, \ldots, \Phi_{m,s_0}$ are already defined so that for all $s \leq s_0$ (38) is verified, one can easily check that for any $\Psi \in C^\infty(I)$,

(40)

$$\exp[-\epsilon^{-1}\Phi_{m,0}(k\epsilon)]\sum_{j=0}^{l}\exp\left\{\epsilon^{-1}\sum_{t=0}^{s_0}\Phi_{m,t}((k+j)\epsilon)\epsilon^t + \epsilon^{s_0+1}\Psi((k+j)\epsilon)\right\}a_j(k\epsilon,\epsilon)$$

$$= \epsilon^{s_0+1}\left[\Psi'(k\epsilon)\sum_{j=0}^{l} j\exp\{j\Phi'_{m,o}(k\epsilon)\}a_j(k\epsilon,0) + H_{s_0}(k\epsilon,\epsilon)\right] + O(\epsilon^{s_0+2}),$$

where H_{s_0} is a smooth function.

Since by (6) and (25)

$$\inf_{x\in I}\left|\sum_{j=0}^{l} j a_j(x,0)\lambda_m^j(x)\right| > 0, \tag{41}$$

one can define a smooth function $\Psi(x) \equiv \Phi_{s_0+1}(x)$ such that the term in square brackets in the right-hand side of (40) vanishes. □

We note at this point that the series $\sum_{s=0}^{\infty}\Phi_{m,s}(x)\epsilon^s$ is usually not convergent and there does not yet follow the existence of a solution asymptotic to it.

Now we address the question of existence of true solutions of the recurrence having the prescribed asymptotic behavior. In what follows, we shall define, by a formal solution, an expression $\tilde{Y} = \exp(\epsilon^{-1}\sum_{s=0}^{\infty}\Phi_{m,s}(x))\epsilon^s$ satisfying the conclusions of Lemma 3.1. Given $\Phi_{m,0}$, the $\Phi_{m,s}$ are uniquely determined up to integration constants.

Assume first $P = 1$ (cf. (26)). Relabelling if necessary, we assume that for $m \leq n$, $|\lambda_m(x)| \leq |\lambda_n(x)|$ on I.

LEMMA 3.2. *Let $\tilde{Y}$ be a formal solution of* (1) *($m \in \{1, \ldots, l\}$ fixed). There exists a sequence $\{Y_{m,k;\epsilon}\}_{k,\epsilon}$ such that for any $\epsilon \leq \epsilon_0$, $Y_{m,k;\epsilon}$ is a solution of* (1) *for $k\epsilon \in I$ having Γ_m as an asymptotic series in the sense that there is a sequence of positive constants $\{C_s\}_s$ such that*

$$\left|Y_{m,k;\epsilon}\exp\left[-\epsilon^{-1}\sum_{t=0}^{s}\Phi_{m,t}(k\epsilon)\epsilon^t\right] - 1\right| < C_s\,\epsilon^{s+1}, \quad \textit{for all } s \in \mathbb{N}. \tag{42}$$

REMARK 2. *The previous lemma can be restated as follows: For each $m = 1, \ldots, l$ there is a function $F_m(x,\epsilon) : I \times [0,\epsilon_0] \to \mathbb{C}$, smooth in each variable, such that $\exp(\epsilon^{-1}F_m(k\epsilon,\epsilon))$ is a solution of the recurrence* (1) *for every ϵ, $k\epsilon \in I$ and, as $\epsilon \to 0$,*

$$F_m(x,\epsilon) \sim \sum_{t=0}^{\infty}\Phi_{m,t}(x)\epsilon^t. \tag{43}$$

This remark follows easily from a classical result (see, e.g., [4, p. 33] and [5]) stating that for any sequence of numbers there is a smooth function having that sequence for its derivatives at the origin and from the easily proven fact that for each sequence $\{(x_n, a_n)\}_{n\in\mathbb{N}}$, $x_n \searrow 0$ and $a_n n^k \to 0$ $\forall k$, one can construct a C^∞ "interpolation" function f such that $f^{(k)}(0) = 0$ $\forall k$ and $f(x_n) = a_n$ for all n.

Some of the estimates that we need for proving Lemma 2 become to a certain extent easier by the remark that a global shift in the estimates defining an asymptotic series is unimportant.

REMARK 3. *Let F be a function such that for a fixed $s_0 \geq 0$ and any s in $\mathbb{N}$*

$$\limsup_{\epsilon\to 0} \epsilon^{-s-1} \left| F(\epsilon) - \sum_{t=0}^{s+s_0} C_t \epsilon^t \right| = D_s. \tag{44}$$

Then

$$\limsup_{\epsilon\to 0} \epsilon^{-s-1} \left| F(\epsilon) - \sum_{t=0}^{s} C_t \epsilon^t \right| \leq D_s + |C_{s+1}|. \tag{45}$$

Thus, to prove (42) we need only show that for some fixed s_0 (we drop the subscript ϵ to ease the notation),

$$\left| Y_{m,k} \exp\left(-\epsilon^{-1} \sum_{t=0}^{s} \Phi_{m,t}(k\epsilon)\epsilon^t \right) - 1 \right| < C_s\ \epsilon^{s+1-s_0}. \tag{46}$$

REMARK 4. *In view of condition* (25), *it is easy to check that, for small ϵ, the solutions* $\exp(\epsilon^{-1}F_m(k\epsilon,\epsilon))$ *are linearly independent, thus forming a fundamental set of solutions on I.*

Proof of Lemma 3.2. The proof is by induction on the order l of the recurrence.

Step 1. For any l, in the same conditions as in Lemma 3.1, the estimate (42) holds for $m = 1$ (recall that $\lambda_1(x)$ is the largest in absolute value). First we show that if a solution is asymptotic to the formal solution corresponding to λ_1 at the left end of I it remains asymptotic to it throughout I. We can assume without loss of generality that the left end of I is at $x = 0$. At this point, we choose appropriate initial conditions.

Let η_p, $p = 1, \ldots, l$ be any functions such that as $\epsilon \to 0$

$$\eta_p(\epsilon) \sim \exp\left(\epsilon^{-1} \sum_{t=0}^{\infty} \Phi_{1,t}(p\epsilon)\epsilon^t \right) \tag{47}$$

(see Remark 2) and let

$$\tilde{Y}_{\epsilon;s}(x) = \exp\left(\epsilon^{-1} \sum_{t=0}^{s} \Phi_{1,t}(x)\epsilon^t \right). \tag{48}$$

Let also $Y_{1,k}$ be the solution of (1) satisfying the initial conditions $Y_{1,p} = \eta_p(\epsilon)$, $p = 1, \ldots, l$.

It is natural to rescale the recurrence relative to its approximate solution. Let

$$C_{k;s} = Y_{1,k}/\tilde{Y}_{\epsilon;s}(k\epsilon). \tag{49}$$

Then the recurrence relation for $C_{k;s}$ can be written

$$\sum_{j=0}^{l} \tilde{a}_{j,s}(k\epsilon,\epsilon) C_{k+j;s} = 0, \tag{50}$$

where $\tilde{a}_{j,s}(k\epsilon,\epsilon) = a_j(k\epsilon,\epsilon)\tilde{Y}_{\epsilon;s}((k+j)\epsilon)/\tilde{Y}_{\epsilon;s}(k\epsilon)$. It can be seen that (50) is of the same type as (1) and that, for ϵ small enough, it satisfies the corresponding assumptions (6), (25), and (26) on I.

Also, $\max\{|\tilde{\lambda}_m(x)|; x \in I,\ m = 1,\dots,l\} = 1$. Now, (46) means that for some fixed s_0,

$$|C_{k;s} - 1| < \text{const}_s\epsilon^{s-s_0+1}. \tag{51}$$

From the definition of $\tilde{Y}_{\epsilon;s}$, it follows that

$$\sum_{j=0}^{l} \tilde{a}_{j,s}(k\epsilon,\epsilon) = O(\epsilon^s). \tag{52}$$

Rewriting the recursion relation (50) in matrix form $\mathbf{C}_{k+1} = \tilde{A}_k\mathbf{C}_k$, where

$$\tilde{A}_k = \begin{pmatrix} \frac{-\tilde{a}_{l-1}(k\epsilon,\epsilon)}{\tilde{a}_l(k\epsilon,\epsilon)} & \frac{-\tilde{a}_{l-2}(k\epsilon,\epsilon)}{\tilde{a}_l(k\epsilon,\epsilon)} & \cdots & \frac{-\tilde{a}_0(k\epsilon,\epsilon)}{\tilde{a}_l(k\epsilon,\epsilon)} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 1 & 0 \end{pmatrix}, \tag{53}$$

and also rewriting (52) as

$$\tilde{A}_k\mathbf{1} = \mathbf{1} + \epsilon^s\mathbf{E}_k, \tag{54}$$

where $\mathbf{1}_j = 1$ and $\|\mathbf{E}_k\| < \text{const}$ uniformly in k, ϵ, we have

$$\mathbf{C}_k = \mathbf{1} + \epsilon^s \sum_{j=1}^{k} \tilde{A}_k\tilde{A}_{k-1}\dots\tilde{A}_{j+1}\mathbf{E}_k. \tag{55}$$

Step 1 is completed by showing (51) (and thus (46)), which follows from the stability lemma below.

LEMMA 3.3. *Let $\tilde{A}_k$ be a family of matrices of the form of* (53), *where $\tilde{a}_j : I \times [0,\epsilon_0] \to \mathbb{C}$ (I is an interval) as in Lemma* 3.2. *Suppose further that the roots $\tilde{\lambda}_m(k\epsilon,\epsilon)$ of the polynomial*

$$\sum_{j=0}^{l} \tilde{a}_j(k\epsilon,\epsilon)\tilde{\lambda}^j = 0, \qquad m = 1,\dots,l \tag{56}$$

satisfy

$$\sup_{k\epsilon\in I}\{|\tilde{\lambda}_m(k\epsilon,\epsilon)|\} \le 1 + \text{const}\ \epsilon,\ m = 1,\dots,l \tag{57}$$

and that the condition corresponding to (25) *is fulfilled on I. Then there is an ϵ-independent constant C such that*

$$\|\tilde{A}_k\tilde{A}_{k-1}\dots\tilde{A}_{j+1}\| \le 1 + C|k-j|\epsilon. \tag{58}$$

Proof of Lemma 3. The eigenvalues of the matrix (53) are the $\tilde{\lambda}_m(k\epsilon,\epsilon)$ and the corresponding eigenvectors matrix is $(\tilde{\Lambda}_k)_{i,j} = \tilde{\lambda}_j(k\epsilon,\epsilon)^{l-i}$. We then write the product on the left-hand side of (58) as

$$\tilde{\Lambda}_k D_k \tilde{\Lambda}_k^{-1}\tilde{\Lambda}_{k-1}D_{k-1}\dots\tilde{\Lambda}_{j+1}D_{j+1}\tilde{\Lambda}_{j+1}^{-1}, \tag{59}$$

where

$$D_p = \operatorname{diag}(\{\tilde{\lambda}_m(p\epsilon,\epsilon)\}_{m=1,\dots,l}), \tag{60}$$

and the proof follows from (57) and the estimate

$$\|\tilde{\Lambda}_p^{-1}\tilde{\Lambda}_{p-1}\| \le 1 + \text{const } \epsilon, \tag{61}$$

which can be checked, for instance, using the following explicit formula, whose elementary proof we omit.

REMARK 5. *Let X and Y be two nonsingular Vandermonde-type matrices $X_{i,j} = x_j^{l-i}$, $Y_{i,j} = y_j^{l-i}$, $i,j = 1,\dots,l$. Then*

$$(X^{-1}Y)_{i,j} = \prod_{n\neq i} \frac{y_j - x_n}{x_i - x_n}. \tag{62}$$

Step 2. The conclusion of Lemma 3.2 for $l = 1$ follows from Step 1. Now we assume that the conclusion of the lemma holds for all recurrences of order less than l and prove it for order l by reduction to the $l-1$ case. In view of the first step, we know that to $|\lambda_1|$ there corresponds a true solution Y_1 for which the asymptotic behavior is the formal solution $\tilde{Y}_1$. We shall use this solution to reduce the order of the recurrence by 1. Let

$$C_k = y_k/Y_{1,k}. \tag{63}$$

The recurrence relation for C_k is then of the form (50), where now

$$\tilde{a}_j(k\epsilon,\epsilon) = a_j(k\epsilon,\epsilon)Y_{1,k+j}/Y_{1,k} \tag{64}$$

and, obviously, the asymptotic behavior to all orders is the same as if we had made the rescaling with respect to a formal solution. The point is that now instead of (52), we have

$$\sum_{j=0}^{l} \tilde{a}_j(k\epsilon,\epsilon) = 0 \tag{65}$$

so that the $y = 1$ is an actual solution. To use this fact, let $d_k = C_{k+1} - C_k$. We get

$$\sum_{s=0}^{l-1} b_s(k\epsilon,\epsilon)d_{k+s} = 0, \tag{66}$$

where $b_s(k\epsilon,\epsilon) = \sum_{j=s+1}^{l} \tilde{a}_j(k\epsilon,\epsilon)$.

The characteristic equation for (66) can be written as

$$\sum_{j=1}^{l} \tilde{a}_j(x,0) \sum_{s=0}^{j-1} \tilde{\lambda}^s = 0$$

or, for $\lambda \neq 1$, as it easily follows from (65),

$$\sum_{j=0}^{l} \tilde{a}_j(x,0)\tilde{\lambda}^j = 0. \tag{67}$$

We first check that the new recurrence satisfies the hypothesis of the lemma. But this is easy since the new coefficients are finite combinations of a_j and in particular $b_0 = \sum_{j=1}^{l} \tilde{a}_j = -a_0$ and $b_{l-1} = a_l$, and in view of (67) the same arguments as in Step 1 apply to ensure that the characteristic roots have the required properties. Then we want to check that we have the required number of appropriate formal solutions. This is also straightforward because we can derive them from the formal solutions of the original equation. Indeed,

$$d_k = y_{k+1}/Y_{1,k+1} - y_k/Y_{1,k}. \tag{68}$$

If we substitute for y a *formal* solution $\tilde{Y}_m$ we obtain the formal expression

$$\begin{aligned} \tilde{d}_k &= \exp\left\{\epsilon^{-1}\sum_{t=0}^{\infty}[\Phi_{m,t}((k+1)\epsilon) - \Phi_{1,t}((k+1)\epsilon)]\epsilon^t\right\} \\ &\quad - \exp\left\{\epsilon^{-1}\sum_{t=0}^{\infty}[\Phi_{m,t}(k\epsilon)) - \Phi_{1,t}(k\epsilon)]\epsilon^t\right\} \\ &= \exp\left\{\epsilon^{-1}(\Phi_m(x) - \Phi_1(x)) + \text{ series}\right\}[\lambda_m(x)/\lambda_1(x)](1+\epsilon\times\text{series}), \end{aligned} \tag{69}$$

which, since $\lambda_m(x)/\lambda_1(x)$ does not vanish, can be written as an exponential of a formal series in the form required by our arguments

$$\exp\left(\epsilon^{-1}\sum_{t=0}^{\infty}\Delta_{m,t}(k\epsilon)\right). \tag{70}$$

Then the expressions (70) for $m = 2, \dots, l$ are formal solutions for the recurrence of order $l-1$ (66) because, by construction, (69) are. It then follows by the induction hypothesis that there exist true solutions of (66) of the form

$$d_{m,k} = \exp(\epsilon^{-1}D_m(k\epsilon, \epsilon)), \tag{71}$$

where $D_m(\,\cdot\,,\,\cdot\,)$ are smooth functions having the asymptotic behavior given by (70).

Step 3. To complete the proof of Lemma 2, it remains only to check that

$$\exp\left(\epsilon^{-1}F_1(k\epsilon,\epsilon)\right)\sum_{p=0}^{k}\exp(\epsilon^{-1}D_m(p\epsilon,\epsilon)) \tag{72}$$

has the asymptotic behavior needed for the original recurrence, i.e.,

$$\exp(\epsilon^{-1}\sum_{t=0}^{\infty}(\Phi_{m,t}(k\epsilon)\epsilon^t)). \tag{73}$$

We let k_1 (k_2) be the left (right, respectively) end of the interval. Both k_1 and k_2 might depend on ϵ. By the definition of the $d_{k;m}$, we have

$$C_{m;k} = C_{m;k_1} + \sum_{i=k_1}^{k} d_{m;i}.$$

With the choice

$$C_{m;k_1} = -\sum_{i=k_1}^{k_2} d_{m;i},$$

we get the particular solution

$$C_{m;k} = \sum_{i=k}^{k_2} d_{m;i},$$

from which, referring to the definition of the $C_{m;k}$, we get a solution of the $l-1$ recurrence in the form (the choice of the sign will become clear later)

$$Y_{m;k} = -Y_{1;k} \sum_{i=k}^{k_2} d_{m;i}$$

whose asymptotic behavior is given by the formal solution $\tilde{Y}_m$. Indeed, for any fixed large s, we have

$$\begin{aligned}
Y_{m;k} &= -Y_{1;k} \left[\sum_{i=k}^{k_2} \exp\left(\epsilon^{-1} \sum_{t=0}^{s} \Delta_t(i\epsilon)\epsilon^t \right) \left(1 + O(\epsilon^{s-1})\right) \right] \\
&= -Y_{1;k} \left[\sum_{i=k}^{k_2} \left(\exp\left(\epsilon^{-1} \sum_{t=0}^{s} \Phi_{m;t}(i\epsilon+\epsilon)\epsilon^t - \Phi_{1;t}(i\epsilon+\epsilon)\epsilon^t \right) \right. \right. \\
&\quad \left. \left. - \exp\left(\epsilon^{-1} \sum_{t=0}^{s} \Phi_{m;t}(i\epsilon)\epsilon^t - \Phi_{1;t}(i\epsilon)\epsilon^t \right) \right) \left(1 + O(\epsilon^{s-1})\right) \right] \\
&= Y_{1;k} \left[\exp\left(\epsilon^{-1} \sum_{t=0}^{s} \Phi_{m;t}(k\epsilon)\epsilon^t - \Phi_{1;t}(k\epsilon)\epsilon^t \right) \right. \\
&\quad \left. - \exp\left(\epsilon^{-1} \sum_{t=0}^{s} \Phi_{m;t}(k_2) - \Phi_{1;t}(k_2)\epsilon^t \right) \right] \\
&\quad + (k_2 - k) \max_k \left| \exp\left(\epsilon^{-1} \sum_{t=0}^{s} \Delta_t(i\epsilon)\epsilon^t \right) \right| O(\epsilon^{s-1}). \qquad (74)
\end{aligned}$$

Because, by assumption, λ_1 has the largest modulus, $\Re(\Phi_{m;0}(k\epsilon) - \Phi_{1;0}(k\epsilon))$ is nonincreasing in k in the given region. Therefore, (74) equals

$$\begin{aligned}
&Y_{m;k} \left(1 + \text{const} \max_k \left\{ \left| \exp\left(\Phi_{m;1}(k\epsilon) - \Phi_{1;1}(k\epsilon) \right) \right| \right\} o(\epsilon^{s-2}) \right) \\
&\quad = Y_{m;k} \left(1 + o(\epsilon^{s-2}) \right) \qquad \square \qquad (75)
\end{aligned}$$

The proof of Proposition 2.2 follows essentially the same steps but is, as expected, more involved in the regions of near breakdown of the asymptotic series. The details are given in the next section.

The proof of Proposition 2.3 is very easy, using an estimate of the form (58) for the matrices corresponding to the original recurrence, an estimate which can be obtained directly from Remark 5 and the hypothesis of the proposition.

4. Proof of Proposition 2.2. We assume at first that the crossing occurs between the two largest characteristic roots and explain at the end of the proof how the general case is reduced to this one.

The layout is as follows. We first study the small region around the crossing point (the interior region), where $\exp(\epsilon^{-1} \sum \Phi_{1,2;t}(x)\epsilon^t)$ fail to be formal solutions (and the series occurring at the exponent cease to be asymptotic series). The new formal solutions are to leading order combinations of Airy functions. Their formal properties (domain of asymptoticity, growth in x) are examined. Next we show that there exist true solutions of the recurrence that are asymptotic to them. It is also shown that there exists a particular true solution which is asymptotic, to leading order, to the function $\mathrm{Ai}(x\epsilon^{-2/3})$ and is important for the matching problem (it gives the exponentially decaying formal solution).

We then show that the formal solutions coming from the exterior region continue to represent correctly the solutions of the recurrence far enough into the interior region (down to $|x| \sim \epsilon^{2/3}$) to allow for matching with the interior ones, which are valid up to $|x| \sim \epsilon^{1/2}$.

4.1. The interior region of the crossing interval. This is the region $|x| \ll \epsilon^{1/2}$; for definiteness we fix $\alpha \in (\frac{1}{2}, \frac{2}{3})$ and take it to be

$$D_\alpha = \{k : |k| < \epsilon^{-\alpha}\}$$

or, in terms of $\xi := k\epsilon^{1/3} := k\delta$, which is, for reasons that will become clear later, the natural variable in this region,

$$D_\alpha = \{\xi : |\xi| < \delta^{1-3\alpha}\}. \tag{76}$$

The basic steps of the proof of existence of solutions with given asymptotics are the same as for Lemma 3.2. We will first obtain a solution corresponding to (one of the two) largest eigenvalues and with it reduce the problem to a lower order, nondegenerate recurrence.

Because the variable of ξ D_α is unbounded, a slight extension of Lemma 3.2 is needed for the interior region. We now allow the interval I in Lemma 3.2 to be of the form (76) but strengthen the other hypothesis. Corresponding to Lemma 3.2, ξ plays the role of x and δ is the counterpart of ϵ. We require the following three conditions in addition to those of Lemma 3.2.

(a) $a_j(\xi, \delta)$ are assumed to have asymptotic series $\sum_t a_{j,t}(\xi)\delta^t$ valid throughout the region D_α which are smooth in the sense that all their formal derivatives with respect to ξ, $\sum_t a_{j,t}^{(m)}(\xi)\delta^t$, exist and $|a_{j,t}^{(m)}(\xi)| < \mathrm{const}\,\delta^{\mathrm{const}'\,t}$.

(b) The roots of the characteristic polynomial

$$\sum_{j=0}^{l} a_j(\xi, 0)\lambda(\xi)^j = 0 \tag{77}$$

are nondegenerate,

$$\inf\nolimits_{D_\alpha} |\lambda_m(\xi) - \lambda_n(\xi)| > \mathrm{const} > 0 \ (m \neq n), \tag{78}$$

and the polynomial itself is nondegenerate in the sense

$$\inf\nolimits_{D_\alpha} \left\{ |a_0(k\delta)|, |a_l(k\delta)|, \frac{1}{|a_j(k\delta)|} \right\} > \mathrm{const} > 0. \tag{79}$$

(c) Finally,

$$|\lambda_m((k+1)\delta, \delta) - \lambda_m(k\delta, \delta)| < \frac{\mathrm{const}}{|k| + 1}$$

in D_α, where $\lambda_m((k+1)\delta,\delta)$ are the roots of the complete polynomial $\sum_{j=0}^{l} a_j(\xi,\delta)\lambda^j$.

LEMMA 4.1. *Under these assumptions for any formal solution of* (1) *of the form*

$$S := \exp\left(\delta^{-1}\Phi_0(\xi) + \sum_{m=0}^{\infty} \Phi_m(\xi)\,\delta^m\right),$$

where the exponent is assumed to be a smooth asymptotic series in the sense defined in (a), *there exists a true solution of the recurrence which is asymptotic to it in* D_a.

Proof. The proof follows closely the proof of Lemma 3.2. We emphasize only the differences: For the recurrence (50), we also have to verify condition (c),

$$\begin{aligned}\left|\tilde{\lambda}_m((k+1)\delta,\delta) - \tilde{\lambda}_m(k\delta,\delta)\right| &= \left|\frac{\lambda_m((k+1)\delta,\delta)}{\lambda_1((k+1)\delta,\delta)} - \frac{\lambda_m(k\delta,\delta)}{\lambda_1(k\delta,\delta)}\right| \\ &< \text{const } (|k|+1)^{-1}.\end{aligned}$$

The equivalent of Lemma 3.3 states now that there is a δ-independent constant C such that

$$\|A_k A_{k-1}\ldots A_{j+1}\| \le (1 + C|k-j|^{\text{const}}). \tag{80}$$

Indeed, by Remark 5 a diagonal term of the matrix $T := \Lambda_p^{-1}\Lambda_{p-1}$ is of the form

$$\begin{aligned}T_{m,m} &= \prod_{n\neq m} \frac{\lambda_m((p-1)\delta,\delta) - \lambda_n(p\delta,\delta)}{\lambda_m(p\delta,\delta) - \lambda_n(p\delta,\delta)} \\ &= \prod_{n\neq m}\left(1 + \frac{\lambda_m((p-1)\delta,\delta) - \lambda_m(p\delta,\delta)}{\lambda_m(p\delta,\delta) - \lambda_n(p\delta,\delta)}\right) = 1 + O((|k|+1)^{-1})\end{aligned} \tag{81}$$

by (c). Similarly, the moduli of the nondiagonal terms are seen to be less than const$/(|k|+1)$. Therefore, $T = \mathbf{I} + R$ where the $\|R\|$ is $O((|k|+1)^{-1})$, and hence the inequality (80) follows.

The last part of the proof of Lemma 3.3 applies here without any significant change. □

Next, we discuss the reduction to the nondegenerate case. We consider the initial recurrence in the neighborhood of a crossing point, say $x = 0$, where $\lambda_1(0) = \lambda_2(0) = 1$ (the value at zero can be chosen through a trivial global rescaling of the recurrence). It is convenient to consider rescaled variables $\delta = \epsilon^{1/3}$ and $\xi = k\delta$. In these variables, the coefficients $a_j(x,\epsilon) = a_j(\xi\delta^2,\delta^3)$ have smooth asymptotic series in δ in D_α which are in fact obtained through series expansion in x from (5)

$$a_j(\xi\delta^2,\delta^3) \sim \sum_{s\ge 0} P_{j,s}(\xi)\delta^s, \tag{82}$$

where

$$P_{j,0} = a_j(0,0), \quad P_{j,1}(\xi) = 0, \quad P_{j,2}(\xi) = (D_x a_j)(0,0)\xi,$$

and in general $P_{j,s}(\xi)$ are polynomials in ξ of degree at most $s/2$ for s even and $(s-3)/2$ for s odd. To avoid complicating the notation we write $a_j(\xi,\delta) \equiv a_j(\xi\delta^2,\delta^3)$. We have first to find formal solutions for this new recurrence.

LEMMA 4.2. *There exist l linearly independent formal solutions in D_α of the form*

$$\exp\left(\delta^{-1}\sum_{t=0}^{\infty}\Psi_{m,t}(\xi)\delta^t\right), \tag{83}$$

where $\Psi_{m,t}(\xi)$ are smooth in ξ and satisfy the estimates

$$|\Psi_{m,t}(\xi)| < \mathrm{const}_{m,t} + \mathrm{const}'_{m,t}|\xi|^{t/2+1}.$$

This means in particular that the domain of formal validity of the power series is then $\xi^{1/2}\delta \ll 1$, i.e., $x \ll 1$. The domain in which it is actually asymptotic to the solution is however much smaller ($x \ll \sqrt{\epsilon}$), as we shall see.

Proof of Lemma 4.2. The formal solutions corresponding to the nondegenerate roots give rise automatically to acceptable formal solutions in the new variables ξ, δ. Indeed,

$$\begin{aligned}\exp\left(\epsilon^{-1}\sum_{t=0}^{\infty}\Phi_{m,t}(x)\epsilon^t\right) &= \exp\left(\delta^{-3}\sum_{t=0}^{\infty}\Phi_{m,t}(\xi\,\delta^2)\delta^{3\,t}\right)\\ &= \exp\left(\delta^{-3}\Phi_{m,t}(0) + \delta^{-1}\sum_{s,t\geq 0}^{\infty}\Phi^{(s)}_{m,t}(0)\xi^s\delta^{2\,s+3\,t-2}\right).\end{aligned} \tag{84}$$

The term in δ^{-3} is merely a multiplicative constant so it can be dropped and we are left with a formal solution of the form

$$\exp\left(\delta^{-1}\sum_{t=0}^{\infty}\chi_{m,t}(\xi)\delta^t\right), \tag{85}$$

where the $\chi_{m,t}(\xi)$ are in fact polynomials in ξ of degree $\leq t/2+1$.

For $m = 1, 2$ it is more convenient to write first the possible formal series solutions for the equation in the form

$$\sum_{t=0}^{\infty}\chi_t(\xi)\delta^t \tag{86}$$

and then show that we can write them in the form (85).

Substituting (86) in the recurrence, we get

$$\sum_{j=0}^{l}a_{j,\xi,\delta}\sum_{t=0}^{\infty}\chi_t(\xi + j\delta)\delta^t.$$

The term of order s in (86) is obtained by differentiating the auxilliary equation

$$\sum_{j=0}^{l}a_j(\xi,\delta)\chi(\xi + j\delta,\delta) = 0$$

s times with respect to δ. We get (see (82))

$$\sum_{j=0}^{l}\sum_{t=0}^{s}P_{s-t}(\xi)\left(D_\xi + D_\delta\right)^t\chi\Big|_{\delta=0} = 0,$$

which, after expansion, change of order of summation, and use of (82), gives

$$\sum_{j=0}^{l}\left[\frac{1}{\binom{s}{s-2}}\sum_{\sigma=0}^{s-3}\sum_{t=\sigma}^{s}\left(\binom{s}{t}\binom{t}{\sigma}P_{s-t}(\xi)j^{t-\sigma}D_\xi^{t-\sigma}\chi_\sigma(\xi)\right)\right.$$

$$\left.+D_x a_j(0,0)\xi\,\chi_{s-2}+a_j(0,0)j^2\chi_{s-2}\right]=0. \tag{87}$$

It follows that χ_0 is obtained as a solution of the homogeneous Airy equation

$$\chi''(\xi)=\Theta^3\xi\,\chi(\xi), \tag{88}$$

where

$$\Theta^3=\frac{\sum_{j=0}^{l}D_x a_j(0,0)}{\sum_{j=0}^{l}a_j(0,0)\,j^2}, \tag{89}$$

and that, given $\chi_0,\dots,\chi_{s-3}$, we get χ_{s-2} as a solution of an inhomogeneous Airy equation of the form

$$\chi''(\xi)=\Theta^3\xi\,\chi(\xi)+R(\xi),$$

where $R(\xi)$ is a linear combination of higher derivatives of $\chi_0,\dots,\chi_{s-3}$. To avoid cumbersome notation, we shall assume in the following that Θ is 1. We can check that the assumption of genericity ($|\lambda_1(x)-\lambda_2(x)|>\text{const}\sqrt{x}$) implies $\sum_{j=0}^{l}D_x a_j(0,0)\neq 0$. We shall assume for definiteness that it is negative.

It follows by an obvious induction that the χ_s are smooth. Now we show that they satisfy the inequalities stated in the Lemma 4.2.

Remark. Consider the inhomogeneous Airy equation $f''(\xi)=\xi f(\xi)+R(\xi)$ and assume $R(\xi)\sim\xi^p\exp(2A/3\xi^{3/2})$ with $A=\pm1$ for $x\to\infty$ and $A=\pm i$ at $-\infty$. Then $f(\xi)\sim\xi^{p-1}\exp(2A/3\xi^{3/2})$. This estimate follows immediately from the explicit form of the solution

$$f(\xi)=\ \mathrm{Ai}(\xi)\int^{\xi}R(t)\mathrm{Bi}(t)\,dt-\mathrm{Bi}(\xi)\int^{\xi}R(t)\mathrm{Ai}(t)\,dt.$$

At this point we can show by induction that the solution $\chi_n(\xi)$ grows at most like $\exp(2A/3\xi^{3/2})\xi^{n/2}$. So we assume that this holds for $s\le n$ and we show that it is true for $n+1$. Using the remark above and (87), the induction step is as follows: with $p_n=n/2$,

$$\max_{0\le\sigma\le n;\sigma\le t\le n+3}\left\{\left[\frac{1}{2}(n+1-t)\right]+\frac{(-1)^n+1}{2}+\frac{1}{2}(t-\sigma)+p_n\right\}\le p_{n+1}+1,$$

which is straightforward.

Finally, we argue that there are two linear independent formal solutions of this type that can be written in the form (85) which is convenient for our approach. For this we have to choose χ_0, which is a solution of the homogeneous Airy equation such that it does not vanish in D_α. Since the Wronskian of the couple $\mathrm{Ai}(\xi)$, $\mathrm{Bi}(\xi)$ is a nonzero constant, any combination with real nonzero constants of the form $C_1\mathrm{Ai}(\xi)+i\,C_2\,\mathrm{Bi}(\xi)$ is an everywhere nonzero solution (and the derivative is also nonzero). We can choose two linear independent solutions in this way, say the ones for which χ_0 =Ai(ξ)$\pm i$Bi(ξ). That they are formally linearly independent with respect to the solutions obtained from (84) follows easily, for instance from the fact that they correspond to different roots of the characteristic polynomial.

To show that there is an actual solution for each formal solution in this region, we first single out a true solution corresponding to the dominant characteristic root and then use it to reduce the problem to a regular one. Then we show that they give the expected asymptotic behavior for the solutions of the original equation.

The ideas are similar to those used in the regular case with the exception that extra care is needed along the degenerate directions.

Choose

$$\chi_0(\xi) = \mathrm{Ai}(\xi) + i\,\mathrm{Bi}(\xi)$$

and consider the nonzero formal solution that has the leading order χ_0. We proceed as in Step 1 of §3 to construct a rescaled recurrence with respect to the truncation of our formal solution. The same argument used there shows that the new coefficients $\tilde{a}$ have smooth asymptotic series.

What is new here is that we must provide for the estimate of the type (c) to which end we examine the complete characteristic equation $P(\xi, \delta; \tilde{\lambda}) = 0$. P is a polynomial in $\tilde{\lambda}$ (actually it is, to leading order, a polynomial with constant coefficients), C^∞ in ξ and δ. P(0,0,$\tilde{\lambda}$) has a double root $\tilde{\lambda} = 1$ but by assumption the second derivative does not vanish so that we can obtain the roots of the polynomial perturbatively. After series expansion, we obtain

$$\sum_{j=0}^{l} \{ [a_j(0,0) \left(1 + \delta\, C_j + \delta^2 E_j + O(\delta^3)\right) + \xi D_x a_j \delta^2 + O(\delta^3)]$$

$$\cdot \left(1 + j\chi_1(\xi)\delta + \left(j\chi_2(\xi)\delta^2 + j(j-1)/2\chi_1^2\right) + O(\delta^3)\right)\} = 0, \tag{90}$$

where we have taken $\tilde{\lambda} \sim 1 + \chi_1(\xi)\delta + \chi_2(\xi)\delta^2 + O(\delta^3)$ and

$$\begin{aligned} C_j &= \frac{1}{\chi_0}\left(j\chi_0' + \chi_1\right), \\ E_j &= \frac{1}{\chi_0^2}\left(j^2\chi_0''\chi_0 + j\chi_1'\chi_0 - j\chi_1\chi_0' - \chi_1^2\right). \end{aligned}$$

Using the relations in (82), we get two solutions for χ_1: $\chi_1 = 0$ (actually, as expected, we get a root $\tilde{\lambda}_1 = 1 + O(\delta^s)$) and $\chi_1 = 1 - 2\chi_0'\chi_0^{-1}\delta + O(\delta^2)$. We see that in the first order in δ there is no root crossing, which is not a surprise since a generic perturbation tends separate coalescing roots. The asymptotic series are uniformly valid in our domain D_α. Now we show that

$$\left| \frac{\lambda_2((k+1)\epsilon) - \lambda_2(k\epsilon)}{\lambda_1(k\epsilon) - \lambda_2(k\epsilon)} \right| < \frac{\text{const}}{|k| + 1} \tag{91}$$

(this explains the condition (c) at the begining of this section).

We have

1.

$$\left|(\log \chi_0(\xi))'\right| > \text{const}\sqrt{\xi} + \text{const}'.$$

This is obvious since it holds asymptotically and the function does not vanish. Hence

$$\left|\tilde{\lambda}_2(\delta\, k)) - \tilde{\lambda}_1(\delta\, k)\right| > \text{const}\sqrt{\delta|k|+1},$$

2.

$$\left|(\log h_0(\xi))''\right| < \text{const}/\sqrt{\xi+1} + \text{const}'.$$

Using the asymptotic series for the $\tilde{\lambda}_{1,2}$ and the above estimates (1) and (2) we see that

$$\left|\frac{\tilde{\lambda}_2(\delta\,(k+1)) - \tilde{\lambda}_2(\delta\, k)}{\tilde{\lambda}_2(\delta\, k) - \tilde{\lambda}_1(\delta\, k)}\right| < \frac{(\sqrt{\xi+1}+\text{const}')\delta^2}{(\text{const}\sqrt{\xi}+\text{const}')\delta} < \frac{\text{const}}{|k|+1}.$$

The similar estimates for the other roots are better but this of course does not improve the overall rate of convergence. $\tilde{\lambda}_1 = 1 + O(\delta^s)$ and can be obviously made less than $1+\text{const}\,\delta^2$ in D_α so that also $|\tilde{\lambda}_1((k+1)\delta) - \tilde{\lambda}_1(k\delta)| < 1+\text{const}'\,\delta^2$, which is enough for our purposes. For $m \neq 1, 2$ we can for instance use the fact that the derivative of the polynomial at these points does not vanish and settle for a crude bound $|\tilde{\lambda}_m(\delta\,(k+1)) - \tilde{\lambda}_m(\delta\, k)| < \text{const}\,\delta^2$, which can be obtained immediately from (90).

Now, to see that there is a true solution of the recurrence which is asymptotic to our formal series starting with $\mathrm{Ai} + i\mathrm{Bi}$, we only have to repeat the same arguments used in the regular case.

The next step is to use this particular solution to lower the order of the recurrence. We mimic the construction done in the proof of Lemma 3.2 to get a lower-order recurrence in the variable d_k

$$\sum_{s=0}^{l-1} b_s(k\delta,\delta)d_{k+s} = 0$$

(see (66)) and want to check that this new recurrence satisfies the hypothesis of Lemma 4.1. To leading order, the characteristic polynomial of the above equation has no double roots (it now has only one root equal to 1).

As in the regular case, $b_0(\xi, 0) = -\tilde{a}(\xi, 0) = -a_0(0, 0)$; $b_{l-1}(\xi, 0) = -\tilde{a}_l(\xi, 0) = -a_l(0, 0)$.

Noting that the coefficients of the recurrence (66) have asymptotic series valid throughout the domain (as finite sums of terms of the series of $\alpha_j(\xi,\delta)Y_{1,k+j}/Y_{1,k}$), the boundedness of the coefficients is also trivial.

As in the proof of Lemma 3.2, the polynomial in b_j has the same roots as the polynomial in $\tilde{a}_j$ (except for the eliminated one), for which we have already obtained the estimates of type (c).

Now we have a regular problem for which we know that to each formal solution there is a genuine solution asymptotic to it.

It remains to check that we can recover the asymptotic behavior of the solutions of the original recurrence from those of the reduced one. For the solutions corresponding to the characteristic roots that are less than 1, exactly the same proof as in Step 3 of Lemma 3.2 works. For any formal solution of our original equation that corresponds to the largest eigenvalue and which does not vanish, the proof is the one given in Lemma 4.1.

In the crossing region however there might be a special interest in finding a particular solution which is not of exponential type and which is small for large ξ (the Airy-like solution). To this end, a slightly different argument is necessary. We can obtain a formal solution of the reduced equation which is, to leading order,

$$\frac{\mathrm{Bi}(\xi+\delta)\mathrm{Ai}(\xi)-\mathrm{Bi}(\xi)\mathrm{Ai}(\xi+\delta)}{(\mathrm{Ai}(\xi)-i\mathrm{Bi}(\xi))(\mathrm{Ai}(\xi+\delta)-i\mathrm{Bi}(\xi+\delta))} \tag{92}$$

(suggested by computing (68) for two formal solutions of the original equation, corresponding to Ai ± iBi)). The asymptotic behavior of (92) for large ξ is

$$\frac{\delta}{(\mathrm{Ai}(\xi)-i\mathrm{Bi}(\xi))^2}(1+\text{power series}) \quad (1 \ll |\xi| \ll \delta^{-1/2}). \tag{93}$$

Writing the asymptotic representation of Ai ± iBi in the form

$$\sim \xi^{-1/4}\exp(2/3\xi^{3/2})(1+\text{series})$$

is sufficient to see this. It is important to note that there are no powers of ξ multiplying the asymptotics (93); its leading order does not vanish, and (93) can be written as an exponential of a formal series, the form required by our arguments.

We now apply the construction in Step 3 of §3 to recover the solutions of the initial recurrence.

Using the asymptotic behavior of the Airy functions for large argument, we get for the reconstructed solution the representation

$$Y \sim \xi^{(-1/4)}\exp\left(\frac{2}{3}\xi^{3/2}\right)\sum_{k=j}^{\delta^{-\alpha}}(k\delta)^{1/2}\exp\left(-\frac{4}{3}(k\delta)^{3/2}\right)(1+\text{power series}),$$

for which the Euler–Maclaurin summation formula gives the asymptotic representation

$$\xi^{-1/4}\exp\left(-\frac{2}{3}\xi^{3/2}\right)$$

for large positive ξ.

In conclusion, there is a true solution of the recurrence which behaves like the Airy function for positive k (also for negative k when this is properly interpreted) and our argument shows what initial conditions have to be chosen to obtain it. For negative k we see that, in fact, all the solutions corresponding to the largest two eigenvalues are comparable.

4.2. The exterior region. Now we want to show that the solutions coming from the exterior region remain asymptotic to the true solutions as long as $|x| \gg \epsilon^{2/3}$.

The problem that arises here is that the characteristic polynomial has a virtually degenerate root for small x and this leads to a lesser smoothness of the asymptotic series and ultimately to its collapse at $|x| \sim \epsilon^{2/3}$. Let β be as in Proposition 2.2 and define the exterior region by

$$E_\beta = \{x : |x| > \epsilon^\beta\}.$$

In what follows, we make the following conventions. We write

$$g(\epsilon) = O(\epsilon^\infty)$$

if $\lim_{\epsilon\to 0}\epsilon^{-k}\, g(\epsilon)=0$ for all k; also $\mathcal{F}(x^{\gamma})$ will denote a generic function such that it and its derivatives of arbitrary order s satisfy the estimates

$$\left|\mathcal{F}(x^{\gamma})^{(s)}(x)\right| < A_s + B_s x^{\gamma-s} \tag{94}$$

with the convention that $|\mathcal{F}(x^0)\equiv\mathcal{F}(\ln(x))| < A + B|\ln|x||$.

The first step is to study the asymptotic properties of the formal solutions, i.e., of the (possibly divergent) expressions for which

$$\sum_{j=0}^{l} a_j(x;\epsilon)\exp\left(\sum_{t=0}^{\infty}\Phi_t(x+j\epsilon)\epsilon^t\right)\equiv 0. \tag{95}$$

We show by induction that Φ_t and their derivatives behave like $x^{\frac{3}{2}(1-t)}$ and its derivatives. We place ourselves in the assumption of genericity of the crossing which means in particular that $\frac{\partial P(\lambda;x)}{\partial\lambda} > \text{const}\sqrt{|x|}$.

LEMMA 4.3. (i) *If in* (95) *the asymptotic series for the* a_j *are of the form* $\sum_{k=0}^{\infty}\mathcal{F}_k(x^{1-\frac{3}{2}k})\epsilon^k$, *then, in the formal solution* (95), *we have* $\Phi_t{=}\mathcal{F}(x^{\frac{3}{2}(1-t)})$.

(ii) *The same conclusion is true if* $\frac{\partial P(\lambda;x)}{\partial\lambda} >$ const > 0 *uniformly in* x *and* $a_j\sim\sum_{k=0}^{\infty}\mathcal{F}_k(x^{\frac{1}{2}-\frac{3}{2}k})\epsilon^k$.

Note that the coefficients a_j of our initial recurrence are smoother than it is assumed in (i) but this smoothness does not withstand a rescaling as done for (65).

The proof of Lemma 4.3 is by induction on t.

(i) It is easy to see from the eikonal equation that $\Phi_0{=}\mathcal{F}(x^{3/2})$. Assume that the conclusions of the lemma hold for all $t'<t$. After a formal series expansion of the exponent of (95), one gets

$$\sum_{j=0}^{l} a_j(x;\epsilon)\exp\left[\epsilon^t\left(j\Phi_t^{(s)}+\sum_{\substack{s\,\geq\,1\\ t'\,+\,s\,=\,t\,+\,1}}\frac{j^s}{s!}\Phi_{t'}^{(s)}\right)+\sum_{k=0}^{t}\epsilon^{k-1}\sum_{\substack{s\,\geq\,1\\ t'\,+\,s\,\leq\,t\,+\,1}}\frac{j^s}{s!}\Phi_{t'}^{(s)}+O(\epsilon^{t+1})\right]=0, \tag{96}$$

which, using the induction hypothesis, can be rewritten as

$$\begin{aligned}\sum_{j=0}^{l}\left(\sum\mathcal{F}_{k;j}(x^{1-\frac{3}{2}k})\epsilon^k\right)\exp\Bigg(j\Phi_0'+\epsilon^t\left(j\Phi_t'+\mathcal{F}(x^{1-\frac{3}{2}t})\right)\\ +\sum_{k=1}^{t-1}\epsilon^k\mathcal{F}_k(x^{\frac{1-3k}{2}})+O(\epsilon^{t+1})\Bigg)=0.\end{aligned} \tag{97}$$

After expanding in powers of ϵ and collecting the term in ϵ^t, we obtain the equation for Φ_t in the form

$$\sum_{j=0}^{l}\left(ja_{j;0}(x)\epsilon^{\Phi_0'(x)}\,\Phi_t'(x)+\mathcal{F}(x^{1-\frac{3t}{2}})\right)=0 \tag{98}$$

or

$$\Phi_t'(x)=\frac{\mathcal{F}(x^{1-\frac{3t}{2}})}{\frac{\partial P(\lambda;x)}{\partial\lambda}}, \tag{99}$$

where the derivative of the polynomial is evaluated at $\epsilon = 0, \lambda = \exp(\Phi_0')$ thus proving (i).

For (ii) the same proof works, replacing everywhere $\mathcal{F}(x^{1-\frac{3t}{2}})$ with $\mathcal{F}(x^{\frac{1}{2}-\frac{3t}{2}})$.

4.2.1. Rescaling first the recurrence with respect to the approximate solution, we show that there is a genuine solution corresponding to the maximal eigenvalue. Let Ψ_m be any function such that $\Psi_m(x;\epsilon) \sim \sum_0^\infty \Phi_{m;t}(x)\epsilon^t$ and take

$$\tilde{Y}_m := \exp(\epsilon^{-1}\,\Psi_m(x;\epsilon)). \tag{100}$$

The existence of a solution corresponding to the asymptotics $\exp(\epsilon^{-1}\,\Psi_1(x;\epsilon))$ is again equivalent to a solution which is 1 to all orders in ϵ of the recurrence

$$\sum_{j=0}^{l} \tilde{a}_{j,s}(k\epsilon,\epsilon)C_{k+j} = 0, \tag{101}$$

where $\tilde{a}_{j,s}(k\epsilon,\epsilon) = a_j(k\epsilon,\epsilon)\tilde{Y}_1((k+j)\epsilon)/\tilde{Y}_1(k\epsilon)$. The formal solutions of the equation (101) are $\tilde{Y}_m/\tilde{Y}_1$. We need the roots of the new characteristic polynomial

$$\tilde{P}(\tilde{\lambda}) := \sum_{j=0}^{l} \tilde{a}_{j,s}(k\epsilon,\epsilon)\tilde{\lambda}^j = 0. \tag{102}$$

It is easy to see that the polynomial (102) has a root which is 1 to all orders in ϵ. Now let $G(x;\epsilon)$ be one of the differences $G_m(x;\epsilon) = \Psi_m(x;\epsilon) - \Psi_1(x;\epsilon)$. We have

$$\sum_{j=0}^{l} \tilde{a}_{j,s}(k\epsilon,\epsilon)\exp(\epsilon^{-1}G(x+j\epsilon;\epsilon)) = o(\epsilon^t)$$

for all t and so, after series expansion,

$$\sum_{j=0}^{l} \tilde{a}_{j,s}(k\epsilon,\epsilon)\lambda_0^j(1+\epsilon H_j(x;\epsilon)) = o(\epsilon^\infty), \tag{103}$$

where $H_j(x;\epsilon)$ are some smooth functions of x,ϵ and $\lambda_0 := \exp(G_x(x;\epsilon))$. Using Lemma 4.3 and the genericity assumptions, it is not difficult to see that

$$|H_j(x;\epsilon)| < \text{const}\,|x|^{-1/2}. \tag{104}$$

If we look for solutions of the characteristic polynomial (102) in the form $\lambda_0 + \gamma$, we get

$$0 = \tilde{P}(\tilde{\lambda}) = \sum_{j=0}^{l} P^{(j)}(\lambda_0)\gamma^j$$

(where the derivatives are taken with respect to λ). Using (103) we obtain γ as the unique small solution of the equation

$$\gamma = \frac{\epsilon\sum_{j=0}^{l} \tilde{a}_j(x;\epsilon)\lambda_0^j H_j(x;\epsilon)}{\sum_{j=1}^{l} P^{(j)}(\lambda_0)\gamma^{j-1}}, \tag{105}$$

which is a contraction for small enough ϵ (and small γ) in the region $|x| > \epsilon^\beta$ as it is easy to check. We then obtain from (104) and (105)

$$|\gamma| < \text{const}\frac{\epsilon}{|x|}, \tag{106}$$

again valid for $|x| > \epsilon^\beta$.

We shall also need estimates for $\gamma(x+\epsilon)-\gamma(x)$. Differentiating (105) with respect to x and using Lemma 4.3,

$$|\gamma(x+\epsilon) - \gamma(x)| < \text{const}\,\epsilon\,|x|^{-2}. \tag{107}$$

Now we proceed as in the regular case in rewriting the recurrence in matrix form and evaluating the terms in the product (59). In the matrices $T := \tilde{\Lambda}_k^{-1}\tilde{\Lambda}_{k-1}$, the off-diagonal elements are estimated by $T_{mn} < \text{const}\,\epsilon|x|^{-1}$ and for the diagonal elements we have $T_{mm} = 1 + O(\epsilon/x)$. Indeed,

$$T_{mm} = \prod_{n\neq m}\left(1 + \frac{\lambda_m((p-1)\epsilon,\epsilon) - \lambda_n(p\epsilon,\epsilon)}{\lambda_m(p\epsilon,\epsilon) - \lambda_n(p\epsilon,\epsilon)}\right).$$

Since by the assumption of genericity the roots of the polynomial are separated by at least $\text{const}\sqrt{(|x|)}$, each term in the product above can be estimated by

$$\begin{aligned}1 + \text{const}\,\epsilon(\lambda_0'(x) + \gamma'(x))|x|^{-1/2} + O\left(\epsilon(\lambda_0'(x) + \gamma'(x))|x|^{-1/2}\right)\\ < 1 + \text{const}\left(\epsilon|x|^{-1} + \epsilon^2|x|^{-5/2}\right) < 1 + \text{const}\,\epsilon|x|^{-1}\end{aligned}$$

in our region E_β. The nondiagonal terms are estimated in a very similar way.

We derive the estimate $\|T(x) - \mathbf{I}\| < K\epsilon/x$ for some constant K. Assume for definiteness that we are on the left of the crossing point. We get

$$\left\|\prod_{k;-k\epsilon>\epsilon^\beta} T(k\epsilon)\right\| < \text{const}\,\epsilon^{-K/3}.$$

Finally, we have to control the product of the norms of the diagonal matrices D_k. Since they all have one eigenvalue equal to 1 to all orders in ϵ and for $i > 2$ $|\tilde{\lambda}_i(x;\epsilon)| < 1 - \text{const}$, the only nontrivial contribution comes from $\tilde{\lambda}_2$ and this only if $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ have the same modulus to leading order in ϵ. Referring to the decomposition $\tilde{\lambda} = \lambda_0 + \gamma$, we have in this case, using the Euler–Maclaurin summation formula,

$$\begin{aligned}\prod_{k;-k\epsilon>\epsilon^\beta} |\lambda_0(k\epsilon)| &= \exp\left(\epsilon^{-1}\sum_k \Re((\Psi_2)_x(k\epsilon;\epsilon) - (\Psi_1)_x(k\epsilon;\epsilon))\right)\\ &\sim \exp\left\{\Re\left(\Phi_{2;1}(-\epsilon^{2/3}) - \Phi_{1;1}(-\epsilon^{2/3})\right)\right\} < \epsilon^{-\text{const}}.\end{aligned}$$

Also,

$$\prod_{k;-k\epsilon>\epsilon^\beta}\left|1 + \frac{\gamma(k\epsilon)}{\lambda_0(k\epsilon)}\right| < \prod_{k;-k\epsilon>\epsilon^\beta}\left|1 + \frac{\text{const}}{|k|}\right| < \epsilon^{-\text{const}}$$

so that also $\prod\tilde{\lambda}$ is less than $\epsilon^{-\mathrm{const}}$. At this point, the same arguments as in the regular case show that there is a true solution behaving asymptotically as the formal solution corresponding to the largest eigenvalue.

4.2.2. Let Y_1 be a solution of the recurrence relation such that $Y_1 \sim \exp(\epsilon^{-1}\cdot \sum \Phi_{1;t}\epsilon^t)$ in E_β. We now follow the same steps that led to equations (65) and (66).

It is a matter of straightforward induction to determine by Lemma 4.3 that the coefficients $\tilde{a}_j$ have the behavior

$$\tilde{a}_j(x;\epsilon) \sim a_{j;0}(x)e^{j\,\Phi_0'(x)} + \sum_{k\geq 1}\mathcal{F}_{j;k}(x^{\frac{1}{2}-\frac{3}{2}k})\epsilon^k,$$

and then clearly

$$b_j(x;\epsilon) \sim \sum_{k\geq 0}\mathcal{F}_{j;k}(x^{\frac{1}{2}-\frac{3}{2}k})\epsilon^k. \tag{108}$$

It is also easy to check that the recurrence (66) is now nondegenerate in the sense that

$$\inf_{E_\beta}\left\{|b_0(k\epsilon)|, |b_{l-1}(k\epsilon)|, \frac{1}{|b_j(k\epsilon)|}\right\} > \mathrm{const} > 0 \tag{109}$$

and the characteristic polynomial of the new recurrence does not have coalescing roots (the root $\lambda = 1$ of (65) has been eliminated in the reduction):

$$\inf_{E_\beta}\{|\,|\lambda_m(x)| - |\lambda_n(x)|\,|\} > \mathrm{const} > 0 \quad (m \neq n). \tag{110}$$

We are now left with a problem of the following type. Taking a recurrence of the form

$$\sum_{j=0}^{l} a_j(x;\epsilon)y_{k+j} = 0 \tag{111}$$

under conditions

$$\tilde{a}_j(x;\epsilon) \sim \sum_{k\geq 0}\mathcal{F}_{j;k}(x^{\frac{1}{2}-\frac{3}{2}k})\epsilon^k, \tag{112}$$

$$\inf_{E_\beta}\left\{|a_0(k\epsilon)|, |a_{l-1}(k\epsilon)|, \frac{1}{|a_j(k\epsilon)|}\right\} > \mathrm{const} > 0, \tag{113}$$

$$\inf_{E_\beta}\{|\,|\lambda_m(x)| - |\lambda_n(x)|\,| > \mathrm{const} > 0\ (m \neq n)\}, \tag{114}$$

where $\lambda_m(x)$ are the roots of the polynomial

$$P(\lambda) := \sum_{j=0}^{l} a_{j,0}(k\epsilon)\lambda^j = 0, \tag{115}$$

we want to show the following.

LEMMA 4.4. *Given a formal solution to* (111)

$$\exp\left(\epsilon^{-1}\sum_{t=0}^{\infty}\epsilon^{t}\Phi_t(x)\right),$$

where $\Phi_t(x)=\mathcal{F}_t(x^{\frac{3}{2}(1-t)})$, *there is a solution asymptotic to it for* $0<x\in E_\beta$ (*and correspondingly, one when* x *is negative*).

Proof. We prove Lemma 4.4 by using induction on l.

(a) We show that we can find a solution corresponding to the root that has the largest modulus (this will simultaneously prove the lemma for $l=1$). All the arguments in §4.2.1 above apply here. Actually, now we could get some better estimates since we do not have small denominators in (106), (107), and in the estimates of the matrices, but this would not affect the final result.

(b) We assume that the conclusion is true for all recurrences of order less than $l-1$ and show it holds for recurrences of order l. By the arguments above, there is a true solution asymptotic to the formal solution defined by the maximum eigenvalue. Using it to reduce the order of the recurrence, we obtain an order $l-1$ scheme, which satisfies the conditions of Lemma 4.4 as it is easy to check and for which we thus know the asymptotic behavior of the solutions. It remains to verify that they can be used to produce solutions of the higher-order recurrence with the stated asymptotic behavior. For definiteness we study the subregion $x<-\epsilon^{\beta}$. All the arguments in Step 3 of §3 apply if we take k_1 to be $-\epsilon^{\beta}$. The only change is that in (75) $\Phi_{m;1}(x)$ are not uniformly bounded. Instead, using Lemma 4.3 we get $|\Phi_{j;1}|=\mathcal{F}(\ln(|x|))<K|\ln\epsilon|$ for some fixed constant K so the right-hand side of (75) changes to $Y_{m;k}\left(1+O(\epsilon^{s-2-K})\right)$.

Finally, after the first reduction we end up with a recurrence that is nondegenerate in the sense of Lemma 4.4 and for which we can control the small-ϵ behavior of the solutions. Now, the reconstruction of the solutions of the original recurrence from the solutions of the reduced one amounts to merely repeating, without any significant change, the construction and estimates in part (b) above. At this point in the proof, it is clear that if the crossing roots are not the largest, one can reduce the order of the recurrence to the actual level at which the roots cross and then apply the arguments above.

REFERENCES

[1] P. DEIFT AND K. MCLAUGHLIN, *A continuum limit of the Toda lattice*, submitted to Amer. Math. Soc. Monogr. Ser.

[2] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, John Wiley, New York, 1962.

[3] A. ERDELYI, *Higher Transcendental Functions*, McGraw–Hill, New York, 1953.

[4] E. BOREL, *Leçons sur les sèries divèrgentes*, Gauthier–Villars, Paris, 1928.

[5] R. REMMERT, *Theory of Complex Functions*, Springer-Verlag, Berlin, New York, 1991.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 135–169, January 1996

009

TIME-PERIODIC QUASI-LINEAR REACTION-DIFFUSION EQUATIONS*

MARY M. LEGNER† AND VICTOR L. SHAPIRO†

Abstract. A resonance result for a time-periodic quasi-linear reaction-diffusion system with (possibly) different diffusion coefficients is established. The driving force of the system may depend upon the dependent variables as well as their derivatives. Also a nonresonance result below the common first associated eigenvalue is stated and proved, and an example of an autocatalytic system illustrating this result is presented. Two other examples for the resonance result are also presented.

Key words. reaction-diffusion, time-periodic, quasi-linear partial differential equations, nonlinear driving force

AMS subject classifications. 35K55, 35K35

1. Introduction.

It is the purpose of this paper to present a resonance-type existence theorem for a system of time-periodic quasi-linear parabolic differential equations with a driving force that depends upon the derivatives of the dependent variables. The result obtained appears to be new, even in the semilinear parabolic case. The quasi-linear parabolic system that we study is of the following form with N_1 a positive integer:

$$(1.1)\qquad \begin{aligned} D_t u_i + \rho_i Q_i u_i = \rho_i \lambda_1^* u_i &+ f_i(x,t,\xi_m(u_1),\dots,\xi_m(u_{N_1})) \\ &+ G_{1i}(D_t u_i) G_{2i}(u_i) g_i(x,t), \qquad i = 1,\dots,N_1, \end{aligned}$$

for $(x,t) \in \Omega \times (-\pi,\pi)$, where each $Q_i u$ is a $2m$th-order quasi-linear elliptic differential operator, λ_1^* is its common associated first eigenvalue, $D_t u = \frac{\partial u}{\partial t}, \xi_m(u) = \{D^\alpha u : |\alpha| \le m\}$, and ρ_i is a positive diffusion coefficient. We shall also give a nonresonance-type result for a quasi-linear system closely related to (1.1).

Systems of the form (1.1) play an important role in applied mathematics and are generally referred to in the literature as reaction-diffusion equations. In particular, in mathematical biology when the diffusion coefficients (i.e., the ρ_i) are different they give rise to the Turing mechanism, which is important in the theory of pattern formation (see the interesting article [14] and a further discussion in [15, pp. 375–379]). Also, in predator–prey problems autocatalytic systems (see [2, pp. 116–120]) are important. In §2 we give a quasi-linear autocatalytic system that is covered by Theorem 2 and two other examples that illustrate Theorem 1.

The two theorems that are proven in this paper are motivated by the work of Shapiro [17], de Figueiredo and Gossez [6], Lefton and Shapiro [12], and de Figueiredo [5]. In [17] the notion of a first eigenvalue for a higher-order quasi-linear elliptic operator is introduced (evidently for the first time), and this gives the aforementioned λ_1^*. (See also (1.5) and (1.6).) In [6], techniques for solving semilinear higher-order elliptic boundary value problems below and at the first eigenvalue (called problems at resonance) are developed. In [12], a time-periodic result at resonance for quasi-linear parabolic operators is obtained; and in [5], a resonance theorem for semilinear

* Received by the editors October 1, 1993; accepted for publication (in revised form) April 6, 1994.

† Department of Mathematics, University of California, Riverside, CA 92521.

elliptic partial differential equations with derivatives in the nonlinear forcing term is established.

To be quite explicit about the various aspects of each of the quasi-linear partial differential equations in (1.1), let $\Omega \subset \mathbb{R}^N, N \geq 1$, be a bounded open connected set. The points of Ω will be designated by $x = (x_1, \dots, x_N)$ and the elementary differential operators by $D^\alpha = \prod_{j=1}^N (\frac{\partial}{\partial x_j})^{\alpha_j}$ for an ordered N-tuple $\alpha = (\alpha_1, \dots, \alpha_N)$ of nonnegative integers with the order of the operator D^α written as $|\alpha| = \sum_{j=1}^N \alpha_j$. To write nonlinear partial differential operators in a convenient form, we introduce the vector space $\mathbb{R}^{s_m}$ whose elements are $\xi_m = \{\xi_\alpha : |\alpha| \leq m\}$ and divide each ξ_m into two parts: $\xi_m = (\eta_{m-1}, \zeta_m)$, where $\eta_{m-1} = \{\eta_\beta : |\beta| \leq m-1\} \in \mathbb{R}^{s_{m-1}}$ is the lower order part of ξ_m and $\zeta_m = \{\zeta_\alpha : |\alpha| = m\}$ is the part of ξ_m corresponding to the mth derivatives. In particular, $\zeta_m \in \mathbb{R}^{q_m}$, where $q_m = s_m - s_{m-1}$. For $u \in W_0^{m,2}(\Omega), \xi_m(u)(x) = \{D^\alpha u(x) : |\alpha| \leq m\}$. (Note $D^{(0,0,\dots,0)}u = u$.)

$Q_i u$ will designate a $2m$th-order differential operator of the generalized divergence form

$$Q_i u = \sum_{1 \leq |\alpha| \leq m} (-1)^{|\alpha|} D^\alpha A_\alpha^i(x, \xi_m(u)) + a_0^i(x, \xi_m(u))u, \tag{1.2}$$

and the following hypotheses will be imposed on the functions $A_\alpha^i(x, \xi_m), 1 \leq |\alpha| \leq m$, and a_0^i appearing in (1.2).

(Q-1) $A_\alpha^i(x, \xi_m) : \Omega \times \mathbb{R}^{s_m} \to \mathbb{R}$ satisfies the Carathéodory conditions (i.e., $A_\alpha^i(x, \xi_m)$ is measurable in x $\forall \xi_m \in \mathbb{R}^{s_m}$ and continuous in ξ_m for a.e. $x \in \Omega$ for $1 \leq |\alpha| \leq m$).

(Q-2) $\exists c_0 > 0$ and $h^\# \in L^2(\Omega)$ with $h^\#$ nonnegative such that $|A_\alpha^i(x, \xi_m)| \leq h^\#(x) + c_0|\xi_m|$ $\forall \xi_m \in \mathbb{R}^{s_m}, 1 \leq |\alpha| \leq m$, and a.e. $x \in \Omega$ for $i = 1, \dots, N_1$.

(Q-3) $\sum_{|\alpha|=m} [A_\alpha^i(x, \eta_{m-1}, \zeta_m) - A_\alpha^i(x, \eta_{m-1}, \zeta_m')](\zeta_\alpha - \zeta_\alpha') > 0$ for a.e. $x \in \Omega$, $\forall (\eta_{m-1}, \zeta_m), (\eta_{m-1}, \zeta_m') \in \mathbb{R}^{s_m}$ with $\zeta_m \neq \zeta_m'$, where $\xi_m = (\eta_{m-1}, \zeta_m)$, and for $i = 1, \dots, N_1$.

(Q-4) $\exists c_1 > 0$ such that $\sum_{|\alpha|=m} A_\alpha^i(x, \xi_m)\zeta_\alpha \geq c_1 |\zeta_m|^2$ $\forall \xi_m \in \mathbb{R}^{s_m}$ and a.e. $x \in \Omega$ for $i = 1, \dots, N_1$.

(Q-5) $a_0^i(x, \xi_m) : \Omega \times \mathbb{R}^{s_m} \to \mathbb{R}$ satisfies the Carathéodory conditions and $a_0^i(x, \xi_m) \in L^\infty(\Omega \times \mathbb{R}^{s_m})$.

The notation we have introduced is the standard type used. (See [4].)

Next, we introduce the Hilbert space $\tilde{H}$ as follows. Set $\mathcal{A} = \{v(x,t) \in C^\infty(\Omega \times \mathbb{R}) : v(x,t)$ is real-valued and satisfies ($\mathcal{A}$-1) and ($\mathcal{A}$-2)$\}$, where

($\mathcal{A}$-1) $v(x,t) = v(x, t + 2\pi)$ $\forall x \in \Omega$ and $\forall t \in \mathbb{R}$,

($\mathcal{A}$-2) $\exists E$, a compact subset of Ω, such that $v(x,t) = 0$ $\forall x \in \Omega \backslash E$ and $\forall t \in \mathbb{R}$.

Set $\tilde{\Omega} = \Omega \times T$ where $T = (-\pi, \pi)$, and introduce the following inner product, where we write L^2 for $L^2(\tilde{\Omega})$,

$$\langle u, v \rangle_{\tilde{H}} = \langle D_t u, D_t v \rangle_{L^2} + \sum_{|\alpha| \leq m} \langle D^\alpha u, D^\alpha v \rangle_{L^2} \tag{1.3}$$

for $u, v \in \mathcal{A}$, where $\langle \cdot, \cdot \rangle_{L^2}$ is the usual L^2-inner product over $\tilde{\Omega}$. With this inner product $\mathcal{A}$ becomes a pre-Hilbert space. $\tilde{H}$ will designate the Hilbert space that we obtain by completing $\mathcal{A}$ using Cauchy sequences with respect to the norm $\|u\|_{\tilde{H}} = \langle u, u \rangle_{\tilde{H}}^{\frac{1}{2}}$.

For our quasi-linear operator $Q_i u$ in (1.2) we shall also suppose

(Q-6)

$$\int_{\tilde{\Omega}} \sum_{1\leq|\alpha|\leq m} A^i_\alpha(x,\xi_m(v))D^\alpha D_t v + \int_{\tilde{H}} a^i_0(x,\xi_m(v))vD_tv = 0 \quad \forall v \in \mathcal{A}.$$

In (Q-6), we note that $\xi_m(v) = \{D^\alpha(v(x,t)) : |\alpha| \leq m\}$.

From (1.1), (Q-1), and (Q-2) we see that the two-form

$$(1.4) \qquad Q_i(u,v) = \sum_{1\leq|\alpha|\leq m} \int_{\tilde{\Omega}} A^i_\alpha(x,\xi_m(u))D^\alpha v + \int_{\tilde{\Omega}} a^i_0(x,\xi_m(u))uv$$

is well defined for $u, v \in \tilde{H}$ and $i = 1,\ldots,N_1$.

Motivated by Shapiro [17, p. 1821], we define

$$(1.5) \qquad \lambda^*_{1i} = \liminf_{\|u\|_{L^2}\to\infty} \frac{Q_i(u,u)}{\|u\|^2_{L^2}}, \quad u \in \tilde{H}.$$

The same proof used in [17, p. 1822] shows that λ^*_{1i} is a finite real-valued number. λ^*_{1i} plays the role of the first (or principal) eigenvalue of Q_i. We shall suppose throughout this paper that

$$(1.6) \qquad \lambda^*_{11} = \cdots = \lambda^*_{1N_1} = \lambda^*_1.$$

Next we set

$$(1.7) \qquad Lu = \sum_{|\alpha|,|\beta|\leq m} (-1)^{|\alpha|}D^\alpha(b_{\alpha\beta}(x)D^\beta u),$$

where we have the following hypotheses on $b_{\alpha\beta}(x)$.

(L-1) $b_{\alpha\beta}$ are real-valued functions in $L^\infty(\Omega)$ for $|\alpha|, |\beta| \leq m$. Also, if $|\alpha| = |\beta| = m$, then $b_{\alpha\beta}$ is uniformly continuous in Ω.

(L-2) $\exists c_2$ such that $\sum_{|\alpha|=|\beta|=m} b_{\alpha\beta}\mu^\alpha\mu^\beta \geq c_2|\mu|^{2m}$ $\forall x \in \Omega$ and $\forall \mu \in \mathbb{R}^N$ where $\mu = (\mu_1,\ldots,\mu_N)$ and $|\mu|^2 = \mu_1^2 + \cdots + \mu_N^2$.

Observe that with (L-1) and (L-2), L is a higher-order linear elliptic operator. Associated to L is the bilinear form

$$(1.8) \qquad \mathcal{L}_\Omega(u,v) = \sum_{|\alpha|,|\beta|\leq m} \int_\Omega b_{\alpha\beta}(x)D^\alpha uD^\beta v \quad \forall u,v \in W^{m,2}_0(\Omega)$$

where $W^{m,2}_0(\Omega)$ is the well-known Sobolev space obtained by taking the closure of $C^m_0(\Omega)$ in $W^{m,2}(\Omega)$ with inner product $\langle u,v\rangle = \int_\Omega \sum_{|\alpha|\leq m} D^\alpha uD^\beta v$.

We will also assume throughout this paper that L satisfies (L-1) and (L-2) as well as the following conditions.

(L-3) L is symmetric; that is, $\mathcal{L}_\Omega(u,v) = \mathcal{L}_\Omega(v,u)$ $\forall u, v \in W^{m,2}_0(\Omega)$.

Also, we have the following well-known facts about L (see [8] and [18, p. 858]):

$\exists\{\lambda_n\}^\infty_{n=1}$ a nondecreasing sequence with $\lambda_n \underset{n\to\infty}{\longrightarrow} \infty$ and a

(1.9) sequence of real-valued functions $\{\phi_n\}^\infty_{n=1}$ in $W^{m,2}_0(\Omega)$ such that

$\mathcal{L}_\Omega(u,\phi_n) = \lambda_n\langle u,\phi_n\rangle_{L^2(\Omega)}$ $\forall u \in W^{m,2}_0(\Omega)$;

$$\{\phi_n\}_{n=1}^{\infty} \text{ is a } CONS \text{ on } L^2(\Omega); \tag{1.10}$$

$$\lambda_1 = \inf_{u\neq 0} \frac{\mathcal{L}_\Omega(u,u)}{\|u\|_{L^2(\Omega)}}, \quad u \in W_0^{m,2}(\Omega). \tag{1.11}$$

Extending the bilinear form $\mathcal{L}_\Omega(\cdot,\cdot)$ from $W_0^{m,2}(\Omega)$ to $\tilde{H}$ by

$$\mathcal{L}(u,v) = \int_T \mathcal{L}_\Omega(u,v) = \int_{\tilde{\Omega}} \sum_{|\alpha|,|\beta|\leq m} b_{\alpha\beta} D^\alpha u D^\beta v \tag{1.12}$$

for $u, v \in \tilde{H}$, where $T = (-\pi, \pi)$, we see from (1.11) that

$$\lambda_1 = \inf_{u\neq 0} \frac{\mathcal{L}(u,u)}{\|u\|_{L^2}}, \quad u \in \tilde{H}. \tag{1.13}$$

Also, we set $\|u\|_{\mathcal{L}}^2 = |\mathcal{L}(u,u)|$ for $u \in \tilde{H}$ and

$$\|u\|^2_{W^{m,2,\sim}} = \int_{\tilde{\Omega}} |\xi_m(u)|^2. \tag{1.14}$$

Let κ be given with $0 \leq \kappa < 1$.

We shall say Q_i is $\kappa*$-related to L if the following conditions hold.

$$\lambda^*_{1i} = \lambda_1. \tag{1.15}$$

$$\begin{aligned} &\text{(i)} \quad u \in \tilde{H}, \liminf_{\|u\|_{W^{m,2,\sim}}\to\infty} \frac{Q_i(u,u) - \mathcal{L}(u,u)}{\|u\|^{1+\kappa}_{W^{m,2,\sim}}} \geq 0. \\ &\text{(ii)} \quad \text{There exist positive constants } c_1^\#, c_2^\# \text{ such that} \\ &\qquad Q_i(u,u) - \mathcal{L}(u,u) \geq -c_1^\# \|u\|^{\kappa+1}_{W^{m,2,\sim}} - c_2^\# \ \forall u \in \tilde{H}. \end{aligned} \tag{1.16}$$

(Q_i being $\kappa*$-related to L comes from the notion introduced in [12, p. 153] where in (1.19) in this last-named reference, $\|u\|_{\mathcal{L}}$ should be replaced by $\|u\|_{W^{m,2,\sim}}$. In case $\lambda_1 > 0$, $\|u\|_{\mathcal{L}}$ and $\|u\|_{W^{m,2,\sim}}$ are equivalent norms.)

Let n_1 be the largest integer such that $\lambda_1 = \cdots = \lambda_{n_1} < \lambda_{n_1+1}$. This implies that

$$\mathcal{L}_\Omega(v,\phi_i) = \lambda_1\langle v,\phi_i\rangle_{L^2(\Omega)} \quad \forall v \in W_0^{m,2}(\Omega) \text{ for } i = 1,\ldots,n_1. \tag{1.17}$$

We set

$$S^\#_{\lambda_1} = \left\{ v = \sum_{j=1}^{n_1} d_j\phi_j : d_j \text{ is a real constant } j = 1,\ldots,n_1 \right\}. \tag{1.18}$$

We shall impose the following hypotheses upon the f_i given in (1.1) for $i = 1,\ldots,N_1$.

$$f_i(x,t,\xi_{m1},\ldots,\xi_{mN_1}) : \tilde{\Omega} \times \underbrace{\mathbb{R}^{s_m} \times \cdots \times \mathbb{R}^{s_m}}_{N_1} \to \mathbb{R} \tag{f-1}$$

satisfies the Carathéodory conditions.

(f-2) $\exists\kappa$ with $0 \le \kappa < 1$ and $\exists$ nonnegative constants c_α and an L^2-function $d(x,t)$ such that

$$|f_i(x,t,\xi_{m1},\dots,\xi_{mN_1})| \le \sum_{|\alpha|\le m} c_\alpha|\xi_{\alpha i}|^\kappa + d(x,t) \;\; \forall(\xi_{m1},\dots,\xi_{mN_1}) \in \underbrace{\mathbb{R}^{s_m}\times\cdots\times\mathbb{R}^{s_m}}_{N_1},$$

a.e. in $\tilde{\Omega}$ for $i = 1,\dots,N_1$ where $\xi_{mi} = \{\xi_{\alpha i}\}_{|\alpha|\le m}$.

(f-3) Let $S_1' = \{\xi_m \in \mathbb{R}^{s_m} : |\xi_m| = 1 \text{ and } \xi_{(0,\dots,0)} \ne 0\}$. $\exists\tilde{\Omega}_0 \subset \tilde{\Omega}$ with meas $(\tilde{\Omega} - \tilde{\Omega}_0) = 0$ such that the following uniform limit holds for each $(x,t) \in \tilde{\Omega}_0$:

$$\text{(1.19)}\quad \begin{aligned} &\lim_{n\to\infty} \frac{f_i(x,t,\xi_{m1},\dots,r_n\xi_{mi}^{(n)},\dots,\xi_{mN_1})}{r_n^\kappa} = h_i(x,t,\xi_{mi}) \\ &\text{uniformly for } (\xi_{m1},\dots,\xi_{m(i-1)},\xi_{m(i+1)},\dots,\xi_{mN_1}) \in \underbrace{\mathbb{R}^{s_m}\times\cdots\times\mathbb{R}^{s_m}}_{N_1-1} \end{aligned}$$

for all sequences $\{r_n\} \subset \mathbb{R}$ and $\{\xi_{mi}^{(n)}\} \subset S_1'$ such that $r_n \to \infty$ and $\xi_{mi}^{(n)} \to \xi_{mi} \in S_1'$, where $h_i(x,t,\xi_{mi})$ is a Carathéodory function on $\tilde{\Omega} \times S_1'$ that is uniformly bounded by a function in $L^2(\tilde{\Omega})$.

For the G's on the right-hand side of (1.1) we make the following assumptions for $i = 1,\dots,N_1$.

(G-1) $G_{1i} : L^2(\tilde{\Omega}) \to \mathbb{R}$ is sequentially weakly continuous and $\exists\kappa$ with $0 \le \kappa < 1$ such that $|G_{1i}(u)| \le c_1'\|u\|_{L^2}^\kappa + c_2'$ $\forall u \in L^2(\tilde{\Omega})$, where c_1' and c_2' are nonnegative constants.

(G-2) $G_{2i} : L^2(\tilde{\Omega}) \to \mathbb{R}$ is strongly continuous and $\exists c_3'$ such that $|G_{2i}(u)| \le c_3'$ $\forall u \in L^2(\tilde{\Omega})$. Also in case $\kappa = 0$ in either Theorem 1 or Theorem 2, $\lim_{\|u\|_{L^2}\to\infty} |G_{2i}(u)| = 0$.

We shall establish the following resonance theorem for the time-periodic quasi-linear parabolic differential system (1.1).

THEOREM 1. *Let $\Omega \subset \mathbb{R}^N, N \ge 1$, be a bounded open connected set; and let κ be given with $0 \le \kappa < 1$. Suppose that the coefficients of Q_i satisfy* (Q-1)–(Q-6); *Q_i is $\kappa*$-related to L; f_i satisfies* (f-1)–(f-3); *$g_i \in L^2(\tilde{\Omega})$; G_{1i} and G_{2i} satisfy* (G-1) *and* (G-2), *respectively; and $\rho_i > 0$ for $i = 1,\dots,N_1$. Also suppose*

$$\text{(1.20)}\quad \int_{\tilde{\Omega}(v)} h_i\left(x,t,\frac{\xi_m(v)}{|\xi_m(v)|}\right)|\xi_m(v)|^\kappa v < 0 \quad \textit{for every nontrivial } v \in S_{\lambda_1}^\#$$

for $i = 1,\dots,N_1$, where h_i is given in (f-3) *and $\tilde{\Omega}(v) = \{(x,t) \in \tilde{\Omega} : v(x,t) \ne 0\}$. Then $\exists(u_1,\dots,u_{N_1}) \in \tilde{H}^{N_1}$, which is a weak solution to the system* (1.1) *on $\tilde{\Omega}$.*

What we mean by $(u_1,\dots,u_{N_1}) \in \tilde{H}^{N_1}$ is a weak solution to the system (1.1) on $\tilde{\Omega}$ is that

$$\text{(1.21)}\quad \begin{aligned} \langle D_tu_i,v\rangle_{L^2} + \rho_iQ_i(u_i,v) = \rho_i\lambda_1^*\langle u_i,v\rangle_{L^2} &+ \langle f_i(\cdot,\xi_m(u_1),\dots,\xi_m(u_{N_1})),v\rangle_{L^2} \\ &+ G_{1i}(D_tu_i)G_{2i}(u_i)\langle g_i,v\rangle_{L^2} \end{aligned}$$

$\forall v \in \tilde{H}$ and $i = 1,\dots,N_1$.

The nonresonance-type theorem that we shall prove will deal with the system

$$\text{(1.22)}\quad \begin{aligned} D_tu_i + \rho_iQ_iu_i = \rho_i(\lambda_1^* - \delta_i)u_i &+ f_i(x,t,\xi_m(u_1),\dots,\xi_m(u_{N_1})) \\ &+ G_{1i}(D_tu_i)G_{2i}(u_i)g_i(x,t), \quad i = 1,\dots,N_1, \end{aligned}$$

where each δ_i is a positive constant. To handle the system (1.22) we do not need the condition (f-3) nor the condition (1.20). In particular, we shall prove the following theorem.

THEOREM 2. *Suppose the conditions in the hypothesis of Theorem* 1 *hold, except for* (f-3) *and the condition in* (1.20). *Then* $\exists(u_1,\ldots,u_{N_1}) \in \tilde{H}^{N_1}$, *which is a weak solution to the system* (1.22).

In Theorems 1 and 2, besides dealing with a system, we also allow derivatives of the dependent variables to appear in the forcing terms on the right-hand side of the equations. Also, we allow derivatives of order m as well as derivatives of lower order of the dependent variable to appear in the coefficients of each of the Q_i on the left-hand side of the equations. Neither of these situations were handled in the paper by Lefton and Shapiro [12], and some new ideas of the type presented in the current paper are needed. In the next section, we give examples of Q's and f's that are relevant for Theorems 1 and 2 but are not covered in [12]. The extension of our results to almost-periodic reaction-diffusion equations is left as an open problem.

2. Some examples. If L satisfies (L-1)–(L-3) and $m = 1$, then L is automatically $\kappa*$-related to itself. For $m \geq 2$, if L also satisfies (L-4) below, then L also is $\kappa*$-related to itself; this is the first situation we deal with in this section. Next, we give two examples of a Q that has derivatives of order lower than m involved in its coefficients (as well as of order m) and a corresponding L to which Q is $\kappa*$-related. Finally we consider three different systems, the first of which is autocatalytic and involves Theorem 2 and the second and third of which illustrate condition (1.20) in Theorem 1.

We first show that if Lu is given by (1.7) and L meets (L-1)–(L-3), then

$$\mathcal{L}(v, D_t v) = 0 \quad \forall v \in \mathcal{A}, \tag{2.1}$$

where $\mathcal{L}(\cdot,\cdot)$ is defined by (1.12).

To see that (2.1) is true, we observe from (L-3) that for $v \in \mathcal{A}, \mathcal{L}(v, D_t v) = \mathcal{L}(D_t v, v)$. Also, we see that $D_t[D^\beta v D^\alpha v] = (D^\alpha D_t v)D^\beta v + (D^\alpha v)D^\beta D_t v$. Hence from (1.12)

$$\begin{aligned} 2\mathcal{L}(v, D_t v) &= \sum_{|\alpha|,|\beta|\leq m} \int_{\tilde{\Omega}} b_{\alpha\beta}[(D^\alpha v)(D^\beta D_t v) + (D^\alpha D_t v)(D^\beta v)] \\ &= \sum_{|\alpha|,|\beta|\leq m} \int_{\Omega} b_{\alpha\beta}(x) \int_T D_t[D^\beta v D^\alpha v] \\ &= 0, \end{aligned}$$

which establishes (2.1). Next, we say L meets (L-4) if

(L-4)

$$\sum_{|\alpha|=|\beta|=m} b_{\alpha\beta}(x)(\zeta_\beta - \zeta'_\beta)(\zeta_\alpha - \zeta'_\alpha) \geq c'_1 \sum_{|\alpha|=m} (\zeta_\alpha - \zeta'_\alpha)^2 \quad \forall \zeta_m, \zeta'_m \in \mathbb{R}^{q_m}$$

where c'_1 is a positive constant and $q_m = s_m - s_{m-1}$.

(In the special case $m = 1$, it is easy to see that (L-2) implies (L-4).)

We now show if

$$\begin{aligned} Lu = &\sum_{|\alpha|=|\beta|=m} (-1)^{|\alpha|} D^\alpha[b_{\alpha\beta}(x)D^\beta u] \\ &+ \sum_{1\leq|\alpha|,|\beta|\leq m-1} (-1)^{|\alpha|} D^\alpha[b_{\alpha\beta}(x)D^\beta u] + b_{00}(x)u \end{aligned} \tag{2.2}$$

and

(2.3) if L meets (L-1)–(L-4), then L is $\kappa*$-related to itself.

In order to establish (2.3), what we have to do is show that L given by (2.2) satisfies (Q-1)–(Q-6). To do this we define

$$\begin{aligned} A_\alpha(x,\xi_m) &= \sum_{|\beta|=m} b_{\alpha\beta}(x)\zeta_\beta \quad \text{for } |\alpha| = m, \\ &= \sum_{1\le|\beta|\le m-1} b_{\alpha\beta}(x)\eta_\beta \quad \text{for } 1 \le |\alpha| \le m-1, \\ a_0(x,\xi_m) &= b_{00}(x). \end{aligned} \tag{2.4}$$

It is clear from (2.4) that $A_\alpha(x,\xi_m)$ satisfies (Q-1), (Q-2), (Q-5), and (Q-6); this latter fact holds because of (2.1). Taking $\zeta'_m = 0$ in (L-4), we see from (2.4) that

$$\sum_{|\alpha|=m} A_\alpha(x,\xi_m)\zeta_\alpha \ge c'_1 \sum_{|\beta|=m} |\zeta_\alpha|^2;$$

so (Q-4) holds. It is also clear from (L-4) that (Q-3) holds. Hence L is indeed $\kappa*$-related to itself since (1.15) and (1.16) obviously hold. So Theorems 1 and 2 hold when Q_i is an L of the form (2.2) satisfying (L-1)–(L-4). If all the Q_i's are L's of this nature, our quasi-linear result is actually a semilinear result and appears to be new even in this case.

For $m = 1$, any Lu of the form (2.2) that also satisfies (L-1)–(L-3) automatically is $\kappa*$-related to itself by (2.3) since (L-2) then implies (L-4) as we noted earlier. For $m \ge 2$, suppose Lu of the form (2.2) satisfies (L-1) and in addition

$$\begin{aligned} &b_{\alpha\beta}(x) = 0 \quad \text{if } |\alpha| = |\beta| = m \text{ and } \alpha \ne \beta, \\ &b_{\alpha\alpha}(x) \ge c'_2 \quad \forall x \in \Omega \text{ if } |\alpha| = m, \\ &b_{\alpha\beta}(x) = b_{\beta\alpha}(x) \quad \text{if } 1 \le |\alpha|, |\beta| \le m-1, \end{aligned} \tag{2.5}$$

where c'_2 is a positive constant. Then it is easy to see from [9, p. 55] that (L-2), (L-3), and (L-4) also hold. Hence by (2.3) such an L is $\kappa*$-related to itself. In particular, we observe that $Lu = (-1)^m\Delta^m u$, where Δ^m is the familiar mth-iterated Laplacian, satisfies (2.5) and, therefore, is $\kappa*$-related to itself.

In [12, pp. 154–160], two quasi-linear examples, one of second order and one of higher order, were given that meet the criteria for the $\kappa*$-relationship. We now give some quasi-linear examples that are not covered by the theorems in [12] because $A_\alpha(x,\xi_m)$ for $|\alpha| = m$ will be allowed to depend on derivatives or order $m-1$. However, these examples will be covered by Theorems 1 and 2.

We first consider the case when Qu is a second-order quasi-linear differential expression, that is, $m = 1$ in (1.2). For this situation, we can designate the A_α when $|\alpha| = 1$ by A_j for $j = 1, \ldots, N$ and ξ_1 by $(\eta_0, \zeta_1, \ldots, \zeta_N)$ with $|\xi_1|^2 = \eta_0^2 + \zeta_1^2 + \cdots + \zeta_N^2$; so $\eta_0(u) = u$ and $\zeta_j(u) = \frac{\partial u}{\partial x_j}$. In particular, we set

$$\begin{aligned} A_j(x,\xi_1) &= \zeta_j + \frac{\zeta_j}{(1+|\xi_1|^2)^{\frac{1}{2}}}, \qquad j = 1, \ldots, N; \\ a_0(x,\xi_1) &= (1+|\xi_1|^2)^{-\frac{1}{2}}. \end{aligned} \tag{2.6}$$

It is clear from the start that (2.6) implies that (Q-1), (Q-2), (Q-4), and (Q-5) are automatically satisfied. To see that (Q-3) holds, we observe that for $c > 0$, $[c^2 + \zeta_1^2 + \cdots + \zeta_N^2]^{\frac{1}{2}}$ is a convex function in $\mathbb{R}^N$. Hence by [11, p. 16], the gradient of this last-named function is monotone. This fact plus an easy computation show that (Q-3) holds for the A_j in (2.6).

To see that (Q-6) holds we observe from (2.6) that for $v \in \mathcal{A}$

$$\begin{aligned}(2.7)\qquad &\sum_{j=1}^{N} A_j(x,\xi_1(v))D_tD_jv + a_0(x,\xi_1(v))vD_tv \\ &\quad = D_t[1+|v|^2+|\nabla v|^2]^{\frac{1}{2}} + \frac{1}{2}D_t|\nabla v|^2.\end{aligned}$$

Hence, the periodicity of v shows that the integral over $\tilde{\Omega}$ of the expression on the left-hand side of the equal sign in (2.7) is zero and (Q-6) is established.

We set $Lu = -\Delta u$ where Δ stands for the Laplace operator. To complete this example we must show that Qu given by (2.6) is $\kappa *$-related to L. An easy computation using (1.4) and (2.6) shows that

$$(2.8)\qquad \mathcal{Q}(u,u) - \mathcal{L}(u,u) = \int_{\tilde{\Omega}} \frac{|\nabla u|^2 + u^2}{(1+u^2+|\nabla u|^2)^{\frac{1}{2}}}.$$

Since the right-hand side of (2.8) is nonnegative, it follows that conditions (1.16) hold. In a similar manner it follows from (1.5), (1.13), and (2.8) that $\lambda_1^* \geq \lambda_1$. Next, with ϕ_1 given in (1.9), we have that $\mathcal{L}(n\phi_1, n\phi_1) = n^2\lambda_1 2\pi$. Consequently, we obtain from (2.8) that

$$\frac{\mathcal{Q}(n\phi_1,n\phi_1)}{\langle n\phi_1,n\phi_1\rangle_{L^2}} \leq \lambda_1 + \int_\Omega \frac{[|\nabla\phi_1|^2+\phi_1^2]^{\frac{1}{2}}}{n}.$$

Taking the lim inf of both sides of this last inequality as $n \to \infty$, we infer from (1.5) that $\lambda_1^* \leq \lambda_1$. Hence $\lambda_1^* = \lambda_1$, (1.15) is established, Q is $\kappa *$-related to L, and our example is complete.

For our next example, we let $m \geq 2$ and let $\alpha = (\alpha_1,\ldots,\alpha_N), \beta = (\beta_1,\ldots,\beta_N)$ designate N-tuples or orders $(m-1)$ and m, respectively. $i\alpha$ will stand for the N-tuple β where $\beta_j = \alpha_j$ for $j \neq i$ and $\beta_i = \alpha_i + 1$. Also, for each β with $|\beta| = m$, let $F_\beta(s)$ be a mapping of the following nature:

$$(2.9)\qquad F_\beta : [0,\infty) \to \mathbb{R} \text{ is continuous and nondecreasing;}$$

$$(2.10)\qquad 0 \leq F_\beta(s) \leq 1 \quad \forall s \in [0,\infty);$$

$$(2.11)\qquad \lim_{s\to\infty} s[1-F_\beta(s)] = 0.$$

Examples of such an $F_\beta(s)$ are

$$\frac{s}{(1+s^2)^{\frac{1}{2}}},\ \frac{s^2}{1+s^2},\quad \text{and}\quad \left[\frac{1+s^2}{2+s^2}\right]^{\frac{1}{2}}.$$

We shall also suppose

$$b_\beta(x) \in C(\bar{\Omega}) \quad \text{with } b_\beta(x) \geq c_1' \ \forall x \in \Omega, \tag{2.12}$$

where c_1' is a positive constant. Next, we set

$$\begin{aligned} &\text{(i)} \quad \Gamma(\beta) = \{\alpha : |\alpha| = m-1 \text{ and } \exists i \text{ such that } i\alpha = \beta\}, \\ &\text{(ii)} \quad \nu(\beta) = \text{ the number of } \alpha \text{ in } \Gamma(\beta). \end{aligned} \tag{2.13}$$

We define

$$A_\beta(x, \xi_m) = \xi_\beta + b_\beta(x) \sum_{\alpha \in \Gamma(\beta)} F_\beta[(\xi_\alpha^2 + \xi_\beta^2)^{\frac{1}{2}}]\xi_\beta \quad \text{for } |\beta| = m, \tag{2.14}$$

$$A_\alpha(x, \xi_m) = \sum_{i=1}^{N} b_{i\alpha}(x) F_{i\alpha}[(\xi_\alpha^2 + \xi_{i\alpha}^2)^{\frac{1}{2}}]\xi_\alpha \quad \text{for } |\alpha| = m-1, \tag{2.15}$$

and

$$Qu = \sum_{|\beta|=m} (-1)^m D^\beta A_\beta(x, \xi_m(u)) + \sum_{|\alpha|=m-1} (-1)^{m-1} D^\alpha A_\alpha(x, \xi_m(u)). \tag{2.16}$$

Now, it is clear that the coefficients Q satisfy (Q-1), (Q-2), (Q-4), and (Q-5). From (2.9), we see that $sF_\beta[(s^2 + c^2)^{\frac{1}{2}}]$ is a nondecreasing function of s for $s \in \mathbb{R}$ and $c \in \mathbb{R}$. Hence it follows from (2.12) and (2.14) that the condition (Q-3) is also satisfied. To see that the condition (Q-6) holds, we observe from (2.13) that the following equality of sets holds:

$$\{(\beta, \alpha) : |\beta| = m, \alpha \in \Gamma(\beta)\} = \{(i\alpha, \alpha) : |\alpha| = m-1, i = 1, \ldots, N\}. \tag{2.17}$$

From (2.17), it consequently follows that for $v \in \mathcal{A}$

$$\begin{aligned} &\sum_{|\beta|=m} \sum_{\alpha \in \Gamma(\beta)} b_\beta F_\beta \left\{ [(D^\alpha v)^2 + (D^\beta v)^2]^{\frac{1}{2}} \right\} D^\beta v D^\beta D_t v \\ &\quad = \sum_{|\alpha|=m-1} \sum_{i=1}^{N} b_{i\alpha} F_{i\alpha} \left\{ [(D^\alpha v)^2 + (D_i D^\alpha v)^2]^{\frac{1}{2}} \right\} D_i D^\alpha v D_i D^\alpha D_t v. \end{aligned} \tag{2.18}$$

It therefore follows from (2.14)–(2.16) and this last fact that

$$\begin{aligned} 2Q(v, D_t v) = &\sum_{|\alpha|=m-1} \sum_{i=1}^{N} \int_\Omega b_{i\alpha} \int_T F_{i\alpha} \left\{ [(D^\alpha v)^2 + (D_i D^\alpha v)^2]^{\frac{1}{2}} \right\} \\ &\times 2[D_i D^\alpha v D_i D^\alpha D_t v + D^\alpha v D^\alpha D_t v]. \end{aligned} \tag{2.19}$$

Set $G_{i\alpha}(s) = \int_0^s F_{i\alpha}(r^{\frac{1}{2}})\,dr$ for $s \geq 0$. Then it is clear that $D_t G_{i\alpha}[(D^\alpha v)^2 + (D_i D^\alpha v)^2] =$ the integrand for $\int_T$ in (2.19). Hence the periodicity of v implies that $Q(v, D_t v) = 0$ for $v \in \mathcal{A}$, and we see that the condition (Q-6) does indeed hold for Q.

Next, we set

$$L_1 u = (-1)^m \sum_{|\beta|=m} D^\beta[D^\beta u] \tag{2.20}$$

and with $\nu(\beta)$ defined in (2.13) (ii),

$$\begin{aligned} Lu = L_1 u &+ (-1)^m \sum_{|\beta|=m} \nu(\beta) D^\beta[b_\beta D^\beta u] \\ &+ (-1)^{m-1} \sum_{|\alpha|=m-1} \sum_{i=1}^{N} D^\alpha[b_{i\alpha} D^\alpha u]. \end{aligned} \tag{2.21}$$

Consequently, it follows from (2.13)–(2.16), (2.20), and (2.21) that for $u \in \tilde{H}$,

$$\begin{aligned} &\mathcal{L}(u,u) - \mathcal{Q}(u,u) \\ &= \sum_{|\beta|=m} \sum_{\alpha \in \Gamma(\beta)} \int_{\tilde{\Omega}} \{1 - F_\beta[(|D^\alpha u|^2 + |D^\beta u|^2)^{\frac{1}{2}}]\} b_\beta |D^\beta u|^2 \\ &+ \sum_{|\alpha|=m-1} \sum_{i=1}^{N} \int_{\tilde{\Omega}} \{1 - F_{i\alpha}[(|D^\alpha u|^2 + |D_i D^\alpha u|^2)^{\frac{1}{2}}]\} b_{i\alpha} |D^\alpha u|^2. \end{aligned} \tag{2.22}$$

Our example will be complete if we can show Q is $\kappa*$-related to L. To accomplish this we first observe that

$$u \in \tilde{H}, \quad \lim_{\|u\|_{W^{m,2,\sim}} \to \infty} \int_{\tilde{\Omega}} \frac{\left\{1 - F_\beta\left[(|D^\alpha u|^2 + |D^\beta u|^2)^{\frac{1}{2}}\right]\right\} |D^\beta u|^2}{\|u\|^{1+\kappa}_{W^{m,2,\sim}}} = 0, \tag{2.23}$$

where $\|u\|^2_{W^{m,2,\sim}}$ is defined in (1.14). Equation (2.23) follows easily from the fact that given $\varepsilon > 0$, there is an s_0 by (2.11) such that the numerator of the integrand in (2.23) is majorized by $\varepsilon|D^\beta u|$ if $|D^\beta u| \geq s_0$. Consequently, using (2.10) and (2.12) joined with (2.23), we see that each of the integrals in the first term on the right-hand side of (2.22) when divided by $\|u\|^{1+\kappa}_{W^{m,2,\sim}}$ tends to zero as $\|u\|_{W^{m,2,\sim}} \to \infty$. A similar situation prevails for each of the integrals in the second term. We conclude that

$$\frac{\mathcal{Q}(u,u) - \mathcal{L}(u,u)}{\|u\|^{1+\kappa}_{W^{m,2,\sim}}} \to 0 \quad \text{as } \|u\|_{W^{m,2,\sim}} \to \infty.$$

Hence (1.16) (i) holds. A similar argument using (2.22) shows that $\exists K > 0$ such that $\mathcal{L}(u,u) - \mathcal{Q}(u,u) \leq \|u\|_{W^{m,2,\sim}} + K$ $\forall u \in \tilde{H}$. Hence (1.16) (ii) holds.

It remains to show that $\lambda_1^* = \lambda_1$. From (2.10), (2.16), and (2.22) we see that $0 \leq \mathcal{Q}(u,u) \leq \mathcal{L}(u,u)$ $\forall u \in \tilde{H}$. Hence for ϕ_1 given in (1.10),

$$\frac{Q(n\phi_1, n\phi_1)}{\|n\phi_1\|^2_{L^2}} \leq \lambda_1 \quad \text{for every positive integer } n.$$

Consequently, from (1.5), we have that

$$0 \leq \lambda_1^* \leq \lambda_1. \tag{2.24}$$

Next, we set

$$\mathcal{L}_1(u,u) = \sum_{|\beta|=m} \int_{\tilde{\Omega}} |D^\beta u|^2 \tag{2.25}$$

and observe from (2.10), (2.12), and (2.14)-(2.16) that

$$\mathcal{L}_1(u,u) \le \mathcal{Q}(u,u) \quad \forall u \in \tilde{H}. \tag{2.26}$$

Also, we see from Poincaré's inequality and (2.25) that there is a positive constant c_2' such that

$$\sum_{|\alpha|=m-1} \int_{\tilde{\Omega}} |D^\alpha u|^2 \le c_2' \mathcal{L}_1(u,u) \quad \forall u \in \tilde{H}. \tag{2.27}$$

Let $\{u_n\}_{n=1}^\infty$ be a sequence of elements in $\tilde{H}$ with the property that

$$\|u_n\|_{L^2} \to \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{\mathcal{Q}(u_n,u_n)}{\|u_n\|_{L^2}^2} = \lambda_1^*. \tag{2.28}$$

Hence, for $n \ge n_2$, $\mathcal{Q}(u_n,u_n) \le (\lambda_1^* + 1)\|u_n\|_{L^2}^2$. Consequently, it follows from (2.25) and this last inequality that $\mathcal{L}_1(u_n,u_n) \le (\lambda_1^*+1)\|u_n\|_{L^2}^2$ for $n \ge n_2$. Therefore, it follows that for $|\beta| = m$

$$\begin{aligned} &\int_{\tilde{\Omega}} \frac{\left\{1 - F_\beta\left[(|D^\alpha u_n|^2 + |D^\beta u_n|^2)^{\frac{1}{2}}\right]\right\} |D^\beta u_n|^2}{\|u_n\|_{L^2}^2} \\ &\quad \le (\lambda_1^*+1) \int_{\tilde{\Omega}} \frac{\left\{1 - F_\beta\left[(|D^\alpha u_n|^2 + |D^\beta u_n|^2)^{\frac{1}{2}}\right]\right\} |D^\beta u_n|^2}{\mathcal{L}_1(u_n,u_n)} \end{aligned} \tag{2.29}$$

for $n \ge n_2$. Now in a manner similar to that used to establish (2.23), we obtain from (2.25) that the right-hand side of the inequality in (2.29) tends to zero as $n \to \infty$. Therefore, the limit as $n \to \infty$ of the left-hand side of (2.29) is zero. In a similar manner, using (2.27) we obtain that

$$\lim_{n\to\infty} \int_{\tilde{\Omega}} \frac{\left\{1 - F_{i\alpha}\left[(|D^\alpha u_n|^2 + |D_i D^\alpha u_n|^2)^{\frac{1}{2}}\right]\right\} |D^\alpha u_n|^2}{\|u_n\|_{L^2}^2} = 0 \quad \text{for } |\alpha| = m-1.$$

We conclude from (2.22) that

$$\lim_{n\to\infty} \frac{\mathcal{Q}(u_n,u_n) - \mathcal{L}(u_n,u_n)}{\|u_n\|_{L^2}^2} = 0. \tag{2.30}$$

But $(\mathcal{L}(u_n,u_n)/\|u_n\|_{L^2}^2) \ge \lambda_1$ and we obtain from (2.28) and (2.30) that $\lambda_1 \le \lambda_1^*$. This fact coupled with (2.24) give that $\lambda_1 = \lambda_1^*$. Hence Q is $\kappa*$-related to L and our example is complete.

Next we take $N_1 = 2$ in (1.1) and give two examples of an f_1 and f_2 that satisfy the conditions of our theorems. Our first example will be a simple autocatalytic system that qualifies under Theorem 2. For this system we also take $N = 1, \Omega = (0, \pi)$,

$$\begin{aligned}
Q_1(u_1) &= -D_x \left[D_x u_1 + \frac{D_x u_1}{(1 + u_1^2 + D_x u_1^2)^{\frac{1}{2}}} \right] + \frac{u_1}{(1 + u_1^2 + D_x u_1^2)^{\frac{1}{2}}}, \\
Q_2(u_2) &= -\frac{1}{2} D_x \left[1 + \frac{|D_x u_2|^2}{(1 + |D_x u_2|^2)} \right] D_x u_2, \\
Lu &= -D_x(D_x u), \\
f_1(x, t, u_1, u_2) &= -\frac{\arctan u_2}{1 + u_1^2} + b_1(x, t), \\
f_2(x, t, u_1, u_2) &= \frac{\tanh u_1}{1 + u_2^2} + b_2(x, t),
\end{aligned}$$

where $b_1(x,t)$ and $b_2(x,t)$ are in $C\{[0,\pi] \times [0, 2\pi], \mathbb{R}\}$. It follows from our earlier discussion in this section and from [12, pp. 156–157] that Q_1 and Q_2 are $\kappa*$-related to L and that $\lambda_{11}^* = \lambda_{12}^* = \lambda_1 = 1$. Hence with

$$\begin{aligned}
F_1(x, t, u_1, u_2) &= \rho_1(1 - \delta_1)u_1 + f_1(x, t, u_1, u_2), \\
F_2(x, t, u_1, u_2) &= \rho_2(1 - \delta_2)u_2 + f_2(x, t, u_1, u_2)
\end{aligned} \tag{2.31}$$

we see from Theorem 2 with $\kappa = 0$ that there exist $(u_1, u_2) \in \tilde{H}^2$, which is a weak solution to the system

$$\begin{aligned}
\frac{\partial u_1}{\partial t} + \rho_1 Q_1 u_1 &= F_1(x, t, u_1, u_2), \\
\frac{\partial u_2}{\partial t} + \rho_2 Q_2 u_2 &= F_2(x, t, u_1, u_2)
\end{aligned} \tag{2.32}$$

on the interval $(0, \pi) \times (0, 2\pi)$, where ρ_1, ρ_2, δ_1, and δ_2 are positive real numbers. An easy computation shows that $\frac{\partial F_1}{\partial u_2} < 0$ and $\frac{\partial F_2}{\partial u_1} > 0$ for all values of u_1, u_2 in $\mathbb{R}$ and (x, t) in $[0, \pi] \times [0, 2\pi]$. Likewise we see that

$$\frac{\partial F_1}{\partial u_1} = \rho_1(1 - \delta_1) + 2u_1 \frac{\arctan u_2}{(1 + u_1^2)^2}. \tag{2.33}$$

Since the second term on the right-hand side of the equality in (2.33) is bounded in absolute value by π, our system (2.32) will be autocatalytic (see [2, p. 116]) as long as $\rho_1 > \pi/(1 - \delta_1)$, where we now also assume that $0 < \delta_1 < 1$.

We next give an example of an f_1 and f_2 with corresponding h_1 and h_2 that meet (f-3) and the condition (1.20) of Theorem 1. For this example we use the notation of (2.6) and take $m = 1, 0 < \kappa < 1, \xi_1 = (\eta_0, \zeta_1, \ldots, \zeta_N)$, and

$$Qu = -\sum_{j=1}^{N} D_j A_j(x, \xi_1(u)) + a_0(x, \xi_1(u))u, \tag{2.34}$$

where $A_j(x, \xi_1)$ and $a_0(x, \xi_1)$ are defined in (2.6). Also, we take $Lu = -\Delta u$ and $\Omega = (0, \pi)^N, N \geq 1$. As we have shown, Q is $\kappa*$-related to L. In the system (1.1) with which we shall be dealing in this example $N_1 = 2, Q_1 = Q_2 = Q$, and $g_1(x,t) = g_2(x,t) = 0$. With $b_1(x) \in C(\bar{\Omega})$ and $b_2(t) \in C(T)$ we take

$$
\begin{aligned}
f_1(x,t,\xi_{11},\xi_{12}) &= -b_1(x)b_2(t)\left[c_{01}|\eta_{01}|^\kappa \arctan \eta_{01} + \sum_{j=1}^{N} c_{j1}|\zeta_{j1}|^\kappa \arctan \zeta_{j1}\right] \\
&\quad + \frac{\arctan(\eta_{02}^2 + \zeta_{12}^2 + \cdots + \zeta_{N2}^2)}{(1+\eta_{01}^2)^{\frac{1}{2}}} + d_1(x,t), \\
f_2(x,t,\xi_{11},\xi_{12}) &= -b_1(x)b_2(t)\left[c_{02}|\eta_{02}|^\kappa \tanh \eta_{02} + \sum_{j=1}^{N} c_{j2}|\zeta_{j2}| \tanh \zeta_{j2}\right] \\
&\quad + \frac{\tanh(\eta_{01}^2 + \zeta_{11}^2 + \cdots + \zeta_{N1}^2)}{(1+\eta_{02}^2)^{\frac{1}{2}}} + d_2(x,t),
\end{aligned} \tag{2.35}
$$

where $c_{0i}, c_{1i}, \ldots, c_{Ni}$ are constants with $c_{0i} > 0$ and $d_i(x,t) \in L^2(\tilde{\Omega})$ for $i = 1,2$. It is clear that f_i, for $i = 1,2$, satisfy (f-1) and (f-2). Also, since $0 < \kappa < 1$, an easy computation using (1.19) in (f-3) in conjunction with (2.35) shows that

$$
\begin{aligned}
h_1(x,t,\xi_1) &= -\frac{\pi}{2} b_1(x)b_2(t)\left[c_{01}|\eta_0|^\kappa \operatorname{sgn} \eta_0 + \sum_{j=1}^{N} c_{j1}|\zeta_j|^\kappa \operatorname{sgn} \zeta_j\right], \\
h_2(x,t,\xi_1) &= -b_1(x)b_2(t)\left[c_{02}|\eta_0|^\kappa \operatorname{sgn} \eta_0 + \sum_{j=1}^{N} c_{j2}|\zeta_j|^\kappa \operatorname{sgn} \zeta_j\right],
\end{aligned} \tag{2.36}
$$

where $|\xi_1| = 1$ and $\eta_0 \neq 0$. Hence f_i also meets (f-3) for $0 < \kappa < 1$ for $i = 1,2$.

We show that condition (1.20) holds for h_1. Similar reasoning shows that it holds for h_2. To show that condition (1.20) holds for h_1, we observe that if $v \in S^{\#}_{\lambda_1}$ is nontrivial, then v is a positive multiple of $\prod_{i=1}^{N} \sin x_i$ or $-\prod_{i=1}^{N} \sin x_i$. It therefore follows that

$$
\frac{\zeta_1(v)}{|\xi_1(v)|} = \pm \frac{\cos x_1 \prod_{i=2}^{N} \sin x_i}{\left\{\prod_{i=1}^{N} \sin^2 x_i + \cos^2 x_1 \prod_{i=2}^{N} \sin^2 x_i + \cdots + \cos^2 x_N \prod_{i=1}^{N-1} \sin^2 x_i\right\}^{\frac{1}{2}}}. \tag{2.37}
$$

Consequently from (2.36), we see that the integral involving $\zeta_1(v)/|\xi_1(v)|$ in (1.20) is a constant multiple of

$$
\int_0^\pi b_2(t)\,dt \underbrace{\int_0^\pi \cdots \int_0^\pi}_{N} b_1(x)|\zeta_1(v)|^\kappa v \operatorname{sgn}\left[\frac{\zeta_1(v)}{|\xi_1(v)|}\right] dx_1 \cdots dx_N. \tag{2.38}
$$

At this point we impose two further conditions on $b_1(x)$ and $b_2(t)$ that we know are in $C(\bar{\Omega})$ and $C(T)$, respectively. We shall also suppose

(2.39)
(i) $b_1(x) > 0 \ \forall x \in \bar{\Omega}$ and $b_2(t) > 0 \ \forall t \in T$;
(ii) in each variable $x_j, j = 1, \ldots, N, b_1(x)$ is even around $\frac{\pi}{2}$.

In particular, for the x_1-variable, (2.39) (ii) means $b_1(\frac{\pi}{2}+x_1,x_2,\dots,x_N)=b_1(\frac{\pi}{2}-x_1,x_2,\dots,x_N)$ for $0<x_1<\frac{\pi}{2}$ and $0<x_j<\pi, j=2,\dots,N$. For example, $b_1(x)=\prod_{j=1}^N \sin^4 x_j$.

It follows from (2.37) that $\zeta_1(v)/|\xi_1(v)|$ is odd in x_1 around $\frac{\pi}{2}$. Hence from (2.39) (ii), the integrand for

$$\underbrace{\int_0^\pi \cdots \int_0^\pi}_{N}$$

in (2.38) is odd in x_1 around $\frac{\pi}{2}$. Hence the corresponding integral is zero. A similar remark is valid for the integrals involving $\zeta_i(v)/|\xi_1(v)|, i=2,\dots,N$. Consequently, it follows that the integral in (1.20) is a positive multiple of

$$\text{(2.40)} \qquad -\int_0^\pi b_2(t)\,dt \underbrace{\int_0^\pi \cdots \int_0^\pi}_{N} b_1(x)|\eta_0(v)|^\kappa v \text{ sgn } \left[\frac{\eta_0(v)}{|\xi_1(v)|}\right] dx_1 \dots dx_N,$$

where $v=\pm\prod_{i=1}^N \sin x_i$ and

$$\frac{\eta_0(v)}{|\xi_1(v)|}=\frac{v}{\left\{\prod_{j=1}^N \sin^2 x_j+\cos^2 x_1 \prod_{j=2}^N \sin^2 x_j+\cdots+\cos^2 x_N \prod_{j=1}^{N-1}\sin^2 x_j\right\}^{\frac{1}{2}}}.$$

It is clear that regardless of the choice of $\pm, v$ sgn $[\eta_0(v)/|\xi_1(v)|]$ is strictly positive. Hence, from (2.39) (i), it follows that the integrand for

$$\underbrace{\int_0^\pi \cdots \int_0^\pi}_{N}$$

is strictly positive. Likewise $\int_0^{2\pi} f_2(t)\,dt>0$. We conclude that the full expression in (2.40) is strictly negative, and therefore the condition in (1.20) prevails for h_1 and likewise for h_2. Hence a weak solution to (1.1) exists where $N_1=2$ and $Q_1=Q_2=Q$ with Q given by (2.31), f_1 and f_2 by (2.35) and (2.39), and $g_1=g_2=0$.

In the previous example, $0<\kappa<1$ and $m=1$. We now allow κ to take on the value 0, so that $0\le\kappa<1$. Also, we take $m\ge 2, N_1$ an arbitrary positive integer, Qu to be given by (2.16), and Lu to be given by (2.21). Furthermore, we set $Q_1=Q_2=\cdots=Q_{N_1}=Q$. With $0=(0,\dots,0)$, we then define
(2.41)

$$f_i(x,t,\xi_{m1},\dots,\xi_{mN_1})=-\frac{b_i(x,t)\xi_{0i}|\xi_{0i}|^\kappa+\sum_{j=1}^{N_1}\sum_{1\le|\alpha|\le m}\frac{c^i_{\alpha j}\xi_{\alpha j}}{\sqrt{1+\xi^2_{\alpha j}}}}{\sqrt{1+\xi^2_{0i}}} \quad i=1,\dots,N_1,$$

where

$$\text{(2.42)} \qquad b_i(x,t)\in C(\tilde\Omega)\cap L^\infty(\tilde\Omega) \quad \text{and} \quad b_i(x,t)>0 \quad \forall (x,t)\in\tilde\Omega.$$

In the example under consideration $\Omega\subset\mathbb{R}^N$ is a general bounded open connected set and the $c^i_{\alpha j}$ in (2.41) are real constants. It is clear that f_i given by (2.41) satisfy (f-1) and (f-2) with $0\le\kappa<1$. Suppose $\xi^{(n)}_{mi}\to\xi_{mi}$, where $|\xi^{(n)}_{mi}|=1, |\xi_{mi}|=1$, and $\xi_{0i}\ne 0$.

Further suppose that $r_n \to \infty$. Then an easy computation using (2.41) shows that (1.19) holds with

$$h_i(x,t,\xi_m) = -b_i(x,t)\frac{\xi_0}{|\xi_0|^{1-\kappa}} \quad \text{for } |\xi_m| = 1 \text{ with } \xi_0 \neq 0. \tag{2.43}$$

Consequently, f_i also meets (f-3) for $i = 1,\dots,N_1$.

Next, we take

$$G_{1i}(u) = \left[\int_{\tilde\Omega} u(x,t)\phi_i(x)\sin it\right]^{\kappa} + 1 \quad \forall u \in L^2(\tilde\Omega) \text{ and } i = 1,\dots,N_1, \tag{2.44}$$

where $\phi_i(x)$ is the ith element in the sequence (1.10) corresponding to the L given by (2.21). It is clear that G_{1i} meets (G-1) for $i = 1,\dots,N_1$.

We define

$$G_{2i}(u) = \frac{1}{\|u\|_{L^2} + 1} \quad \forall u \in L^2(\tilde\Omega) \text{ and } i = 1,\dots,N_1. \tag{2.45}$$

It is clear that $G_{2i}(u)$ meets (G-2) for $i = 1,\dots,N_1$. Hence with Q_i defined as above; with f_i, G_{1i}, and G_{2i} defined, respectively, by (2.41), (2.44), and (2.45), and with g_i arbitrary functions in $L^2(\tilde\Omega)$ to show that $\exists(u_1,\dots,u_{N_1}) \in \tilde H^{N_1}$, which is a weak solution to (1.1), it remains to demonstrate that the h_i given by (2.43) each satisfies the condition in (1.20).

To show this, let $v \in S^{\#}_{\lambda_1}$ with v nontrivial, i.e.,

$$\int_{\tilde\Omega} v^2 \neq 0. \tag{2.46}$$

Now for $(x,t) \in \tilde\Omega(v), \xi_0(v(x,t)) = v(x,t)$. Hence it follows from (2.43) that

$$\int_{\tilde\Omega(v)} h_i\left(x,t,\frac{\xi_m(v)}{|\xi_m(v)|}\right)|\xi_m(v)|^{\kappa}v = -\int_{\tilde\Omega(v)} b_i(x,t)|v|^{1+\kappa} \tag{2.47}$$

for $i = 1,\dots,N_1$. From (2.42), (2.46), and the definition of $\tilde\Omega(v)$, it follows that the integral on the right-hand side of the equality in (2.47) is strictly positive. Hence it follows that the integral on the left-hand side is strictly negative and the condition in (1.20) is established for every i. We conclude that with the conditions just enumerated above, a weak solution to system (1.1) in $\tilde\Omega$ exists and our final example is complete.

3. Fundamental lemmas. To establish the theorems in this paper we use a Galerkin technique that depends upon a *CONS* for $\tilde\Omega$ that comes from our elliptic operator L introduced in (1.8) and (1.9). To do this, we introduce a *CONS* on $\tilde\Omega$ in the following manner.

Define

$$\begin{aligned}
&\text{(i)} \quad \tilde\phi^c_{nk}(x,t) = \frac{\phi_n(x)\cos kt}{\sqrt{\pi}}, \quad n,k = 1,2,\dots, \\
&\qquad\qquad\qquad\;\; = \frac{\phi_n(x)}{\sqrt{2\pi}}, \quad k = 0, n = 1,2,\dots; \\
&\text{(ii)} \quad \tilde\phi^s_{nk}(x,t) = \frac{\phi_n(x)\sin kt}{\sqrt{\pi}}, \quad n,k = 1,2,\dots.
\end{aligned} \tag{3.1}$$

Note that $\tilde{\phi}^c_{nk}, \tilde{\phi}^s_{nk} \in \tilde{H}$.

For $u \in L^2(\tilde{\Omega})$, set

$$\text{(3.2)} \qquad \begin{aligned} &\text{(i)} \quad \hat{u}^c(n,k) = \langle u, \tilde{\phi}^c_{nk} \rangle_{L^2} = \int_{\tilde{\Omega}} u(x,t)\tilde{\phi}^c_{nk}(x,t).,\\ &\text{(ii)} \quad \hat{u}^s(n,k) = \langle u, \tilde{\phi}^s_{nk} \rangle_{L^2}. \end{aligned}$$

From (1.9) we see that if $\hat{u}^c(n,0) = 0$ for $n = 1, 2, \ldots$ and $\hat{u}^c(n,k) = \hat{u}^s(n,k) = 0$ for $n, k = 1, 2, \ldots$, then $u = 0$ a.e. on $\tilde{\Omega}$. Therefore,

$$\text{(3.3)} \qquad \{\tilde{\phi}^c_{nk}\}^{\infty,\infty}_{n=1,k=0} \cup \{\tilde{\phi}^s_{nk}\}^{\infty,\infty}_{n=1,k=1} \quad \text{is a } CONS \text{ for } L^2(\tilde{\Omega}).$$

We first state the following lemma proved by Lefton and Shapiro in [12, pp. 160–162].

LEMMA 1. *Let Lu be the elliptic operator given by* (1.7) *satisfying* (L-1)–(L-3), *$\mathcal{L}(u,v)$ be the bilinear form given by* (1.12), *and $\{\tilde{\phi}^c_{nk}\}^{\infty,\infty}_{n=1,k=0} \cup \{\tilde{\phi}^s_{nk}\}^{\infty,\infty}_{n=1,k=1}$ be the $CONS$ for $L^2(\tilde{\Omega})$ given by* (3.1). *With λ_1 defined by* (1.11), *suppose furthermore that $\lambda_1 > 0$. For $v \in L^2(\tilde{\Omega})$, set*

$$\text{(3.4)} \qquad \sigma_n(v) = \sum_{j=1}^{n} \hat{v}(j,0)\tilde{\phi}^c_{j0} + \sum_{j=1}^{n}\sum_{k=1}^{n} [\hat{v}^c(j,k)\tilde{\phi}^c_{jk} + \hat{v}^s(j,k)\tilde{\phi}^s_{jk}],$$

where $\hat{v}^c(n,k)$ and $\hat{v}^s(n,k)$ are given by (3.2). *Then*

$$\text{(3.5)} \qquad \begin{aligned} &\text{(i)} \quad \lim_{n\to\infty} \int_{\tilde{\Omega}} [D_t\sigma_n(v) - D_t v]^2 = 0 \quad \forall v \in \tilde{H},\\ &\text{(ii)} \quad \lim_{n\to\infty} \int_{\tilde{\Omega}} \sum_{|\alpha|\le m} |D^\alpha[\sigma_n(v) - v]|^2 = 0 \quad \forall v \in \tilde{H}. \end{aligned}$$

Next, we observe that the following easily obtainable facts hold.

Fact 1. For $0 \le \kappa < 1$ and some positive constant $K_1(\kappa)$,

$$\text{(3.6)} \qquad \||u|^\kappa\|_{L^2} \le K_1\|u\|^\kappa_{L^2} \quad \forall u \in L^2.$$

Fact 2. Set $\|u\|^2_{W^{m,2,\sim}} = \int_{\tilde{\Omega}} \sum_{|\alpha|\le m} |D^\alpha u|^2 = \int_{\tilde{\Omega}} |\xi_m(u)|^2$ (see (1.14)) for $u \in \tilde{H}$. Then for $0 \le \kappa < 1$,

$$\text{(3.7)} \qquad \sum_{|\alpha|\le m} \|D^\alpha u\|^{2\kappa}_{L^2} \le s_m\|u\|^{2\kappa}_{W^{m,2,\sim}} \quad \forall u \in \tilde{H},$$

where s_m is the constant introduced in the fourth paragraph of §1.

Also, we record for future use the following fact, which is a consequence of the well-known Gårding's inequality on Ω [8, p. 34].

Fact 3. There are two positive constants c_0^* and c_1^* such that

$$\text{(3.8)} \qquad c_1^*\|u\|^2_{W^{m,2,\sim}} \le \mathcal{L}(u,u) + c_0^*\|u\|^2_{L^2} \quad \forall u \in \tilde{H}.$$

Next, we prove the following lemma.

LEMMA 2. *Suppose that $Q, L, f, g, G_{1i}, G_{2i}, h_i$, and ρ_i meet the conditions stated in the hypothesis of Theorem 2. (Hence by (1.6), $\lambda_1^* = \lambda_{1i} = \lambda_1 > 0$ for $i = 1, \ldots, N_1$.) Then for each positive integer n, $\exists(u_{1n}, \ldots, u_{N_{1}n}) \in S_n^{N_1}$ such that*

$$\begin{aligned} &D_t\langle u_{in}, v\rangle_{L^2} + \rho_i Q_i(u_{in}, v) \\ &\quad = \rho_i\left(\lambda_1^* - \frac{1}{n}\right)\langle u_{in}, v\rangle_{L^2} + \langle f_i(\cdot, \xi_m(u_{1n}), \ldots, \xi_m(u_{N_{1}n})), v\rangle_{L^2} \\ &\quad + G_{1i}(D_t u_{in})G_{2i}(u_{in})\langle g_i, v\rangle_{L^2} \quad \forall v \in S_n \end{aligned} \tag{3.9}$$

for $i = 1, \ldots, N_1$, where we define for $n \geq 1$

$$S_n = \left\{ v = \sum_{j=1}^{n} \tau_{j0}^c \tilde{\phi}_{j0}^c + \sum_{j=1}^{n}\sum_{k=1}^{n} (\tau_{jk}^c \tilde{\phi}_{jk}^c + \tau_{jk}^s \tilde{\phi}_{jk}^s), \right. \\ \left. \textit{where } \tau_{jk}^c, \tau_{jk}^s \in \mathbb{R}, j = 1, \ldots, n, k = 0, 1, \ldots n \right\}. \tag{3.10}$$

To prove the lemma we first note from (3.1) that

$$\begin{aligned} D_t\tilde{\phi}_{n0}^c &= 0, \\ D_t\tilde{\phi}_{nk}^c &= -k\tilde{\phi}_{nk}^s \quad \text{for } k \geq 1, \\ D_t\tilde{\phi}_{nk}^s &= k\tilde{\phi}_{nk}^c \quad \text{for } k \geq 1. \end{aligned} \tag{3.11}$$

Hence by (2.10)

$$v \in S_n \Rightarrow D_t v \in S_n. \tag{3.12}$$

For ease of notation, set $\{\psi_j\}_{j=1}^{2n^2+n} = \{\tilde{\phi}_{jk}^c\}_{j=1,k=0}^{n,n} \cup \{\tilde{\phi}_{jk}^s\}_{j=1,k=1}^{n,n}$, where $\{\psi_j\}_{j=1}^{n^2}$ corresponds to $\{\tilde{\phi}_{jk}^c\}_{j=1,k=1}^{n,n}$, $\{\psi_j\}_{j=n^2+1}^{2n^2}$ corresponds to $\{\tilde{\phi}_{jk}^s\}_{j=1,k=1}^{n,n}$, and $\{\psi_j\}_{j=2n^2+1}^{2n^2+n}$ corresponds to $\{\tilde{\phi}_{j0}^c\}_{j=1}^{n}$. Hence we can rewrite S_n as

$$S_n = \left\{ v \in \tilde{H} : v = \sum_{j=1}^{2n^2+n} \gamma_j\psi_j \text{ where } \gamma_j \in \mathbb{R}, j = 1, \ldots, 2n^2 + n \right\}. \tag{3.13}$$

Let $\gamma = (\gamma_k^i)_{k=1,\ldots,2n^2+n, i=1,\ldots,N_1} \in \mathbb{R}^{2N_1n^2+N_1n}$. Set

$$\begin{aligned} B_k^i(\gamma) &= \langle D_t(\gamma_j^i\psi_j), \psi_k\rangle_{L^2} + \rho_i Q(\gamma_j^i\psi_j, \psi_k) + \rho_i\left(\frac{1}{n} - \lambda_1^*\right)\langle\gamma_j^i\psi_j, \psi_k\rangle_{L^2} \\ &\quad - \langle f_i(\cdot, \xi_m(\gamma_j^1\psi_j), \ldots, \xi_m(\gamma_j^{N_1}\psi_j)), \psi_k\rangle_{L^2} \\ &\quad - G_{1i}(D_t\gamma_j^i\psi_j)G_{2i}(\gamma_j^i\psi_j)\langle g_i, \psi_k\rangle_{L^2} \end{aligned} \tag{3.14}$$

for $k = 1, \ldots, 2n^2 + n, i = 1, \ldots, N_1$, where we use the summation convention in (3.14) for $j = 1, \ldots, 2n^2 + n$. We shall use the summation convention also on $k = 1, \ldots, 2n^2 + n$.

Set $B(\gamma) = (B_k^i(\gamma))_{k=1,\dots,2n^2+n,i=1,\dots,N_1}$. Observing that $\mathcal{Q}_i(\cdot,\cdot)$ is linear in the second variable and using the summation convention on j and k, we have

$$\begin{aligned} B(\gamma)\cdot\gamma &= \sum_{i=1}^{N_1} B_k^i(\gamma)\gamma_k^i \\ &= \sum_{i=1}^{N_1}\Big[\langle D_t(\gamma_j^i\psi_j),\gamma_k^i\psi_k\rangle_{L^2} + \rho_i\mathcal{Q}_i(\gamma_j^i\psi_j,\gamma_k^i\psi_k) \\ &\qquad + \rho_i\left(\frac{1}{n}-\lambda_1^*\right)\langle\gamma_j^i\psi_j,\gamma_k^i\psi_k\rangle_{L^2} \\ &\qquad - \langle f_i(\cdot,\xi_m(\gamma_j^1\psi_j),\dots,\xi_m(\gamma_j^{N_1}\psi_j)),\gamma_k^i\psi_k\rangle_{L^2} \\ &\qquad - G_{1i}(D_t\gamma_j^i\psi_j)G_{2i}(\gamma_j^i\psi_j)\langle g_i,\gamma_k^i\psi_k\rangle_{L^2}\Big]. \end{aligned} \tag{3.15}$$

From (Q-1), (Q-2), (f-1), (f-2), (G-1), and (G-2) we see that

$$B = (B_k^i)_{k=1,\dots,2n^2+n,i=1,\dots,N_1} : \mathbb{R}^{2N_1n^2+N_1n} \to \mathbb{R}^{2N_1n^2+N_1n}$$

is continuous in γ. Also, it follows from the orthogonality of the $\{\psi_j\}_{j=1}^{2n^2+n}$ that there is a positive constant c_0' such that

$$\|\gamma_j^i\psi_j\|_{L^2} \geq c_0'[\gamma_1^{i2}+\cdots+\gamma_{2n^2+n}^{i2}]^{\frac{1}{2}}. \tag{3.16}$$

We propose to show that for γ sufficiently large

$$B(\gamma)\cdot\gamma > 0. \tag{3.17}$$

To show that (3.17) is true, we first observe that $0 = \int_{\tilde{\Omega}} D_tv^2 = 2\langle D_tv, v\rangle_{L^2}$ $\forall v \in \mathcal{A}$. Hence it follows that

$$\sum_{i=1}^{N_1}\langle D_t\gamma_j^i\psi_j,\gamma_k^i\psi_k\rangle_{L^2} = 0 \quad \forall\gamma\in\mathbb{R}^{2N_1n^2+N_1n}. \tag{3.18}$$

Also, it is easy to see from (f-1), (f-2), Fact 1, and Fact 2 that

$$\begin{aligned} &\sum_{i=1}^{N_1}|\langle f_i(\cdot,\xi_m(\gamma_j^1\psi_j),\dots,\xi_m(\gamma_j^{N_1}\psi_i)),\gamma_k^i\psi_k\rangle_{L^2}| \\ &\leq\left(C_1\sum_{i=1}^{N_1}\|\gamma_k^i\psi_k\|_{W^{m,2,\sim}}^{\kappa}+C_2\right)\sum_{i=1}^{N_1}\|\gamma_k^i\psi_k\|_{L^2} \quad \forall\gamma\in\mathbb{R}^{2N_1n^2+N_1n}, \end{aligned} \tag{3.19}$$

where C_1 and C_2 are positive constants.

Next, we note from (1.9), (1.10), (3.1), and (3.3) that

$$\begin{aligned} \mathcal{L}(\tilde{\phi}_{j_1k_1}^c,\tilde{\phi}_{j_2k_2}^s) &= 0 \quad \forall(j_1,k_1),(j_2,k_2), \\ \mathcal{L}(\tilde{\phi}_{j_1k_1}^c,\tilde{\phi}_{j_2k_2}^c) &= 0 \quad \forall(j_1,k_1)\neq(j_2,k_2), \\ \mathcal{L}(\tilde{\phi}_{j_1k_1}^s,\tilde{\phi}_{j_2k_2}^s) &= 0 \quad \forall(j_1,k_1)\neq(j_2,k_2), \end{aligned} \tag{3.20}$$

$$\mathcal{L}(\tilde{\phi}^c_{j_1k_1},\tilde{\phi}^c_{j_1k_1}) = \lambda_{j1}, \qquad \mathcal{L}(\tilde{\phi}^s_{j_1k_1},\tilde{\phi}^s_{j_1k_1}) = \lambda_{j1}.$$

Since n is fixed and we are assuming that $\lambda_1 > 0$, it follows from (1.13) and (3.20) that

$$\begin{aligned}&\exists C_3 > 0 \text{ such that } \lambda_1|\nu|^2 \le \mathcal{L}(\nu_j\psi_j,\nu_k\psi_k) \le C_3|\nu|^2\\ &\forall \nu = (\nu_1,\ldots,\nu_{2n^2+n}) \in \mathbb{R}^{2n^2+n}.\end{aligned}\tag{3.21}$$

Consequently, we see from Fact 3 and (3.21) coupled with (3.19) that there are positive constants C_4 and C_5 such that

$$\begin{aligned}&\sum_{i=1}^{N_1} |\langle f_i(\cdot,\xi_m(\gamma^1_j\psi_j),\ldots,\xi_m(\gamma^{N_1}_j\psi_j)),\gamma^i_k\psi_k\rangle_{L^2}| \le C_4|\gamma|^{\kappa+1} + C_5|\gamma|\\ &\forall\gamma \in \mathbb{R}^{2N_1n^2+N_1n}.\end{aligned}\tag{3.22}$$

Next, we note from (3.1), (3.3), and (3.11) that there exists a positive constant C_6 such that

$$\|D_t(\nu_j\psi_j)\|_{L^2} \le C_6|\nu| \quad \forall\nu \in \mathbb{R}^{2n^2+n}.$$

Hence it follows from (G-1) and this last inequality that there is a positive constant C_7 such that

$$\sum_{i=1}^{N_1} |G_{1i}(D_t\gamma^i_j\psi_j)| \le C_7|\gamma|^{\kappa} + C_7 \quad \forall\gamma \in \mathbb{R}^{2N_1n^2+N_1n}.$$

Using this last inequality in conjunction with (G-2) and the fact that $g_i \in L^2(\tilde{\Omega})$, we finally obtain that

(3.23)

$$\sum_{i=1}^{N_1} |G_{1i}(D_t\gamma^i_j\psi_j)G_{2i}(\gamma^i_j\psi_j)\langle g_i,\gamma^i_k\psi_k\rangle_{L^2}| \le C_8|\gamma|^{\kappa+1} + C_8|\gamma| \quad \forall\gamma \in \mathbb{R}^{2N_1n^2+N_1n},$$

where C_8 is a positive constant.

Next we see from (3.21), Fact 3, and the fact that $\lambda_1 > 0$ that $\|\nu_j\psi_j\|^2_{W^{m,2,\sim}} \le c^*_2\mathcal{L}(\nu_j\psi_j,\nu_k\psi_k)$ $\forall\nu \in \mathbb{R}^{2n^2+n}$, where c^*_2 is a positive constant. Therefore, it follows from (1.16) (ii) that there are positive constants C'_3 and C'_4 such that

$$\mathcal{Q}_i(\gamma^i_j\psi_j,\gamma^i_k\psi_k) - \mathcal{L}(\gamma^i_j\psi_j,\gamma^i_k\psi_k) \ge -C'_3|\mathcal{L}(\gamma^i_j\psi_j,\gamma^i_k\psi_k)|^{\frac{(\kappa+1)}{2}} - C'_4\tag{3.24}$$

for $i = 1,\ldots,N_1$. Also from the fact that $\lambda^*_1 = \lambda_1$ we obtain from (1.13) that

$$\mathcal{L}(\gamma^i_j\psi_j,\gamma^i_k\psi_k) - \lambda^*_1\langle\gamma^i_j\psi_j,\gamma^i_k\psi_k\rangle_{L^2} \ge 0.$$

We conclude from the above inequality and (3.24) that

$$\begin{aligned}&\sum_{i=1}^{N_1}\left[\rho_i\mathcal{Q}_i(\gamma^i_j\psi_j,\gamma^i_k\psi_k) - \rho_i\left(\lambda^*_1 - \frac{1}{n}\right)\langle\gamma^i_j\psi_j,\gamma^i_k\psi_k\rangle_{L^2}\right]\\ &\quad\ge \frac{\rho_0}{n}|\gamma|^2 - C'_3\sum_{i=1}^{N_1}\rho_i|\mathcal{L}(\gamma^i_j\psi_j,\gamma^i_k\psi_k)|^{\frac{(\kappa+1)}{2}} - C'_4\sum_{i=1}^{N_1}\rho_i,\end{aligned}\tag{3.25}$$

where $\rho_0 = \min(\rho_1, \dots, \rho_{N_1}) > 0$. Now it follows from (3.21) that there is a constant $C_5' > 0$ such that

$$C_3' \sum_{i=1}^{N_1} \rho_i |\mathcal{L}(\gamma_j^i \psi_j, \gamma_k^i \psi_k)|^{\frac{(\kappa+1)}{2}} \leq C_5' |\gamma|^{\kappa+1}.$$

Using this last inequality in conjunction with (3.15), (3.18), (3.22), (3.23), and (3.25), we obtain that

$$B(\gamma) \cdot \gamma \geq \frac{\rho_0}{n} |\gamma|^2 - (C_4 + C_8 + C_5') |\gamma|^{\kappa+1} - (C_5 + C_8)|\gamma| - C_4' \sum_{i=1}^{N_1} \rho_i$$
$$\forall \gamma \in \mathbb{R}^{2N_1 n^2 + N_1 n}.$$

Since $0 \leq \kappa < 1$, we conclude that there exists $\Gamma_0 > 0$ such that

$$B(\gamma) \cdot \gamma \geq \frac{\rho_0}{2n} |\gamma|^2 \quad \text{for } |\gamma| \geq \Gamma_0. \tag{3.26}$$

This establishes the inequality in (3.17).

Since $B(\gamma)$ is a continuous map from $\mathbb{R}^{2N_1 n^2 + N_1 n}$ into $\mathbb{R}^{2N_1 n^2 + N_1 n}$, it follows from (3.26) and a well-known theorem in nonlinear analysis (see [10, p. 219] or [16, p. 18]) that $\exists \gamma^\# = (\gamma_j^{\# i})_{j=1,\dots,2n^2+n, i=1,\dots,N_1}$ with $|\gamma^\#| < \Gamma_0$ such that $B_k^i(\gamma^\#) = 0$ for $k = 1, \dots, 2n^2 + n$ and $i = 1, \dots, N_1$. Hence

$$\begin{aligned} D_t(\gamma_j^{\# i} \psi_j, \psi_k) &+ \rho_i Q_i(\gamma_j^{\# i} \psi_j, \psi_k) \\ &= \rho_i \left(\lambda_1^* - \frac{1}{n} \right) \langle \gamma_j^{\# i} \psi_j, \psi_k \rangle_{L^2} \\ &+ \langle f_i(\cdot, \xi_m(\gamma_j^{\# 1} \psi_j), \dots, \xi_m(\gamma_j^{\# N_1} \psi_j)), \psi_k \rangle_{L^2} \\ &+ G_{1i}(D_t \gamma_j^{\# i} \psi_j) G_{2i}(\gamma_j^{\# i} \psi_j) \langle g_i, \psi_k \rangle_{L^2} \end{aligned} \tag{3.27}$$

for $k = 1, \dots, 2n^2 + n$. Since every element in S_n is a finite linear combination of the ψ_k's, the conclusion to the lemma follows immediately from (3.27).

Remark 1. It is clear from the proof of Lemma 2 that $\frac{1}{n}$ in (3.9) can be replaced by $\delta_i > 0$ for $i = 1, \dots, N_1$ and the conclusion of Lemma 2 will still prevail. So, in particular, we have that under the conditions in the hypothesis of Lemma 2 with $\delta_i > 0$ for $i = 1, \dots, N_1, \exists (u_{1n}, \dots, u_{N_1 n}) \in S_n^{N_1}$ such that (3.9) holds with $\left(\lambda_1^* - \frac{1}{n}\right)$ replaced by $(\lambda_1^* - \delta_i)$.

4. Proof of Theorem 1. We notice that if we add a large positive constant $c*$ to $a_0^i(x, \xi_m(u))$ in (1.2) and to $b_{0,0}(x)$ in (1.7) where $0 = (0, \dots, 0)$ and $\rho_i c * u_i$ to the right-hand side of (1.1), then with no loss in generality we can assume that

$$\begin{aligned} a_0^i(x, \xi_m) &\geq \varepsilon_0 > 0 \quad \forall \xi_m \in \mathbb{R}^{s_m} \text{ and a.e. } x \in \Omega, \\ b_{00}(x) &\geq \varepsilon_0 > 0 \quad \text{a.e. } x \in \Omega \end{aligned} \tag{4.1}$$

and that

$$\lambda_{1i}^* = \lambda_1 > 0 \quad \text{for } i = 1, \dots, N_1. \tag{4.2}$$

Since $\lambda_1 > 0$, we see from (L-1), (1.13), and (3.8) in Fact 3 that there is a positive constant c_3^* such that

$$\frac{\|u\|^2_{W^{m,2,\sim}}}{c_3^*} \leq \|u\|^2_{\mathcal{L}} \leq c_3^* \|u\|^2_{W^{m,2,\sim}} \quad \forall u \in \tilde{H}, \tag{4.3}$$

where $\|u\|^2_{\mathcal{L}} = \mathcal{L}(u,u) > 0$ for $u \neq 0$. As a consequence of (4.3), we see that (1.16) (i) is equivalent to

$$u \in \tilde{H}, \qquad \liminf_{\|u\|_{\mathcal{L}} \to \infty} \frac{Q_i(u,u) - \mathcal{L}(u,u)}{\|u\|^{1+\kappa}_{\mathcal{L}}} \geq 0 \quad \text{for } i = 1, \dots, N_1. \tag{4.4}$$

To prove Theorem 1, we invoke Lemma 2 and have a sequence $\{(u_{1n}, \dots, u_{N_1 n})\}_{n=1}^{\infty}$ with each $u_{in} \in S_n$ (defined in (3.10)) such that (3.9) holds. We claim there is a constant K_1 such that

$$\|u_{in}\|_{\mathcal{L}} \leq K_1 \quad \forall n \text{ and for } i = 1, \dots, N_1. \tag{4.5}$$

For ease of notation, we shall establish (4.5) for the case $i = 1$. A similar proof will prevail for the other values of i. Suppose then (4.5) is false when $i = 1$. Then without loss of generality, we can suppose

$$\lim_{n \to \infty} \|u_{1n}\|_{\mathcal{L}} = \infty. \tag{4.6}$$

We propose to show that (4.6) leads to a contradiction of the inequality in (1.20) when $i = 1$. To accomplish this we put u_{1n} in place of v in (3.9) to obtain

$$\begin{aligned} \rho_1 Q_1(u_{1n}, u_{1n}) = \rho_1 \left(\lambda_1^* - \frac{1}{n} \right) \|u_{1n}\|^2_{L^2} + \langle f_1(\cdot, \xi_m(u_{1n}), \dots, \xi_m(u_{N_1 n})), u_{1n} \rangle_{L^2} \\ + G_{11}(D_t u_{1n}) G_{21}(u_{1n}) \langle g_1, u_{1n} \rangle_{L^2}, \end{aligned} \tag{4.7}$$

where we have used the fact established in (3.18) that

$$\langle D_t v, v \rangle_{L^2} = 0 \quad \forall v \in \tilde{H}. \tag{4.8}$$

With $\varepsilon > 0$, we have from (4.4) and (4.6) that

$$Q_1(u_{1n}, u_{1n}) \geq \mathcal{L}(u_{1n}, u_{1n}) - \varepsilon \|u_{1n}\|^{1+\kappa}_{\mathcal{L}} \quad \text{for } n \geq n_2. \tag{4.9}$$

Also, we note from (3.2) that

$$\mathcal{L}(u_{1n}, \tilde{\phi}^c_{jk}) = \lambda_j \langle u_{1n}, \tilde{\phi}^c_{jk} \rangle_{L^2} = \lambda_j \hat{u}^c_{1n}(j,k) \tag{4.10}$$

with a similar situation for $\tilde{\phi}^s_{jk}$. Thus

$$\mathcal{L}(u_{1n}, u_{1n}) = \sum_{j=1}^{n} \lambda_j \left\{ \sum_{k=0}^{n} |\hat{u}^c_{1n}(j,k)|^2 + \sum_{k=1}^{n} |\hat{u}^s_{1n}(j,k)|^2 \right\}. \tag{4.11}$$

Now $\lambda_1 = \cdots = \lambda_{n_1}$ and $\exists \gamma > 0$ such that

$$\gamma \lambda_j < (\lambda_j - \lambda_1) \quad \text{for } j \geq n_1 + 1. \tag{4.12}$$

Thus, we see from (4.2), (4.7), (4.11), and (4.12) that

$$\begin{aligned}\rho_1\gamma \sum_{j=n_1+1}^{n} \lambda_j &\left\{ \sum_{k=0}^{n} |\hat{u}^c_{1n}(j,k)|^2 + \sum_{k=1}^{n} |\hat{u}^s_{1n}(j,k)|^2 \right\} \\ &\le \langle f_1(\cdot,\xi_m(u_{1n}),\dots,\xi_m(u_{N_1n})),u_{1n}\rangle_{L^2} \\ &+ G_{11}(D_tu_{1n})G_{21}(u_{1n})\langle g_1,u_{1n}\rangle_{L^2} + \varepsilon\rho_1\|u_{1n}\|_{\mathcal{L}}^{1+\kappa} \\ &\forall n \ge \max(n_1,n_2),\end{aligned} \tag{4.13}$$

where $\gamma > 0$.

Next, we note from (f-2), (4.3), and (4.6) that there is a constant $K_2 > 0$ such that

$$|\langle f_1(\cdot,\xi_m(u_{1n}),\dots,\xi_m(u_{N_1n})),u_{1n}\rangle_{L^2}| \le K_2[\|u_{1n}\|^\kappa_{\mathcal{L}} + 1]\|u_{1n}\|_{L^2} \quad \forall n. \tag{4.14}$$

Likewise from (G-1) and (G-2), we see there is a constant K_3 such that

$$|G_{11}(D_tu_{1n})G_{21}(u_{1n})\langle g_1,u_{1n}\rangle_{L^2}| \le K_3[\|D_tu_{1n}\|^\kappa_{L^2} + 1]\|u_{1n}\|_{L^2} \quad \forall n. \tag{4.15}$$

We define

$$y_{1n} = \sum_{j=1}^{n_1} \left\{ \sum_{k=0}^{n} \hat{u}^c_{1n}(j,k)\tilde{\phi}^c_{jk} + \sum_{k=1}^{n} \hat{u}^s_{1n}(j,k)\tilde{\phi}^s_{jk} \right\} \tag{4.16}$$

and

$$z_{1n} = u_{1n} - y_{1n}. \tag{4.17}$$

Since u_{1n} has the same form as y_{1n} in (4.16) with n_1 replaced by n we see from (3.20) that

$$\|u_{1n}\|^2_{\mathcal{L}} = \|y_{1n}\|^2_{\mathcal{L}} + \|z_{1n}\|^2_{\mathcal{L}}. \tag{4.18}$$

Next, we observe from (3.11) that $D_tu_{1n} \in S_n \ \forall n$. Hence taking $v = D_tu_{1n}$ in (3.9) with $i = 1$, we see from (Q-6) and (4.8) that

$$\begin{aligned}\|D_tu_n\|^2_{L^2} &= \langle f_1(\cdot,\xi_m(u_{1n}),\dots,\xi_m(u_{N_1n})),D_tu_{1n}\rangle_{L^2} \\ &+ G_{11}(D_tu_{1n})G_{21}(u_{1n})\langle g_1,D_tu_{1n}\rangle_{L^2}.\end{aligned} \tag{4.19}$$

It therefore follows from (f-2), (4.6), (G-1), and (G-2) that there is a positive constant K_4 such that

$$\|D_tu_{1n}\|_{L^2} \le K_4\|u_{1n}\|^\kappa_{\mathcal{L}} + K_4\|D_tu_{1n}\|^\kappa_{L^2} + K_4 \quad \forall n.$$

It follows from this last inequality, (4.6), and the fact that $0 \le \kappa < 1$ that

$$\lim_{n\to\infty} \frac{\|D_tu_{1n}\|_{L^2}}{\|u_{1n}\|_{\mathcal{L}}} = 0. \tag{4.20}$$

From (4.13) and (4.16)–(4.18), we see that

$$\rho_1\gamma\|z_{1n}\|^2_{\mathcal{L}} \le \text{the right-hand side of (4.13)},$$

where $\gamma > 0$. Consequently, it follows from this last inequality, (4.14), (4.15), and (4.20) that

$$\rho_1 \gamma \lim_{n\to\infty} \|Z_{1n}\|_{\mathcal{L}}^2 \leq 0, \tag{4.21}$$

where

$$U_{1n} = \frac{u_{1n}}{\|u_{1n}\|_{\mathcal{L}}}, \quad Y_{1n} = \frac{y_{1n}}{\|u_{1n}\|_{\mathcal{L}}}, \quad Z_{1n} = \frac{z_{1n}}{\|u_{1n}\|_{\mathcal{L}}}. \tag{4.22}$$

Since $\gamma > 0$ and $\rho_1 > 0$, we conclude from (4.21) that

$$\lim_{n\to\infty} \|Z_{1n}\|_{\mathcal{L}} = 0. \tag{4.23}$$

Next, we observe from (1.3) and (1.14) that

$$\|U_{1n}\|_{\tilde{H}}^2 = \|D_t U_{1n}\|_{L^2}^2 + \|U_{1n}\|_{W^{m,2,\sim}}^2. \tag{4.24}$$

From (4.20) and (4.22) we have that

$$\lim_{n\to\infty} \|D_t U_{1n}\|_{L^2} = 0. \tag{4.25}$$

We conclude from this last fact in conjunction with (4.3), (4.22), and (4.24) that

$$\{\|U_{1n}\|_{\tilde{H}}\}_{n=1}^{\infty} \text{ is a bounded sequence.} \tag{4.26}$$

It follows from (4.26), standard Hilbert space theory, and the compact imbedding theorem for Sobolev spaces that there exists a subsequence (which for ease of notation we take to be the full sequence) and

$$U \in \tilde{H} \tag{4.27}$$

such that

$$\lim_{n\to\infty} \|U_{1n} - U\|_{L^2} = 0; \tag{4.28}$$

$$\lim_{n\to\infty} U_{1n}(x,t) = U(x,t) \quad \text{a.e. in } \tilde{\Omega}; \tag{4.29}$$

$$\lim_{n\to\infty} \int_{\tilde{\Omega}} D^\alpha U_{1n}(x,t) w = \int_{\tilde{\Omega}} D^\alpha U(x,t) w, \qquad 1 \leq |\alpha| \leq m, \forall w \in L^2(\tilde{\Omega}). \tag{4.30}$$

Since $\|Z_{1n}\|_{L^2}^2 \leq (\lambda_1)^{-1} \|Z_{1n}\|_{\mathcal{L}}^2$ and $Y_{1n} = U_{1n} - Z_{1n}$, it follows from (4.23) and (4.28) that

$$\lim_{n\to\infty} \|Y_{1n} - U\|_{L^2} = 0. \tag{4.31}$$

Next, we observe from (4.16) that $\hat{Y}_{1n}^c(j,k) = \hat{Y}_{1n}^s(j,k) = 0$ for $j \geq n_1 + 1$. Consequently, it follows from (4.31) that

$$\hat{U}^c(j,k) = \hat{U}^s(j,k) = 0 \quad \text{for } j \geq n_1 + 1 \text{ and } k = 0, 1, 2, \ldots. \tag{4.32}$$

Next, we observe from (3.1) that for $j \geq 1$ and $k \geq 1$

$$\begin{aligned} k\hat{U}^s_{1n}(j,k) &= -\int_{\tilde{\Omega}} U_{1n}(x,t) D_t \tilde{\phi}^c_{jk}(x,t) \\ &= \int_{\tilde{\Omega}} \tilde{\phi}^c_{jk}(x,t) D_t U_{1n}(x,t). \end{aligned}$$

Consequently, it follows from (4.25) that for $j \geq 1$ and $k \geq 1, \lim_{n\to\infty} \hat{U}^s_{1n}(j,k) = 0$. We conclude from (4.28) that

$$\hat{U}^c(j,k) = \hat{U}^s(j,k) = 0 \quad \text{for } j,k \geq 1. \tag{4.33}$$

From (4.32), (4.33), and (3.3), it follows that

$$U(x,t) = \sum_{j=1}^{n_1} \hat{U}^c(j,0)\frac{\phi_j(x)}{\sqrt{2\pi}} \quad \text{a.e. in } \tilde{\Omega}. \tag{4.34}$$

Next, we note from (4.16) that

$$\mathcal{L}(Y_{1n}, Y_{1n}) = \lambda_1 \langle Y_{1n}, Y_{1n}\rangle_{L^2}$$

and from (4.22) that

$$1 = \|U_{1n}\|^2_{\mathcal{L}} = \|Y_{1n}\|^2_{\mathcal{L}} + \|Z_{1n}\|^2_{\mathcal{L}}.$$

From these last two equalities and (4.23), we see that

$$\lim_{n\to\infty} \|Y_{1n}\|^2_{L^2} = \lambda_1^{-1}.$$

Hence it follows from (4.31) that

$$\|U\|^2_{L^2} = \lambda_1^{-1} > 0. \tag{4.35}$$

From (1.18), (4.34), and (4.35), we see that U is a nontrivial function in $S^{\#}_{\lambda_1}$. With $\tilde{\Omega}(U)$ defined in the statement of Theorem 1, we set

$$\tilde{\Omega}_1(U) = \tilde{\Omega}(U) \cap \tilde{\Omega}_0, \tag{4.36}$$

where $\tilde{\Omega}_0$ is defined in (f-3).

Next, we return to (4.16) and write

$$y_{1,1n} = y_{1n} - y_{2,1n}, \tag{4.37}$$

where

$$y_{2,1n} = \sum_{j=1}^{n_1} \left\{ \sum_{k=1}^{n} \hat{u}^c_{1n}(j,k)\tilde{\phi}^c_{jk} + \hat{u}^s_{1n}(j,k)\tilde{\phi}^s_{jk} \right\}.$$

Then with $Y_{i,1n} = \frac{y_{i,1n}}{\|u_{1n}\|_{\mathcal{L}}}$ for $i = 1,2$, we see from (4.37) that

$$Y_{1,1n}(x,t) = \sum_{j=1}^{n_1} \hat{U}^c_{1n}(j,0)\frac{\phi_j(x)}{\sqrt{2\pi}}. \tag{4.38}$$

From (4.28), $\hat{U}^c_{1n}(j,0) \to \hat{U}^c(j,0)$ for $j = 1, \ldots, n_1$ as $n \to \infty$. Consequently, we obtain from (4.34) that

$$\text{(4.39)} \qquad \begin{aligned} &\text{(i)} \quad \lim_{n\to\infty} D^\alpha Y_{1,1n}(x,t) = D^\alpha U(x,t) \text{ a.e. in } \tilde{\Omega} \text{ for } |\alpha| \le m, \\ &\text{(ii)} \quad \lim_{n\to\infty} \|D^\alpha Y_{1,1n} - D^\alpha U\|_{L^2} = 0 \text{ for } |\alpha| \le m. \end{aligned}$$

From (4.31) and (4.37), we also have that

$$\|Y_{2,1n}\|_{L^2} \le \|Y_{1n} - U\|_{L^2} + \|U - Y_{1,1n}\|_{L^2}.$$

But then it follows from (4.31) and (4.39) (ii) that $\lim_{n\to\infty} \|Y_{2,1n}\|_{L^2} = 0$. On the other hand, it follows from (4.37) that $\mathcal{L}(Y_{2,1n}, Y_{2,1n}) = \lambda_1 \|Y_{2,1n}\|^2_{L^2}$. We conclude that

$$\text{(4.40)} \qquad \lim_{n\to\infty} \|Y_{2,1n}\|_{\mathcal{L}} = 0.$$

From (4.23) and (4.40) joined with (4.3), we in turn obtain that

$$\text{(4.41)} \qquad \begin{aligned} &\text{(i)} \quad \lim_{n\to\infty} [\|D^\alpha Y_{2,1n}\|_{L^2} + \|D^\alpha Z_{1n}\|_{L^2}] = 0 \text{ for } |\alpha| \le m, \\ &\text{(ii)} \quad \lim_{n\to\infty} D^\alpha Y_{2,1n}(x,t) = 0 \text{ a.e. in } \Omega \text{ for } |\alpha| \le m, \\ &\text{(iii)} \quad \lim_{n\to\infty} D^\alpha Z_{1n}(x,t) = 0 \text{ a.e. in } \Omega \text{ for } |\alpha| \le m \end{aligned}$$

where we have used full sequences in (4.41) (ii) and (iii) rather than subsequences for ease of notation.

Since $U_{1n} = Y_{1,1n} + Y_{2,1n} + Z_{1n}$, we see from (4.39) and (4.41) that (4.28), (4.29), and (4.30) can be replaced by the stronger assertions

$$\text{(4.42)} \qquad \lim_{n\to\infty} \|D^\alpha U_{1n} - D^\alpha U\|_{L^2} = 0 \quad \text{for } |\alpha| \le m,$$

$$\text{(4.43)} \qquad \lim_{n\to\infty} D^\alpha U_{1n}(x,t) = D^\alpha U(x,t) \quad \text{a.e. in } \tilde{\Omega} \text{ for } |\alpha| \le m.$$

Next, we return to (4.13) and observe that since ρ_1, γ, and λ_1 are all positive, the left-hand side of the inequality is nonnegative. Consequently, upon dividing both sides by $\|u_{1n}\|^{1+\kappa}_{\mathcal{L}}$, we obtain that

(4.44)

$$-\varepsilon\rho_1 \le \frac{\langle f_1(\cdot, \xi_m(u_{1n}), \ldots, \xi_m(u_{N_1 n})), U_{1n}\rangle_{L^2} + G_{11}(D_t u_{1n}) G_{21}(u_{1n}) \langle g_1, U_{1n}\rangle_{L^2}}{\|u_{1n}\|^\kappa_{\mathcal{L}}}.$$

Since

$$\frac{\|u_{1n}\|_{L^2}}{\|u_{1n}\|_{\mathcal{L}}} \xrightarrow[n\to\infty]{} \lambda_1^{-\frac{1}{2}}$$

by (4.22), (4.35), and (4.42), we have that $\lim_{n\to\infty} \|u_{1n}\|_{L^2} = \infty$. Consequently it follows from (G-1), (G-2), (4.20), and (4.42) that

$$\lim_{n\to\infty} \frac{G_{11}(D_t u_{1n}) G_{21}(u_{1n}) \langle g_1, U_{1n}\rangle_{L^2}}{\|u_{1n}\|^\kappa_{\mathcal{L}}} = 0.$$

We therefore conclude from (4.44) that

$$0 \leq \limsup_{n\to\infty} \frac{\langle f_1(\cdot, \xi_m(u_{1n}), \dots, \xi_m(u_{N_1 n})), U_{1n}\rangle_{L^2}}{\|u_{1n}\|_{\mathcal{L}}^{\kappa}} \tag{4.45}$$

because ε is an arbitrary positive constant.

Next, we observe that

$$\left\{ \frac{f_1(\cdot, \xi_m(u_{1n}), \dots, \xi_m(u_{N_1 n}))U_{1n}}{\|u_{1n}\|_{\mathcal{L}}^{\kappa}} \right\}_{n=1}^{\infty} \tag{4.46}$$

is an absolutely equi-integrable sequence;

that is, given $\varepsilon > 0, \exists\delta$ such that if $E \subset \tilde{\Omega}$ and meas $(E) < \delta$, then

$$\int_E \frac{|f_1(\cdot, \xi_m(u_{1n}), \dots, \xi_m(u_{N_1 n}))U_{1n}|}{\|u_{1n}\|_{\mathcal{L}}^{\kappa}} < \varepsilon \quad \forall n.$$

To establish (4.46), we see from (f-2) that it is sufficient to show

(4.47) $\{|D^\alpha(U_{1n})|^\kappa |U_{1n}|\}_{n=1}^\infty$ is an absolutely equi-integrable sequence for $0 \leq |\alpha| \leq m$.

From (4.42) we see that $\{|U_{1n}|^2\}_{n=1}^\infty$ is an absolutely equi-integrable sequence and that $\{|D^\alpha U_{1n}|^{2\kappa}\}_{n=1}^\infty$ is a bounded sequence in $L^1(\tilde{\Omega})$. Equation (4.47) follows immediately from Schwarz's inequality and these last two facts. Hence, (4.46) is established.

Next, with $\tilde{\Omega}(U)$ defined in the statement of Theorem 1, we see from (4.43) that

$$\lim_{n\to\infty} D^\alpha U_{1n}(x,t)^\kappa U_{1n}(x,t) = 0 \quad \text{a.e. in } \tilde{\Omega}\backslash\tilde{\Omega}(U) \text{ for } |\alpha| \leq m.$$

Consequently, from (f-2) and this last fact, it follows that

$$\lim_{n\to\infty} \frac{f_1(x,t,\xi_m(u_{1n}), \dots, \xi_m(u_{N_1 n}))U_{1n}(x,t)}{\|u_{1n}\|_{\mathcal{L}}^{\kappa}} = 0 \quad \text{a.e. in } \tilde{\Omega}\backslash\tilde{\Omega}(U).$$

Hence, it follows from (4.46) and Vitali's theorem [7, p. 325] that

$$\lim_{n\to\infty} \int_{\tilde{\Omega}\backslash\tilde{\Omega}(U)} \frac{f_1(x,t,\xi_m(u_{1n}), \dots, \xi_m(u_{N_1 n}))U_{1n}}{\|u_{1n}\|_{\mathcal{L}}^{\kappa}} = 0. \tag{4.48}$$

Observing from (f-3) that meas $(\tilde{\Omega}\backslash\tilde{\Omega}_0) = 0$, we conclude from (4.36), (4.45), and (4.48) that

$$0 \leq \limsup_{n\to\infty} \int_{\tilde{\Omega}_1(U)} \frac{f_1(x,t,\xi_m(u_{1n}), \dots, \xi_m(u_{N_1 n}))U_{1n}}{\|u_{1n}\|_{\mathcal{L}}^{\kappa}}. \tag{4.49}$$

Next, set

$$\tilde{\Omega}_2(U) = \{(x,t) \in \tilde{\Omega}_1(U) : \lim_{n\to\infty} D^\alpha U_{1n}(x,t) = D^\alpha U(x,t) \text{ for } |\alpha| \leq m\}. \tag{4.50}$$

From (4.43), we have $\operatorname{meas}(\tilde{\Omega}_1(U)\backslash\tilde{\Omega}_2(U)) = 0$. We consequently obtain from (4.49) that

$$0 \leq \limsup_{n\to\infty} \int_{\tilde{\Omega}_2(U)} \frac{f_1(x,t,\xi_m(u_{1n}),\ldots,\xi_m(u_{N_1 n}))U_{1n}}{\|u_{1n}\|_{\mathcal{L}}^{\kappa}}. \tag{4.51}$$

Next, we observe from (4.36) and (4.50) that

$$\begin{aligned} &\text{(i)} \quad U(x,t) \neq 0 \text{ for } (x,t) \in \tilde{\Omega}_2(U), \\ &\text{(ii)} \quad \lim_{n\to\infty} \xi_m(U_{1n}(x,t)) = \xi_m(U(x,t)) \text{ for } (x,t) \in \tilde{\Omega}_2(U). \end{aligned} \tag{4.52}$$

Let $(x,t) \in \tilde{\Omega}_2(u)$. It follows from (4.52) that $|U_{1n}(x,t)| > 0$ for $n \geq n_0(x,t)$. Consequently $(\xi_m(U_{1n}(x,t))/|\xi_m(U_{1n}(x,t))|) \in S_1'$ as defined in (f-3) for $n \geq n_0(x,t)$. Also from (4.52) for $(x,t) \in \tilde{\Omega}_2(U)$,

$$\lim_{n\to\infty} \frac{\xi_m(U_{1n}(x,t))}{|\xi_m(U_{1n}(x,t))|} = \frac{\xi_m(U(x,t))}{|\xi_m(U(x,t))|}, \tag{4.53}$$

where the right-hand side of (4.53) is in S_1'. Next, set

$$r_n = \|u_{1n}\|_{\mathcal{L}} |\xi_m(U_{1n}(x,t))| \quad \text{for } (x,t) \in \tilde{\Omega}(U). \tag{4.54}$$

It follows from (4.6) and (4.52) that $\lim_{n\to\infty} r_n = \infty$. Consequently, from (f-2), (4.53), and (4.54) we have that

$$\lim_{n\to\infty} \frac{f_1\left(x,t,r_n\frac{\xi_m(U_{1n})}{|\xi_m(U_{1n}(x,\ t))|},\ldots,\xi_m(u_{N_1 n})\right)}{r_n^{\kappa}} = h_1\left(x,t,\frac{\xi_m(U(x,t))}{|\xi_m(U(x,t))|}\right)$$

for $(x,t) \in \tilde{\Omega}_2(U)$. Observing that

$$\xi_m(u_{1n}(x,t)) = |\xi_m(U_{1n}(x,t))| \frac{\xi_m(u_{1n}(x,t))}{|\xi_m(U_{1n}(x,t))|} = r_n \frac{\xi_m(U_{1n}(x,t))}{|\xi_m(U_{1n}(x,t))|},$$

we see that it then follows from (4.53), (4.44), and this last fact that

$$\begin{aligned} &\lim_{n\to\infty} \frac{f_1(x,t,\xi_m(u_{1n}(x,t)),\ldots,\xi_m(u_{N_1 n}))U_{1n}(x,t)}{\|u_{1n}\|_{\mathcal{L}}^{\kappa}} \\ &\quad = h_1\left(x,t,\frac{\xi_m(U(x,t))}{|\xi_m(U(x,t))|}\right) |\xi_m(U(x,t))|^{\kappa} U(x,t) \end{aligned} \tag{4.55}$$

for $(x,t) \in \tilde{\Omega}_2(U)$. We conclude from (4.46), (4.55), and Vitali's theorem that

$$\begin{aligned} &\lim_{n\to\infty} \int_{\tilde{\Omega}_2(U)} \frac{f_1(x,t,\xi_m(u_{1n}(x,t)),\ldots,\xi_m(u_{N_1 n}(x,t)))U_{1n}(x,t)}{\|u_{1n}\|_{\mathcal{L}}^{\kappa}} \\ &\quad = \int_{\tilde{\Omega}_2(U)} h_1\left(x,t,\frac{\xi_m(U)}{|\xi_m(U)|}\right) |\xi_m(U)|^{\kappa} U. \end{aligned} \tag{4.56}$$

From (4.51) we conclude that the integral on the right-hand side of (4.56) is nonnegative, that is, ≥ 0. But $U \in S_{\lambda_1}^{\#}$ and U is nontrivial by (4.35). Also meas

$(\tilde{\Omega}(U)\backslash\tilde{\Omega}_2(U)) = 0$. Hence by (1.20) the integral on the right-hand side of (4.56) is negative, that is, strictly < 0. We have arrived at a contradiction. We conclude that (4.6) is false and that (4.5) is indeed true when $i = 1$. A proof similar to that above prevails for $i = 2, \ldots, N_1$. Hence, we have that (4.5) holds for all values of i.

Next, we return to (4.19) and obtain from (4.3), (4.5), (f-2), (G-1), and (G-2) that a positive constant K_5 exists such that

$$\|D_t u_{in}\|_{L^2} \le K_5 \|D_t u_{in}\|_{L^2}^{\kappa} + K_5 \quad \text{for } i = 1, \ldots, N_1 \text{ and } n = 1, 2, \ldots.$$

But $0 \le \kappa < 1$. Therefore, from this last inequality, we obtain that $\{\|D_t u_{in}\|_{L^2}\}_{n=1}^{\infty}$ is a bounded sequence for $i = 1, \ldots, N_1$. This fact coupled with (4.5) implies that

$$\|u_{in}\|_{\tilde{H}} \le K_6 \quad \text{for } i = 1, \ldots, N_1 \text{ and } n = 1, 2, \ldots, \tag{4.57}$$

where K_6 is a positive constant.

We next use Lemma A in §6 (the Appendix) and standard Hilbert space theory to obtain the following from (4.57): There exists a subsequence of $\{(u_{1n}, \ldots, u_{N_1 n})\}_{n=1}^{\infty}$ (which for ease of notation we take to be the full sequence) and functions

$$(u_1, \ldots, u_{N_1}) \in \tilde{H}^{N_1} \tag{4.58}$$

such that

$$\lim_{n\to\infty} \|D^{\alpha} u_{in} - D^{\alpha} u_i\|_{L^2} = 0 \quad \text{for } |\alpha| \le m - 1, \tag{4.59}$$

$$\lim_{n\to\infty} D^{\alpha} u_{in}(x,t) = D^{\alpha} u_i(x,t) \quad \text{a.e. in } \tilde{\Omega} \text{ for } |\alpha| \le m - 1, \tag{4.60}$$

$$\lim_{n\to\infty} \langle D_t u_{in}, w\rangle_{L^2} = \langle D_t u_i, w\rangle_{L^2} \quad \forall w \in L^2, \tag{4.61}$$

$$\lim_{n\to\infty} \langle D^{\alpha} u_{in}, w\rangle_{L^2} = \langle D^{\alpha} u_i, w\rangle_{L^2} \quad \forall w \in L^2 \text{ and } |\alpha| = m, \tag{4.62}$$

for $i = 1, \ldots, N_1$.

We next propose to show that for a subsequence (4.60) also holds for $|\alpha| = m$, that is,

$$\begin{gathered} \exists \{u_{in_k}\}_{k=1}^{\infty} \text{ such that } \lim_{k\to\infty} \zeta_m(u_{in_k}(x,t)) = \zeta_m(u_i(x,t)) \\ \text{a.e. in } \tilde{\Omega} \text{ for } i = 1, \ldots, N_1. \end{gathered} \tag{4.63}$$

Once (4.63) is shown, it will be an easy matter when this fact is coupled with (4.60) to show that the conclusion of Theorem 1 holds.

To establish (4.63) for the case $i = 1$, it is sufficient to establish the following two facts:

$$\begin{gathered} \exists \text{ a subsequence } \{u_{1n_k}\}_{k=1}^{\infty} \text{ such that} \\ \lim_{k\to\infty} \sum_{|\alpha|=m} [A_{\alpha}(x, \eta_{m-1}(u_{1n_k}), \zeta_m(u_{1n_k})) - A_{\alpha}(x, \eta_{m-1}(u_{1n_k}), \zeta_m(u_1))] \\ \cdot [D^{\alpha} u_{1n_k} - D^{\alpha} u_1] = 0 \quad \text{a.e. in } \tilde{\Omega}; \end{gathered} \tag{4.64}$$

$$\begin{array}{l}\text{With } \{u_{1n_k}\}_{k=1}^{\infty} \text{ designating the same subsequence as in (4.64)}\\ \{|\zeta_m(u_{1n_k}(x,t))|\}_{k=1}^{\infty} \text{ is pointwise bounded for a.e. } (x,t) \text{ in } \tilde{\Omega}.\end{array} \tag{4.65}$$

Using (4.60) in conjunction with (4.64) and (4.65) to obtain (4.63) via the monotonicity assumption (Q-3) is standard fare in nonlinear partial differential equations (see [4, p. 30] or [13, p. 104]). It therefore remains to show that (4.64) and (4.65) hold.

To establish (4.64), the key idea is to show that the summation in (4.64) with u_{1n_k} replaced by u_{1n} converges to zero in L^1-norm over $\tilde{\Omega}$ and then to apply the monotonicity condition (Q-3) in conjunction with a standard theorem in Lebesgue theory to reach the conclusion in (4.64). The technique for accomplishing this follows the lines of [17, pp. 1830–1833] and [12, pp. 168–171] and makes use of Lemma 1. We leave the details to the reader.

Next, we show that (4.65) for $i = 1$ follows from (4.60) and (4.64). To accomplish this, we observe from (Q-4) that

$$c_1|\zeta_m(u_{1n}(x,t))|^2 \leq \sum_{|\alpha|=m} A^1_\alpha(x, \eta_{m-1}(u_{1n}), \zeta_m(u_{1n}))D^\alpha u_{1n} \tag{4.66}$$

for a.e. $(x,t) \in \tilde{\Omega}$ where $c_1 > 0$. Also, we see that

$$\begin{aligned} &A^1_\alpha(x, \eta_{m-1}(u_{1n}), \zeta_m(u_{1n}))D^\alpha u_{1n}(x,t)\\ &\quad = A^1_\alpha(x, \eta_{m-1}(u_{1n}), \zeta_m(u_{1n}))D^\alpha u_1(x,t)\\ &\quad + A^1_\alpha(x, \eta_{m-1}(u_{1n}), \zeta_m(u_1))[D^\alpha u_{1n}(x,t) - D^\alpha u_1(x,t)]\\ &\quad + [A^1_\alpha(x, \eta_{m-1}(u_{1n}), \zeta_m(u_{1n}) - A^1_\alpha(x, \eta_{m-1}(u_{1n}), \zeta_m(u_1)))]\\ &\quad \times [D^\alpha u_{1n}(x,t) - D^\alpha u_1(x,t)]. \end{aligned} \tag{4.67}$$

If (x,t) is a point such that (4.60) holds for $|\alpha| \leq m-1$, where $D^\alpha u_1(x,t)$ is finite-valued for $|\alpha| \leq m$, and such that (4.64) holds, and furthermore such that the inequalities given by (Q-2) hold, then it follows on dividing both sides of (4.66) by $|\zeta_m(u_{1n}(x,t))|^{\frac{3}{2}}$ and using (4.67) that a subsequence of $|\zeta_m(u_{1n_k}(x,t))|$ cannot tend to $+\infty$, for then we would have that this same subsequence also tends to zero. Since these aforementioned conditions hold almost everywhere in $\tilde{\Omega}$, (4.65) is established for $i = 1$. A similar argument prevails for $i = 1, \dots, N_1$. Hence, (4.65) holds for $i = 2, \dots, N_1$. As we have stated earlier, (4.63) then holds because it is a standard consequence of (4.64) and (4.65).

We now show that (4.57)–(4.63) along with the fact that $\{(u_{1n}, \dots, u_{N_1 n})\}_{n=1}^{\infty}$ satisfies (3.9) gives (1.21), which is our desired result.

To accomplish this, let n_3 be a fixed but arbitrary positive integer and let $w \in S_{n_3}$. Then it follows that

$$\text{(3.9) holds for } i = 1, \dots, N_1 \text{ and } n \geq n_3 \text{ when } v \text{ is replaced by } w. \tag{4.68}$$

We observe that (f-1), (4.60), and (4.63) imply that

$$f_i(x,t,\xi_m(u_{1n}), \dots, \xi_m(u_{N_1 n}))w(x,t) \underset{n\to\infty}{\longrightarrow} f_i(x,t,\xi_m(u_1), \dots, \xi_m(u_{N_1}))w(x,t)$$

for a.e. $(x,t) \in \tilde{\Omega}$. Also, we see from (f-2), (4.57), and the fact that $w \in L^2(\tilde{\Omega})$ that $\{f_i(x,t,\xi_m(u_{1n}), \dots, \xi_m(u_{N_1 n}))w\}_{n=1}^{\infty}$ is an absolutely equi-integrable sequence for $i = 1, \dots, N_1$. Hence it follows from Vitali's theorem [7, p. 325] that

$$\lim_{n\to\infty} \langle f_i(\cdot, \xi_m(u_{1n})), \dots, \xi_m(u_{N_1 n}), w\rangle_{L^2} = \langle f_i(\cdot, \xi_m(u_1), \dots, \xi_m(u_{N_1})), w\rangle_{L^2}. \tag{4.69}$$

In a similar manner using (Q-1) and (Q-2), it follows that

$$\begin{aligned}\lim_{n\to\infty}\langle A^i_\alpha(\cdot,\xi_m(u_{in})),D^\alpha w\rangle_{L^2}&=\langle A^i_\alpha(\cdot,\xi_m(u_i)),D^\alpha w\rangle_{L^2},\\ \lim_{n\to\infty}\langle a^i_0(\cdot,\xi_m(u_{in}))u_{in},w\rangle_{L^2}&=\langle a^i_0(\cdot,\xi_m(u_i))u_i,w\rangle_{L^2}\end{aligned}\tag{4.70}$$

for $i=1,\dots,N_1$.

Also, we have from (4.61) that $\lim_{n\to\infty}G_{1i}(D_tu_{in})=G_{1i}(D_tu_i)$ and from (4.59) that $\lim_{n\to\infty}G_{2i}(u_{in})=G_{2i}(u_i)$. In a similar manner, we have that $\lim_{n\to\infty}\langle D_tu_{in}, w\rangle_{L^2}=\langle D_tu_i,w\rangle_{L^2}$. Putting all these facts together, we conclude from (4.68), (1.4), (4.60), and (4.68)–(4.70) upon taking the limit as $n\to\infty$ of both sides of (3.9) with v replaced by w that

(4.71)

$$\begin{aligned}\langle D_tu_i,w\rangle_{L^2}+\rho_iQ_i(u_i,w)&=\lambda^*_1\rho_i\langle u_i,w\rangle_{L^2}\\ &\quad+\langle f_i(\cdot,\xi_m(u_1),\dots,\xi_m(u_{N_1})),w\rangle_{L^2}\\ &\quad+G_{1i}(D_tu_1)G_{2i}(u_i)\langle g_i,w\rangle_{L^2}\quad\text{for } i=1,\dots,N_1,\end{aligned}$$

and for $w\in S_{n_3}$. However, n_3 was an arbitrary positive integer. So we see that (4.71) holds $\forall w\in\bigcup_{n=1}^{\infty}S_n$.

Next, let $v\in\tilde H$. Then it follows from (4.71) and our last observation that (4.71) holds when w is replaced by $\sigma_n(v)$ where $\sigma_n(v)$ is defined in (3.4). From (3.5) (ii) and (1.4), it follows that $\lim_{n\to\infty}Q_i(u_i,\sigma_n(v))=Q_i(u_i,v)$. We consequently obtain from (3.5) (i) and this last remark upon replacing w in (4.71) by $\sigma_n(v)$ and taking the limit of both sides as $n\to\infty$ that

$$\begin{aligned}\langle D_tu_i,v\rangle_{L^2}+\rho_1Q_i(u_i,v)&=\lambda^*_1\rho_i\langle u_i,v\rangle_{L^2}+\langle f_i(\cdot,\xi_m(u_1),\dots,\xi_m(u_{N_1})),v\rangle_{L^2}\\ &\quad+G_{1i}(D_tu_i)G_{2i}(u_i)\langle g_i,v\rangle_{L^2}\quad\forall v\in\tilde H.\end{aligned}$$

But this is precisely (1.21), and the proof of Theorem 1 is complete.

5. Proof of Theorem 2. To prove Theorem 2 we can assume as in the proof of Theorem 1 that (4.1) and (4.2) hold. As a consequence we have that (4.3) and (4.4) hold. Next we invoke Remark 1 and have a sequence $\{(u_{1n},\dots,u_{N_1n})\}_{n=1}^{\infty}$ with each $u_{in}\in S_n$ such that

$$\begin{aligned}\langle D_tu_{in},v\rangle_{L^2}+\rho_iQ_i(u_{in},v)&=\rho_i(\lambda^*_1-\delta_i)\langle u_{in},v\rangle_{L^2}\\ &\quad+\langle f_i(\cdot,(\xi_m(u_{1n}),\dots,\xi_m(u_{N_1n}))),v\rangle_{L^2}\\ &\quad+G_{1i}(D_tu_{in})G_{2i}(u_{in})\langle g_i,v\rangle_{L^2}\\ &\quad\forall v\in S_n\text{ and } i=1,\dots,N_1.\end{aligned}\tag{5.1}$$

We claim there is a constant K_1 such that

$$\|u_{in}\|_{\mathcal L}\le K_1\quad\forall n\text{ and for } i=1,\dots,N_1.\tag{5.2}$$

For ease of notation, we shall establish (5.2) for the case $i=1$. A similar proof will prevail for the other values of i. Suppose then that (5.2) is false for $i=1$. Then with no loss in generality we can suppose

$$\lim_{n\to\infty}\|u_{1n}\|_{\mathcal L}=\infty.\tag{5.3}$$

We propose to show that (5.3) is false, and we shall do this without using (f-3) and (1.20). First, we observe as in the proof of Theorem 1 that $D_t u_{1n} \in S_n$. Hence taking $v = D_t u_{1n}$ in (5.1) for $i = 1$, we obtain that (4.19) holds and, as a consequence, since (4.6) is the same as (5.3), that (4.20) holds. We record this as

$$\lim_{n\to\infty} \frac{\|D_t u_{1n}\|_{L^2}}{\|u_{1n}\|_{\mathcal{L}}} = 0. \tag{5.4}$$

Next, we claim that $\exists K_2 > 0$ such that

$$\|u_{1n}\|_{L^2} \leq K_2 \|u_{1n}\|_{\mathcal{L}}^{\kappa} \quad \forall n. \tag{5.5}$$

To see that (5.5) is true, we first observe from (5.1) with $v = u_{1n}$ and $i = 1$ that

$$\begin{aligned} \rho_1 \mathcal{Q}_1(u_{1n}, u_{1n}) = {} & \rho_1(\lambda_1^* - \delta_1)\|u_{1n}\|_{L^2}^2 \\ & + \langle f_1(\cdot, (\xi_m(u_{1n}), \ldots, \xi_m(u_{N_1 n}))), u_{1n}\rangle_{L^2} \\ & + G_{11}(D_t u_{1n}) G_{21}(u_{1n}) \langle g_1, u_{1n}\rangle_{L^2}. \end{aligned} \tag{5.6}$$

If $\|u_{1n}\|_{L^2} \to \infty$, it follows from (1.5) that

$$\rho_1\left(\lambda_1^* - \frac{\delta_1}{2}\right)\|u_{1n}\|_{L^2}^2 \leq \rho_1 \mathcal{Q}_1(u_{1n}, u_{1n}) \quad \text{for } n \geq n_2.$$

But then from (5.4) and (5.6) we have that there is a constant K_3 such that

$$\rho_1 \frac{\delta_1}{2}\|u_{1n}\|_{L^2}^2 \leq K_3 \|u_{1n}\|_{\mathcal{L}}^{\kappa}\|u_{1n}\|_{L^2} + K_3\|u_{1n}\|_{L^2} \quad \text{for } n \geq n_2, \tag{5.7}$$

where we have also used (f-2), (G-1), and (G-2) to establish this last inequality. From (5.3) and (5.7), we infer that (5.5) is indeed true.

Now we have that Q_1 is $\kappa*$-related to L. So it follows from (1.16) (i) and (4.3) that for $n \geq n_3$

$$\|u_{1n}\|_{\mathcal{L}}^2 \leq \mathcal{Q}_1(u_{1n}, u_{1n}) + \delta_1 \|u_{1n}\|_{\mathcal{L}}^{1+\kappa}.$$

Using this last inequality in conjunction with (f-2), (G-1), (G-2), and (5.3)–(5.6), we obtain that there is a constant K_4 such that

$$\|u_{1n}\|_{\mathcal{L}}^2 \leq K_4[\|u_{1n}\|_{\mathcal{L}}^{2\kappa} + \|u_{1n}\|_{\mathcal{L}}^{1+\kappa}] \quad \text{for } n \geq n_3.$$

As a consequence of this last inequality and the fact that $0 \leq \kappa < 1$, we obtain that $\|u_n\|_{\mathcal{L}}^{1-\kappa} \leq 2K_4$ for $n \geq n_3$. But this is a contradiction to (5.3). Hence (5.2) does indeed hold for $i = 1$. A similar proof prevails for $i = 2, \ldots, N_1$. Hence (5.2) holds for $i = 1, \ldots, N_1$.

But this last situation is the same as (4.5) in Theorem 1. After (4.5) was established, the rest of the proof of Theorem 1 did not depend upon (f-3) and the condition in (1.20). So the remainder of the proof of Theorem 2 follows exactly the lines of Theorem 1 and the proof of Theorem 2 is complete.

6. Appendix.

In this section, we shall prove the following lemma.

LEMMA A. *Let $\{u_n\}_{n=1}^{\infty}$ be a sequence of elements in $\tilde{H}$ with*

$$\|u_n\|_{\tilde{H}} \leq K \quad for\, n = 1, 2, \ldots. \tag{6.1}$$

Then $\exists u \in \tilde{H}$ *and a subsequence* $\{u_{n_q}\}_{q=1}^{\infty}$ *such that*

$$\lim_{n\to\infty} \int_{\tilde{\Omega}} |D^\alpha u_{n_q} - D^\alpha u|^2 = 0 \quad \textit{for}\, |\alpha| \le m-1.$$

For $m = 1$, the above lemma is an easy consequence of the familiar compact imbedding theorem for Sobolev spaces [1, p. 97]. Therefore, to prove the lemma, we can assume from the start that the m used in the definition $\tilde{H}$ is a positive integer with $m \ge 2$. Also since for $u_n \in \tilde{H}, \exists v_n \in \mathcal{A}$ such that $\|u_n - v_n\|_{\tilde{H}} \le n^{-1}$, we can assume from the start that

$$(6.2) \qquad u_n \in \mathcal{A} \quad \text{for } n = 1, 2 \ldots.$$

Furthermore, since Ω is a bounded open set, we can suppose that Ω is contained in the interior of an open cube with each side of length 2π. In particular, we shall suppose $\Omega \subset (-\pi,\pi)^N$. Since by (6.2) each $u_n \in \mathcal{A}$, we have that for each $n, \exists$ a compact set $E_n \subset (-\pi,\pi)^N$ such that $u_n(x,t) = 0$ in $(-\pi,\pi)^N \backslash E_n$ and $t \in \mathbb{R}$. Also, $u_n(x,t)$ is $C^\infty((-\pi,\pi)^N \times \mathbb{R})$ and periodic of period 2π in t. Hence we can view each $u_n(x,t)$ as a real-valued function in $C^\infty(T^N \times T) = C^\infty(T^{N+1})$, where T^N is the N-dimensional torus.

Next, we introduce the complete orthogonal trigonometric series $\{e^{i[(j,x)+kt]}\}$, where $j = (j_1, \ldots, j_N)$ with $j_1, \ldots, j_N$ and k integers running from $-\infty$ to ∞ and $(j,x) = j_1x_1 + \cdots + j_Nx_N$. We set $\hat{u}_n(j,k) = (2\pi)^{-(N+1)} \int_T \int_{T^N} u_n(x,t)e^{i[(j,x)+kt]}$. From (6.1) and [9, p. 55] we have that

$$(6.3) \qquad \sum_{(j,k)\in M} [|j|^{2m} + k^2]|\hat{u}_n(j,k)|^2 \le K^2 \quad \forall n = 1, 2, \ldots$$

where $|j|^2 = (j,j)$ and $M = \{(j,k) : -\infty < j_l < \infty, -\infty < k < \infty, l = 1, \ldots, N\}$ is the set of integral lattice points in $\mathbb{R}^{N+1}$. Also for future use we record the fact that we are assuming

$$(6.4) \qquad m \ge 2 \text{ is a positive integer.}$$

Since $\int_{T^{N+1}} |u_n|^2 \le K^2$, we see there exists a subsequence (which for ease of notation, we take to be the full sequence) and a $u \in L^2(T^N \times T)$ such that

$$\lim_{n\to\infty} \int_T \int_{T^N} u_n w = \int_T \int_{T^N} uw \quad \forall w \in L^2(T^N \times T).$$

In particular, taking $w = e^{-i[(j,x)+kt]}$, we have that

$$(6.5) \qquad \lim_{n\to\infty} \hat{u}_n(j,k) = \hat{u}(j,k) \quad \forall (j,k) \in M.$$

Next, let $\rho > 0$ be any positive number. It follows from (6.3) and (6.5) that $\sum_{|j|^2+k^2\le\rho} [|j|^{2m} + k^2]|\hat{u}(j,k)|^2 \le K^2$. Hence,

$$(6.6) \qquad \sum_{(j,k)\in M} [|j|^{2m} + k^2]|\hat{u}(j,k)|^2 \le K^2.$$

We therefore conclude that $D^\alpha u$ exists in the distributional sense [3, Chap. 3] and $D^\alpha u \in L^2(T^N \times T)$ for $|\alpha| \leq m$. The conclusion to Lemma A will follow if we can show

$$\lim_{n\to\infty} \int_T \int_{T^N} |D^\alpha u_n - D^\alpha u|^2 = 0 \quad \text{for } |\alpha| \leq m-1. \tag{6.7}$$

Now from a Fourier series point of view

$$D^\alpha u_n - D^\alpha u = \sum_{(j,k)\in M} (ij_1)^{\alpha_1} \dots (ij_N)^{\alpha_N} [\hat{u}_n(j,k) - \hat{u}(j,k)] e^{i[(j,x)+kt]}.$$

Therefore (6.7) will follow if we can show

$$\lim_{n\to\infty} \sum_{(j,k)\in M} j_1^{2\alpha_1} \dots j_N^{2\alpha_n} |\hat{u}_n(j,k) - \hat{u}(j,k)|^2 = 0 \quad \text{for } |\alpha| \leq m-1. \tag{6.8}$$

Now

$$j_1^{2\alpha_1} \dots j_N^{2\alpha_N} \leq j_1^{2(m-1)} + \dots + j_N^{2(m-1)} \leq N|j|^{2(m-1)} \quad \text{for } |\alpha| \leq m-1 \text{ and } (j,k) \in M.$$

Hence (6.8) will follow if we can show

$$\lim_{n\to\infty} \sum_{(j,k)\in M} |j|^{2(m-1)} |\hat{u}_n(j,k) - \hat{u}(j,k)|^2 = 0. \tag{6.9}$$

We now show that (6.9) follows from (6.3)–(6.5).

To accomplish this, we first set

$$y(s) = s^m - \rho^{\frac{1}{m}} s^{m-1} + \rho, \tag{6.10}$$

where $\rho \geq 2$ and $m \geq 2$. Then

$$y(s) > 0 \quad \text{for } 0 \leq s \leq \rho. \tag{6.11}$$

To see that this is the case, we observe that both $y(0)$ and $y(\rho)$ are positive numbers. Also the only place where $y'(s) = 0$ inside the interval $(0, \rho)$ is when $s = ((m-1)\rho^{1/m})/m$. But

$$y\left[\frac{(m-1)\rho^{\frac{1}{m}}}{m}\right] \geq \left[1 - \left(\frac{m-1}{m}\right)^{m-1}\right] \rho > 0.$$

Hence, $y(s)$ does not have a nonpositive minimum inside $(0, \rho)$ and (6.11) is established.

To show that (6.9) is indeed true let I_n be the summation in (6.9) and let $\varepsilon > 0$ be given. (6.9) will follow if we can show

$$\overline{\lim_{n\to\infty}} I_n \leq 2\varepsilon. \tag{6.12}$$

Choose $\rho \geq 2$ such that also

$$\frac{4K^2}{\rho^{\frac{1}{m}}} \leq \varepsilon, \tag{6.13}$$

and set

$$
\begin{aligned}
&M_1(\rho) = \{(j,k) \in M : |j|^2 + k^2 \leq 2\rho\},\\
(6.14) \qquad &M_2(\rho) = \{(j,k) \in M : |j|^2 > \rho\},\\
&M_3(\rho) = \{(j,k) \in M : |j|^2 \leq \rho, k^2 > \rho\}.
\end{aligned}
$$

It is clear that $M = M_1(\rho) \cup M_2(\rho) \cup M_3(\rho)$ because $[M_1(\rho) \cup M_2(\rho)]^C \subset M_3(\rho)$. Set

$$(6.15) \qquad I_n^{(p)} = \sum_{(j,k)\in M_p(\rho)} |j|^{2(m-1)}|\hat{u}_n(j,k) - \hat{u}(j,k)|^2, \qquad p = 1,2,3.$$

Therefore

$$(6.16) \qquad I_n \leq I_n^{(1)} + I_n^{(2)} + I_n^{(3)}.$$

Since there are only a finite number of lattice points in $M_1(\rho)$, it follows from (6.5) and (6.15) that

$$(6.17) \qquad \lim_{n\to\infty} I_n^{(1)} = 0.$$

Also from the fact that $|a - b|^2 \leq 2a^2 + 2b^2$, it follows from (6.3) and (6.6) that

$$(6.18) \qquad \sum_{(j,k)\in M} [|j|^{2m} + k^2]|\hat{u}_n(j,k) - \hat{u}(j,k)|^2 \leq 4K^2.$$

This fact coupled with the fact that

$$\frac{|j|^{2(m-1)}}{|j|^{2m} + k^2} \leq \rho^{-1} \quad \text{for } (j,k) \in M_2$$

give that $\overline{\lim}_{n\to\infty} I_n^{(2)} \leq \frac{4K^2}{\rho}$. But both $\rho \geq 2$ and $m \geq 2$. It follows therefore from (6.13) that

$$(6.19) \qquad \overline{\lim}_{n\to\infty} I_n^{(2)} \leq \varepsilon.$$

Next, we observe from (6.14) that

$$(6.20) \qquad \frac{|j|^{2(m-1)}}{|j|^{2m} + k^2} \leq \frac{|j|^{2(m-1)}}{|j|^{2m} + \rho}$$

for $(j,k) \in M_3(\rho)$. Also, we observe from (6.10) and (6.11) that for $|j|^2 \leq \rho$, the expression on the right-hand side of the inequality in (6.20) in majorized by $\rho^{-\frac{1}{m}}$. Therefore we have that

$$\frac{|j|^{2(m-1)}}{|j|^{2m} + k^2} \leq \rho^{-\frac{1}{m}} \quad \text{for } (j,k) \in M_3(\rho).$$

Consequently, we obtain from (6.15), (6.18), and this last inequality that

$$\overline{\lim_{n\to\infty}} I_n^{(3)}(\rho) \leq \frac{4K^2}{\rho^{1/m}}.$$

But this fact coupled with (6.13), (6.16), (6.17), and (6.19) give that $\overline{\lim}_{n\to\infty} I_n \leq 2\varepsilon$. Hence, (6.12) is established and the proof of Lemma A is complete.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces,* Academic Press, San Diego, CA, 1978.

[2] E. BELTRAMI, *Mathematics for Dynamic Modeling,* Academic Press, San Diego, CA, 1987.

[3] L. BERS, F. JOHN, AND M. SCHECTER, *Partial differential equation,* in Lectures in Appl. Math., John Wiley & Sons, New York, 1964.

[4] F. E. BROWDER, *Existence theorem in partial differential equation,* in Proc. Sympos. Pure Math., vol. XVI, American Mathematical Society, Providence, RI, 1970, pp. 1–60.

[5] D. G. DE FIGUEIREDO, *The Dirichlet problem for nonlinear elliptic equations: A Hilbert space approach,* in Partial Differential Equations and Related Topics, Lecture Notes in Mathematics 4, Springer-Verlag, New York, 1977.

[6] D. G. DE FIGUEIREDO AND J. P. GOSSEZ, *Nonlinear partial differential equations near the first eigenvalue,* J. Differential Equations, 30 (1978), pp. 1–19.

[7] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part* I, *General Theory,* John Wiley & Sons, New York, 1988.

[8] A. FRIEDMAN, *Partial Differential Equations,* Holt, Rinehart and Winston, New York, 1969.

[9] F. JOHN, *Partial Differential Equations,* 4th Ed., Springer-Verlag, New York, 1982.

[10] S. KESAVAN, *Topics in Functional Analysis and Applications,* John Wiley & Sons, New York, 1989.

[11] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications,* Academic Press, San Francisco, CA, 1980.

[12] L. LEFTON AND V. L. SHAPIRO, *Resonance and quasilinear parabolic partial differential equations,* J. Differential Equations, 101 (1993), pp. 148–177.

[13] J. LERAY AND J. L. LIONS, *Quelques résultats de Višik sur les problèms elliptiques non linéaires par les méthodes de Minty-Browder,* Bull. Soc. Math. France, 93 (1965), pp. 97–107.

[14] J. D. MURRAY, *How the leopard gets its spots,* Scientific American, 258 (1988), pp. 80–87.

[15] ———, *Mathematical Biology,* Springer-Verlag, New York, 1990.

[16] L. NIRENBERG, *Topics in Nonlinear Functional Analysis,* Lecture Notes, Courant Institute of Mathematical Sciences, New York University, New York, 1974.

[17] V. L. SHAPIRO, *Quasilinear ellipticity and the* 1*st eigenvalue,* Comm. Partial Differential Equations, 16 (1991), pp. 1819– 1855.

[18] ———, *Resonance, distributions and semilinear elliptic partial differential equations,* Nonlinear Anal., 8 (1984), pp. 857–871.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 170–203, January 1996

010

PERIODIC AND POSITIVE WAVE FRONT SOLUTIONS OF SEMILINEAR DIFFUSION EQUATIONS*

JOSÉ M. FRAILE† AND JOSÉ SABINA DE LIS‡

Abstract. The objective of this work is to study the existence of positive and T-periodic undulatory solutions of the form $u(x,t) = u(x - c(t))$ to the equation

$$\frac{\partial u}{\partial t} = \Delta u + b(t)\nabla u + f(u), \quad x \in \Omega,$$

where $b = b(t)$ is T-periodic and the reaction term $f = f(u)$ is supposed to satisfy $f(u) \le 0$ for $0 < u_o \le u$. The waves are analyzed in bounded domains Ω wherein they are subject to special homogeneous Dirichlet conditions. The average μ of $b(t)$ over the interval $(0, T)$ and the size of Ω are observed as bifurcation parameters. The one-dimensional version of this Dirichlet problem is deeply and geometrically studied by means of chains of travelling waves to the equation $u_t = u_{xx} + f(u)$, which connect (infinitely many in some cases) zeros of $f(u)$.

Key words. travelling waves, subsolutions, supersolutions, semilinear elliptic equations, variational methods, phase space analysis

AMS subject classifications. 35K57, 35B10, 35J65, 34B15, 34A26

1. Introduction. In this work we analyze the existence and some properties of bifurcation of a certain kind of periodic undulatory solution to semilinear parabolic equations, with *periodic transport* terms, with the specific form,

$$\frac{\partial u}{\partial t} = \Delta u + b(t)\nabla u + f(u). \tag{1.1}$$

In (1.1)

$$\begin{aligned} u: \quad & \Omega \times \mathbf{R} \to \mathbf{R}, \\ & (x,t) \to u(x,t) \end{aligned}$$

where $\Omega \subset \mathbf{R}^n$ is a bounded domain (i.e., an open connected set), Δ and ∇ stand for the usual Laplacian $\Delta = \Sigma_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}$ and gradient $\nabla = (\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n})$ n-dimensional operators, respectively. We will assume that

$$\begin{aligned} b: \quad & \mathbf{R} \to \mathbf{R}^n \\ & t \to b(t) \end{aligned}$$

is a continuous and periodic function with period $T > 0$ and average $\mu = \frac{1}{T}\int_0^T b(\tau)\, d\tau$. As for $f = f(u)$ it will be assumed that it belongs to a certain general kind of C^1 nonlinearities (see §3.1).

* Received by the editors October 27, 1993; accepted for publication (in revised form) June 3, 1994. The research contained in this paper was partially supported by Direccion General de Investigación Científica y Tecnología PB90-0258.

† Departamento de Matemática Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain.

‡ Departamento de Análisis Matemático, Universidad de La Laguna, 38271 La Laguna, Spain.

The main objective of the work is to study the undulatory solutions

$$w(x,t) = u(x - c(t)) = u(y), \quad x \in \Omega, \quad t \in \mathbf{R} \tag{1.2}$$

where $c = c(t)$ is a C^1 periodic function with period $T > 0$. Notice that the solutions with the form (1.2) are those profiles which are stationary regarding the mobil reference frame,

$$\begin{cases} y = x - c(t), \\ t' = t. \end{cases}$$

However, a suitable choice of $c(t)$ must be made.

The study of periodic undulatory solutions to reaction–diffusion systems has been a very active research area in recent years. The plane wave fronts, i.e., solutions with the form $w(x,t) = u(kx - ct) = u(\theta)$, $k \in \mathbf{R}^n$, $|k| = 1$, and $c \in \mathbf{R}$, have been perhaps the most studied. In this case the periodicity of the wave comes from the character of u as a periodic function of θ (a free oscillation of an ordinary differential equation). See for instance [Ba-P], [Co-H-M], [Du], [Fr-S,89], [Ko-H,73], [Ko-H,75], and the books by Murray [M] and Fife [F]. However, it should be remarked that the dimension of the variable x is not preserved in the phase $y = kx - ct$ of the plane wave fronts, except when $x \in \mathbf{R}$. Therefore, these kinds of waves are incompatible with any types of boundary conditions. If $x \in \mathbf{R}$, $u(x - ct)$ is called a *travelling wave* (see [F]) and their role as asymptotic states to parabolic equations (1.1) has been widely analyzed (see [Ar-W,75], [F-M,77], [F-M,81], [F], and [He] for the cases $b(t) \equiv 0$ and $\Omega = \mathbf{R}$).

Wave solutions with an n-dimensional phase y (as those of type (1.2)) have already been considered in the literature. For instance [A-A], [Be-L,83], [St] deal with $w(x,t) = u(x - ct)$, $x, c \in \mathbf{R}^n$ and $b(t) \equiv 0$ in (1.1). However, in [Fr-S,84], [S-F,87], and [S] a new aspect exhibited by the more general kinds of waves $u(x,t) = u(Kx - ct)$, with K an $n \times n$ matrix, $c \in \mathbf{R}^n$, was studied. Namely, the possibility of subjecting those waves to boundary conditions when they are observed in bounded domains $\Omega \subset \mathbf{R}^n$. In those references, boundary conditions of Dirichlet and Neumann type were defined on a certain piece Γ of the boundary $\partial\Omega$ and the bifurcation properties and asymptotic behaviour of the corresponding problems were analyzed under the assumptions $b(t) \equiv 0$ or $b(t) = \textit{constant}$. It should be remarked that such boundary value problems (BVPs) lead to the analysis of semilinear elliptic BVPs in cylindrical domains of $\mathbf{R}^n$ (the unboundedness being caused by the condition $t \in \mathbf{R}$). In the context of waves such types of cylindrical BVPs have been considered in [G] and more recently in [V] and [Ca-M-S].

In the present work the compatibility of the waves (1.2), $w(x,t) = u(x - c(t))$, with certain kinds of boundary conditions defined in bounded domains Ω is studied. However, such boundary conditions will be defined in a time T-periodic varying piece Γ_t of the boundary $\partial\Omega$, and the corresponding BVPs will be lead again to semilinear elliptic BVPs in bounded domains Ω_ω for the phase $y = x - c(t)$, rather than in cylindrical ones. Notice that the T-periodicity is an intrinsic property of $w(x,t)$ in (1.2). As for the applications, it should be remarked that the study of waves (1.2) under periodic convection has been suggested by several reaction–diffusion models related to respiratory phenomena occurring in living beings (see [B], [N-N-B], [Sh-A-S]). Specifically, the reacting substances (hemoglobin and hemoglobin compounds) are periodically pumped into the chemical reactor (the lungs' capillarities) in the model considered in [Sh-A-S]. Fluid dynamics and combustion theory also provide a natural source of models where the waves of type (1.2) under periodic convection could be observed [Li].

The organization of the material and the main results in the work are now described. The study of the boundary conditions to be imposed to the waves $w(x,t)$ is performed in §2. This study entails a careful analysis of the geometry of the boundary $\partial\Omega_\omega$ of the domain $\Omega_\omega = \{y = x - c(t)/x \in \Omega, t \in \mathbf{R}\}$ of the phase y to connect $\partial\Omega_\omega$ with a periodic region $\Gamma_t \subset \partial\Omega$. Precise conditions to determine Γ_t are also given in §2. The homogeneous Dirichlet conditions are introduced in §2.2 and the corresponding BVP for the waves (1.2) to equation (1.1) is stated there as follows:

$$\text{(DW)} \qquad \begin{cases} \frac{\partial w}{\partial t} = \Delta w + b(t)\nabla w + f(w), & (x,t) \in \Omega \times \mathbf{R}, \\ w(x,t) = u(x - c(t)), & \\ w(x,t) = 0, & t \in \mathbf{R}, x \in \Gamma_t. \end{cases}$$

In this work the existence and other qualitative properties of positive solutions $w = w(x,t)$ to (DW) are studied in detail. It is important to remark that the analysis of (DW) leads to the following semilinear elliptic problem:

$$\text{(D)} \qquad \begin{cases} \Delta u + \mu\nabla u + f(u) = 0 & \text{in } G, \\ u = 0 & \text{on } \partial G, \end{cases}$$

where $u(y)$ is the wave $w(x,t)$ observed in the phase y, $G = \Omega_\omega$ (the domain of the phase), and $\mu = \frac{1}{T}\int_0^T b(\sigma)\, d\sigma$. Thus, the results in this work are also concerned with the existence of positive solutions to (D) and its bifurcation properties regarding μ and the domain G.

An important role in the work is played by the one-dimensional version of (DW), which is introduced in §2.2 and leads to the one-dimensional version of (D)

$$\text{(D)}_1 \qquad \begin{cases} u_{yy} + \mu u_y + f(u) = 0, & y \in (-b, b), \\ u(-b) = u(b) = 0. \end{cases}$$

The main results of the paper are concerned with the one-dimensional Dirichlet problem for waves $w(x,t)$ and will be summarized below. Observe that the equation in (D)_1 is that satisfied by the travelling wave solutions $u(x,t) = u(x - \mu t)$ to semilinear diffusion equations of the form

$$\text{(1.3)} \qquad \frac{\partial u}{\partial t} = u_{xx} + f(u).$$

Therefore part of the results in this paper give existence and other qualitative properties of travelling waves to equation (1.3).

The statements of the results, the definitions to be used, and the class of nonlinearities to be considered are given in §3. Regarding $f = f(u)$ we will assume that $f(u) < 0$ for u greater than certain u_o. In addition we assume the positivity of $\int_0^u f(s)\, ds$ for certain $u = u_1 > 0$. However we should remark that $f(u)$ will be allowed to have infinitely many zeros in $u > 0$. The main features concerning the n-dimensional (DW) can be described as follows. (DW) does not admit positive solutions in any bounded domain Ω provided $|\mu|$ is greater than a certain positive value μ_1, which only depends on the nonlinearity f. On the other hand, it is possible to find domains Ω where (DW) exhibits positive solutions, provided $|\mu|$ is not too large. These facts motivate the introduction of the parameter μ_o which is the largest $|\mu|$ so that (DW) has a positive solution $w(x,t)$ in some domain Ω. The precise expression for μ_1, the positivity of μ_o, as well as an optimum a priori estimate for all positive solutions to (DW) are stated in Theorem 1, whose proof is postponed until §4.

The main results of the work focus on giving a geometrical interpretation of the value μ_o. This is done for the one-dimensional (DW). First, μ_o is identified as the propagation velocity of a monotone travelling wave when $f(u)$ is the logistic or the bistable nonlinearity (see Theorem 2). Actually, this result remains valid if the nonlinearity $f(u)$ exhibits a finite number of zeros. However, this is no longer true for general nonlinearities $f(u)$, in particular if $f(u)$ exhibits infinitely many zeros. Therefore, objects more general than monotone travelling waves must be introduced for describing μ_o. In Definition 2 of §3, the required generalization of travelling waves, the monotone chains of travelling waves, is presented. In the main result of the work, Theorem 3, μ_o is identified as the propagation velocity of a monotone chain to (1.3), which could connect infinitely many zeros of $f(u)$. Properties of continuity and monotonicity of monotone chains are stated in Theorem 4. The proof of Theorem 3 is based upon phase space analysis and the strategy consists of two steps. First, we study the one-dimensional (DW) when $\mu = 0$. Then we study the perturbation of this problem to positive values of μ, $0 < \mu < \mu_o$. Regarding the first step, existence and multiplicity results of solutions to (DW) are contained in Theorem 5. To describe the structure of the set of solutions to (D), with $\mu = 0$, it is also important to take into account certain distinguished zeros for $f(u)$, which are defined here as *critical* zeros (see Definition 3), together with their associate invariant stable and unstable manifolds (see Definition 5). That kind of zero, when isolated, coincides with the standard saddle points (see [C-H]). As for the second step, a study of the perturbation of those manifolds is required to analyze the existence of solutions to (D) for $0 < \mu < \mu_o$. The precise concept of perturbation needed for the work is introduced in Definition 6. Actually, the properties of perturbation of the unstable manifolds of critical zeros are the required information for the existence of solutions to (D), $\mu > 0$. However, the perturbation of stable manifolds is considered here for the sake of completeness. In fact, a sharp version of a well-known result due to Kannel (see [F]) concerning the perturbability and monotonicity of invariant manifolds is presented (see Proposition 1 and Lemma 3). The main technical results for the proof of Theorem 3 show the relation between the existence of solutions to (D), the perturbability of unstable manifolds, and the formation of monotone chains. Roughly speaking, it is shown that the progressive loss of solutions to (D) when $\mu > 0$ increases (see Propositions 3 and 4) is caused by the monotone chains generated by perturbation of unstable manifolds in the presence of infinitely many zeros for $f(u)$. On the other hand, a characterization of perturbability of unstable manifolds is given in Proposition 2 (Examples 3.1 and 5.1 furnish situations of nonperturbability). The proofs of Theorems 3 and 4 are contained in §6. A preliminary version of these theorems without proofs was presented in [S-F,89].

2. Boundary value conditions for the undulatory solutions.

2.1. The domain of the phase: Structure and smoothness properties. Let us begin with finding out the equation for classical solutions to (1.1) with the form (1.2), i.e., $w(x,t) = u(x - c(t)) = u(y)$. Since it will always be assumed in this work that $b = b(t)$ is continuous and $b(t+T) = b(t)$ for all $t \in \mathbf{R}$ (T-periodicity), then $b(t)$ can be written in the form

$$b(t) = \mu + b_o(t),$$

where $\mu = \frac{1}{T}\int_0^T b(\sigma)\, d\sigma \in \mathbf{R}^n$ is the average of $b(t)$ and the average of $b_o(t)$ is zero. Thus, the waves (1.2) satisfy the equation,

$$\Delta u + (\mu + b_o(t) + c'(t))\nabla u + f(u) = 0 \tag{2.1}$$

or, equivalently,

$$\Delta u + \mu \nabla u + f(u) = 0, \tag{2.2}$$

provided that the choice $c(t) = c_o - \int_o^t b_o(\sigma)\, d\sigma$ has been made. Remark that $c(t)$ is also T-periodic and unique up to an additive constant c_o. Notice also that the existence of the waves (1.2) requires that selection of $c(t)$. In (2.1) and (2.2) derivatives are computed with regard to the phase variable y.

Now, let us study the domain of the phase of the waves. It will be assumed that the spatial variable x belongs to a bounded domain (open and connected) $\Omega \subset \mathbf{R}^n$. We are interested in *permanent* (see [F]) waves, i.e., we want $w(x,t)$ to be defined for all $x \in \Omega$ and $t \in \mathbf{R}$, as is usual in the theory of reaction–diffusion systems. Therefore, the phase $y = x - c(t)$ runs the domain

$$\Omega_\omega = \bigcup_{t \in [0,T]} \Omega_{\omega,t},$$

where $\Omega_{\omega,t} = \Omega - c(t) = \{y \in \mathbf{R}^n / \exists x \in \Omega \text{ with } y = x - c(t)\}$. On the other hand, every regular solution $u = u(y)$ to (2.2), $y \in \Omega_\omega$, defines a wave $w(x,t)$, $x \in \Omega$, $t \in \mathbf{R}$, by means of the identity $w(\cdot,t) = u_{|\Omega_\omega,t}$ ($|$ = restriction). Moreover, remark that the perturbation $u(y_o)$, $y_o \in \Omega_\omega$, propagates with T-periodicity in Ω. In fact, $x = y_o + c(t)$ is the propagation path and $c'(t)$ is the propagation velocity. This shows the undulatory and T-periodic character of the solutions $w(x,t)$ to (1.1) of type (1.2).

Now we are going to prepare the way to endow the waves (1.2), $w(x,t)$, with boundary conditions. The first step consists of finding subregions $\Gamma_t \subset \partial\Omega$ ($\partial\Omega$ the boundary of Ω) so that

$$\partial\Omega_\omega = \bigcup_{t \in [0,T]} \Gamma_t - c(t).$$

This equality must be taken as the definition of the family $\{\Gamma_t\}_{t\in[0,T]}$, which turns out to be necessarily T-periodic in t. Thus the knowledge of $\{\Gamma_t\}_{t\in[0,T]}$ permits us to define certain kinds of boundary conditions on $\Gamma_t \subset \partial\Omega$. The rule is to consider the same type of boundary condition in $\partial\Omega_\omega$ regarding the equation (2.2) (see [Fr-S,84] and [S-F,87] for related ideas).

Therefore, it is important to analyze the structure and regularity of $\partial\Omega_\omega$ and give analytical rules to determine $\{\Gamma_t\}_{t\in[0,T]}$ in $\partial\Omega$. A necessary condition for determining Γ_t, together with some basic properties of Ω_ω, are given in the next result. We recall that a bounded domain $\Omega \subset \mathbf{R}^n$ is said to be C^1 and to be located on one side of its boundary $\partial\Omega$ (see [W]) if there exists a finite open covering $\{U_i\}_{1\le i\le N}$ and a family of C^1 functions $\varphi_i : U_i \to \mathbf{R}$, $\nabla\varphi_i(x) \neq 0$ for all $x \in U_i$, such that $\Omega \cap U_i = \{x \in U_i/\varphi_i(x) < 0\}$, $\partial\Omega \cap U_i = \{x \in U_i/\varphi_i(x) = 0\}$, and $(\mathbf{R}^n\backslash\overline{\Omega}) \cap U_i = \{x \in U_i/\varphi_i(x) > 0\}$ for each $1 \le i \le N$.

LEMMA 1. *Let Ω be a bounded domain of $\mathbf{R}^n$, $b : \mathbf{R} \to \mathbf{R}^n$ a continuous T-periodic function, and $c(t)$ as introduced above. Then the following hold:*

(i) $\Omega_\omega = \cup_{t\in[0,T]} \Omega_{\omega,t}$ *is a bounded domain of* $\mathbf{R}^n$.

(ii) $\overline{\Omega}_\omega = \cup_{t\in[0,T]} \overline{\Omega}_{\omega,t}$ *($\overline{A}$ = the adherence of A).*

(iii) *A certain point $y \in \partial\Omega_\omega$ if and only if $\exists x \in \partial\Omega, \exists t \in [0,T]$ such that $y = x - c(t)$. In other words,*

$$\Gamma_t = \{x \in \partial\Omega / x - c(t) \in \partial\Omega_\omega\}$$

for each $t \in [0,T]$.

(iv) *Assume that Ω is a C^1 domain. With the notation introduced above*

$$\text{(2.3)} \qquad \textit{if a point } x \in \Gamma_t \Rightarrow \exists i \in \{1, \ldots, n\} \textit{ such that } \nabla\varphi_i(x) \cdot c'(t) = 0.$$

Remark 2.1. (a) The proof of points (i)–(iii) is straightforward and that of (iv) is standard and is not given for the sake of briefness.

(b) The structure of Γ_t could undergoe strong variations depending on the geometry of Ω. In this sense, the one-dimensional case exhibits a particular behaviour (see §2.3). The situation is better described by the following example.

Example 2.1. Let $\Omega = B_\epsilon$ be an open ball centered at $(0,0)$ with radius $\epsilon > 0$. Let us take $c(t) = (\cos t, \sin t)$. Then $\Gamma_t = \{(\epsilon\cos t, \epsilon \sin t), (-\epsilon\cos t, -\epsilon\sin t)\}$ for $0 < \epsilon \ \leq 1$ and $t \in [0, 2\pi]$. For ϵ in this range, $\Omega_\omega = B_{1+\epsilon} \backslash B_{1-\epsilon}$. Observe that a piece of $\partial\Omega_\omega$ collapses to $(0,0)$ at $\epsilon = 1$. If $\epsilon > 1$, then $\Gamma_t = \{(-\epsilon\cos t, -\epsilon\sin t)\}$ for $t \in [0, 2\pi]$ and $\Omega_\omega = B_{1+\epsilon}$.

(c) Condition (2.3) is not sufficient, in general, to ensure that an $x \in \partial\Omega$ belongs to Γ_t. For instance, Example 2.1 with $\epsilon > 1$ exhibits points in $\partial\Omega$ satisfying (2.3) and not lying in Γ_t.

In the next result, some conditions on $c(t)$ and Ω are given to show that Γ_t is characterized by (2.3). As a consequence of the proof, the smooth character of the domain Ω_ω is also shown. For the sake of simplicity, we will only consider the case where Ω is a ball in $\mathbf{R}^n$. In this case $\partial\Omega_\omega$ is the so-called tubular surface around the curve $\gamma(t) = -c(t)$ (see [Do]). However, the same study can be developed for bounded smooth enough domains $\Omega \subset \mathbf{R}^n$.

It will be assumed that $\gamma(t) = -c(t)$ defines a *Jordan* curve of class C^n such that the condition

$$\text{(2.4)} \qquad \left\{\frac{d\gamma}{dt}, \ldots, \frac{d^{n-1}\gamma}{dt^{n-1}}\right\} \text{ is linearly independent for every } t \in [0, T]$$

is satisfied. By performing the change $t \to s$, $s(t) = \int_0^t |\frac{d\gamma}{dt}|\, dt$ ($| \cdot |$ the euclidean norm, s the arc length), it is well known that (2.4) entails the existence of a mobil orthonormal reference $\Sigma_s = \{t_1(s), \ldots, t_n(s)\}$ at every point of the curve γ. Moreover, the so-called Frenet Formulae follow from (2.4), i.e.,

$$\frac{dt_i}{ds} = -k_{i-1}(s)t_{i-1}(s) + k_i(s)t_{i+1}(s), \quad 1 \leq i \ \leq n$$

where the functions $k_{-1} \equiv 0$, $k_n \equiv 0$, and the so-called curvature functions $k_1, \ldots, k_{n-1}$ are positive everywhere (see [Kl]).

LEMMA 2. *Assume that $\gamma(t) = -c(t)$ defines a Jordan curve in $\mathbf{R}^n$, $n \geq 2$, which satisfies* (2.4) *and has arc length $\Lambda > 0$. Let Ω be the open ball $B_R(O)$. Provided that the condition*

$$\text{(2.5)} \qquad |\gamma(s) - \gamma(s_o)|^2 \geq \ 2(\gamma(s) - \gamma(s_o)) \cdot \left(\sum_{i=2}^{n} \sigma_i t_i(s_o)\right)$$

holds for every $s, s_o \in [0, \Lambda)$ and $\overline{\sigma} = (\sigma_2, \ldots, \sigma_n) \in \mathbf{R}^{n-1}$ with $|\overline{\sigma}| = 1$, then, for each $R > 0$ such that $R < \inf_{[0,\Lambda]} {k_1}^{-1}(\cdot)$, the set Ω_ω is a C^1 bounded domain whose boundary is given by

$$\text{(2.6)} \qquad \partial\Omega_\omega = \left\{ y \in \mathbf{R}^n / y = \gamma(s) + \sum_{i=2}^{n} \sigma_i t_i(s), s \in [0, \Lambda), |\overline{\sigma}| = R \right\}.$$

Remark 2.2. (a) The Jordan curve hypothesis on $\gamma(t)$ is clearly needed to avoid self-intersections of $\partial\Omega_\omega$. On the other hand, it can be checked that (2.5) holds if $R < \inf_{[0,\Lambda]} {k_1}^{-1}(\cdot)$ is small enough. For a two-dimensional curve $\gamma(t)$, (2.5) means that $\gamma(t)$ does not meet any interior tangent circle with radius R except at the tangent point. For instance, if $\gamma(t) = (a\cos t, b\sin t)$ (an ellipse), (2.5) holds for $0 < R < \frac{b^2}{a}$.

(b) The estimate $0 < \min_{[0,\Lambda]} {k_1}^{-1}(\cdot)$ in Lemma 2 is optimum (recall that $\rho_1(t) = {k_1}^{-1}(t)$ is the curvature radius of $\gamma(t)$ at t, see [Do]). For instance, consider any plane Jordan curve $\gamma(t)$, $t \in [0,T]$, so that $\gamma(t) = A + r(\cos t, \sin t)$ for $0 < t_1 < t < t_2$ and certain t_1, t_2, $t_2 - t_1 = \frac{\pi}{2}$, $r > 0$ and $A = (a_1, a_2) \in \mathbf{R}^2$. Assume in addition that $\rho_1(t) > r$ for $t \notin [t_1, t_2]$ (see Fig. 1). Then for $\Omega = B_R(O)$, $R = r$, the boundary $\partial\Omega_\omega$ of Ω_ω has a *corner* point in A.

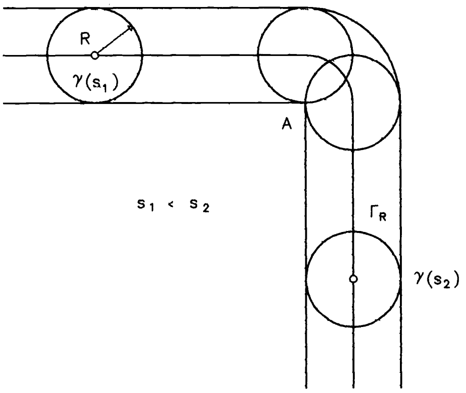


FIG. 1.

(c) We remark that under the hypotheses of Lemma 2, the sets Γ_t are characterized by the condition (2.3).

Proof of Lemma 2. First, let us prove that Ω_ω can be written as

$$\Omega_\omega = \left\{ y \in \mathbf{R}^n / y = \gamma(s) + \sum_{i=2}^{n} \sigma_i t_i(s), s \in [0,\Lambda), |\overline{\sigma}| < R \right\},$$

where $\overline{\sigma} = (\sigma_2, \dots, \sigma_n) \in \mathbf{R}^{n-1}$. In fact, it is clear that the right-hand side of the equality is contained in Ω_ω. On the other hand, if $y \in \Omega_\omega$ then $y \in B_R(\gamma(s_o))$ for some $s_o \in [0,\Lambda)$. Thus, the distance $d(y,\gamma)$ from y to the trace γ of $\gamma(s)$ satisfies $d(y,\gamma) < R$. The periodicity of $\gamma(s)$ implies that $\exists s^* \in (-\epsilon, \Lambda + \epsilon)$ such that $d(y,\gamma) = |y - \gamma(s^*)|$. Hence,

$$\frac{d}{ds}|y - \gamma(s)|^2_{s=s^*} = 0. \tag{2.7}$$

Therefore, (2.7) implies that $y - \gamma(s^*) = \sum_{i=2}^{n} \sigma_i t_i(s^*)$ for certain $\overline{\sigma} = (\sigma_2, \dots, \sigma_n)$, $|\overline{\sigma}| < R$.

Obtaining the desired expression for $\partial\Omega_\omega$ is a little more complicated. Observe that $\partial\Omega_\omega = \{y \in \mathbf{R}^2 / y = (-\cos s - \epsilon\cos s, -\sin s - \epsilon \sin s)\}$ in Example 2.1, $\epsilon > 1$. Thus $\partial\Omega_\omega$ does not agree with the expression (2.6).

If $y \in \partial\Omega_\omega$, point (ii) in Lemma 1 and the same argument employed above give $d(y,\gamma) = R$. As in (2.7), $\exists s^* \in [0,\Lambda), \overline{\sigma} \in \mathbf{R}^{n-1}$ with $|\overline{\sigma}| = R$ such that $y = \gamma(s^*) + \sum_{i=2}^{n} \sigma_i t_i(s^*)$. On the other hand, $y \in \overline{\Omega_\omega}$ provided that $y = \gamma(s^*) + \sum_{i=2}^{n} \sigma_i t_i(s^*)$

for some $s^* \in [0, \Lambda)$ and $\overline{\sigma} \in \mathbf{R}^{n-1}$, $|\overline{\sigma}| = R$. If such $y \in \Omega_\omega$ then $\exists s_o \neq s^*$ such that $y = \gamma(s_o) + \sum_{i=2}^n \sigma_i^o t_i(s_o)$ for certain $\overline{\sigma}_o = (\sigma_i^o)$, $|\overline{\sigma}_o| < R$. This implies

$$|\gamma(s^*) - \gamma(s_o)|^2 < 2(\gamma(s^*) - \gamma(s_o)) \cdot \left(\sum_{i=2}^n \sigma_i t_i(s_o) \right),$$

where $\overline{\sigma} \in \mathbf{R}^{n-1}$, $|\overline{\sigma}| = R$. This contradicts (2.5).

Let us now prove the smoothness of $\partial\Omega_\omega$. By the compacity of Ω_ω it is sufficient to find a neighbourhood $U(y_o)$ of every $y_o \in \partial\Omega_\omega$ and a C^1 function $\varphi : U(y_o) \to \mathbf{R}$ with the properties introduced just before the statement of Lemma 1. To accomplish this, for a fixed $y_o \in \partial\Omega_\omega$ let us write

$$y_o = \gamma(s_o) + \sum_{n=2}^n \sigma_i^o t_i(s_o),$$

with $\overline{\sigma}_o = (\sigma_i^o) \in B_R^{n-1}$ (B_R^{n-1} the ball with radius $R > 0$ and center O in $\mathbf{R}^{n-1}$). Define the map

$$\begin{aligned} T : (-\epsilon, \Lambda) \times B_{R-\epsilon}^{n-1} &\longrightarrow \quad \mathbf{R}^n, \\ (s, \sigma) \quad &\longrightarrow \quad y = T(s, \sigma) = \gamma(s) + \textstyle\sum_{i=2}^n \sigma_i t_i(s). \end{aligned}$$

T is locally invertible at (s_o, σ_o). In fact,

$$dT(s_o, \sigma_o) = col \left[\frac{\partial T}{\partial s}, \frac{\partial T}{\partial \sigma_2}, \dots, \frac{\partial T}{\partial \sigma_n} \right].$$

Since,

$$\begin{aligned} \frac{\partial T}{\partial s} &= (1 - \sigma_2 k_1) t_1 + \sum_{i=3}^n (\sigma_{i-1} k_{i-1} - \sigma_{i+1} k_i) t_i + \sigma_{n-1} k_{n-1} t_n, \\ \frac{\partial T}{\partial \sigma_i} &= t_i, \quad 2 \leq i \leq n, \end{aligned}$$

then

$$\left| \det col \left[\frac{\partial T}{\partial s}, \frac{\partial T}{\partial \sigma_2}, \dots, \frac{\partial T}{\partial \sigma_n} \right] \right| = (1 - \sigma_2^o k_1(s_o)) \geq \ 1 - R k_1(s_o) > 0.$$

By using the inverse function theorem, $\exists \delta > 0$ such that, if we set $V(s_o, \sigma_o) = (s_o - \delta, s_o + \delta) \times \prod_{i=2}^n (\sigma_i^o - \delta, \sigma_i^o + \delta)$, then $U(y_o) = T(V(s_o, \sigma_o))$ is a neighbourhood of y_o and $G = T^{-1}$ is a well-defined C^1 function. By writing $G(y) = (s, \sigma)$ with $s = G_1(y)$ and $\sigma = G_2(y)$, it is sufficient to define $\varphi(y) = |G_2(y)|^2 - R^2$. This completes the proof of Lemma 2. □

2.2. Boundary value conditions for the undulatory solutions. Let us now define the homogeneous Dirichlet and Neumann problems for the undulatory solutions (1.2), $w(x,t) = u(x - c(t))$, to (1.1).

A wave (1.2), $w(x,t) = u(x - c(t))$, is said to satisfy the homogeneous Dirichlet condition in the domain $\Omega \subset \mathbf{R}^n$ provided that

$$w(x,t) = 0, \quad \forall t \in \mathbf{R}, \forall x \in \Gamma_t.$$

A wave (1.2), $w(x,t) = u(x-c(t))$, satisfies the homogeneous Neumann condition in $\Omega \subset \mathbf{R}^n$ if the relation

$$\frac{\partial w}{\partial \nu}(x,t) = 0, \quad \forall t \in \mathbf{R}, \forall x \in \Gamma_t$$

holds ($\nu = \nu(x)$ stands for the exterior unit normal to $\partial\Omega$ at $x \in \partial\Omega$).

Thus, the homogeneous Dirichlet problem for the waves (1.2) is defined as

$$\text{(DW)} \qquad \begin{cases} \frac{\partial w}{\partial t} = \Delta w + b(t)\nabla w + f(w), & (x,t) \in \Omega \times \mathbf{R}, \\ w(x,t) = u(x-c(t)), & \\ w(x,t) = 0, & t \in \mathbf{R}, x \in \Gamma_t. \end{cases}$$

Similarly, the homogeneous Neumann problem for the waves (1.2) consists in solving,

$$\text{(NW)} \qquad \begin{cases} \frac{\partial w}{\partial t} = \Delta w + b(t)\nabla w + f(w), & (x,t) \in \Omega \times \mathbf{R}, \\ w(x,t) = u(x-c(t)), & \\ \frac{\partial w}{\partial \nu}(x,t) = 0, & t \in \mathbf{R}, x \in \Gamma_t. \end{cases}$$

In the present work we will restrict ourselves to analyzing positive and classical solutions to the Dirichlet problem (DW). According to §2.1 it will always be assumed that the necessary choice $c'(t) = -b_o(t)$ has been made. Therefore, it will always be assumed in the work that a wave (1.2), $w(x,t) = u(x-c(t)) = u(y)$, is a solution to (DW) if and only if $u(y)$ solves

$$\text{(D)} \qquad \begin{cases} \Delta u + \mu\nabla u + f(u) = 0, & y \in \Omega_\omega, \\ u = 0, & y \in \partial\Omega_\omega. \end{cases}$$

We should remark that a similar study could have been developed for the Neumann problem or even for another kind of homogeneous boundary condition. On the other hand, it is also possible to define and study the notion of *weak* solution (in the Sobolev spaces sense) to (DW) and (NW). However, the consideration of these facts is beyond the scope of the present work.

The main results of this paper concern the one-dimensional version of (DW). For this reason, the special features exhibited by this problem deserve a separate section.

2.3. The one-dimensional Dirichlet problem. In the one-dimensional case Ω is an open interval, $\Omega = (-a,a)$, $a > 0$. However, in general, the sets Γ_t are either empty or are $\{a\}$ or $\{-a\}$ for finitely many $t \in [0,T]$. More specifically, the constant c_o in §2.1 will always be chosen so that $M = \max_{[0,T]} -c(t) = -\min_{[0,T]} -c(t)$, $M > 0$. Observe that this choice amounts to a phase translation in $u(y)$ and the equation for $u(y)$ is invariant under phase translations. After this normalization, $\Omega_\omega = (-a-M, a+M)$ and the boundary sets are $\Gamma_t = \{-a\}$ for $t \in M_1$ and $\Gamma_t = \{a\}$ for $t \in M_2$, where $M_1 = \{t \in \mathbf{R}/c(t) = M\}$ and $M_2 = \{t \in \mathbf{R}/c(t) = -M\}$. Thus, the one-dimensional (DW) can be written as

$$\text{(DW)} \qquad \begin{cases} w_t = w_{xx} + b(t)w_x + f(w), & (x,t) \in (-a,a) \times \mathbf{R}, \\ w(x,t) = u(x-c(t)), & \\ \begin{cases} w(-a,t) = 0, & t \in M_1, \\ w(a,t) = 0, & t \in M_2. \end{cases} \end{cases}$$

The classical positive solutions to (DW) are precisely the classical solutions to the problem

$$\text{(D)}_1 \qquad \begin{cases} u_{yy} + \mu u_y + f(u) = 0, & y \in (-a,a), \\ u(\pm(a+M)) = 0, & \end{cases}$$

where $y = x - c(t)$ designates the phase. To obtain a geometrical description of the behaviour of a solution to (DW), assume that the profile of a positive solution $u(y)$ to (D) is as in Fig. 2a, being $c(t)$ as in Fig. 2b.

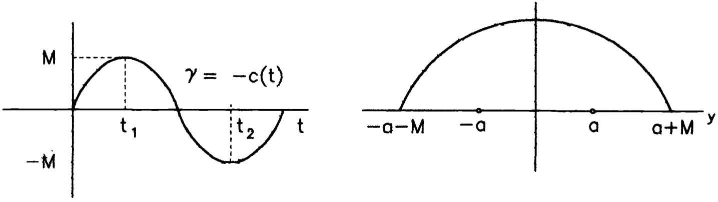
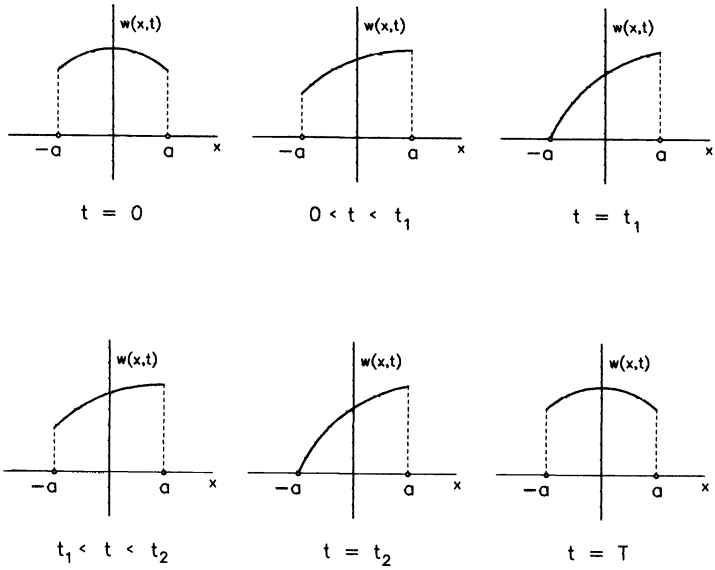


FIG. 2.

This implies that $M_i = \{t_i + kT/k \in \mathbf{Z}\}$, $i = 1, 2$. The behaviour of $w(\cdot, t)$ is given by the restriction $u(\cdot)|_{I(t)}$, where $I(t) = (-c(t) - a, c(t) + a)$. A sketch of the wave along a period $t \in [0, T]$ is given in Fig. 3.



$t = 0$ $0 < t < t_1$ $t = t_1$

$t_1 < t < t_2$ $t = t_2$ $t = T$

FIG. 3.

Remark 2.3. In the study of (DW), i.e., of $(D)_1$, it is sufficient to consider μ nonnegative. In fact, a positive $u(y)$ solves $(D)_1$ in $\Omega = (-a, a)$ with $\mu = \mu_1 < 0$, provided that $\tilde{u}(y) = u(-y)$ solves $(D)_1$ in Ω with $\mu = -\mu_1 > 0$.

3. Statement of the main results. First let us introduce the hypotheses required for the nonlinearity $f(u)$. It will always be assumed that $f : \mathbf{R} \to \mathbf{R}$ is a C^1 function that satisfies

$$(H_f) \qquad \begin{cases} \text{(i) } f(0) = 0, \\ \text{(ii) } \exists u_o > 0 \text{ such that } f(u) < 0 \text{ for } u \geq u_o, \\ \text{(iii) } \exists u_1 > 0 \text{ such that} \int_0^{u_1} f(s)\, ds > 0. \end{cases}$$

As is usual in the literature, the following functions associated with $f(u)$ will be used in the work:

$$V(u) = \int_0^u f(s)\, ds, \quad E(u, v) = \frac{1}{2}v^2 + V(u). \tag{3.1}$$

Let us also introduce two parameters associated to $f(u)$ for later reference. They are

$$V_M = \max_{[0,+\infty)} V(u) \quad \text{and} \quad u_M = \min\{u \in \mathbf{R}^+/V(u) = V_M\}.$$

Observe that $u = u_M$ is a zero of $f(u)$.

Remark 3.1. (a) The hypotheses (H_f) on $f(u)$ are usual in the literature of positive solutions to semilinear elliptic and parabolic equations. See for instance [Am], [Be-L,83], [L], [R], [Sm-W,86], [Sm-W,87]. As it will be seen later, the class of nonlinearities $f(u)$ satisfying (H_f) considerably enlarges the kind of nonlinear terms $f(u)$ considered in the theory of travelling waves to semilinear parabolic equations, as those in [Ar-W,75], [Ar-W,78], [F-M,77], [F-M,81], [F]. For later use it is worthwhile to mention two outstanding examples in that class, i.e., the *logistic* (or *Fisher's*) nonlinearity

$$f_1(u) = u(1-u),$$

and the *bistable* nonlinearity

$$f_2(u) = -u(\alpha - u)(1-u), \quad \alpha \in \left(0, \frac{1}{2}\right),$$

which have also been deeply studied in the dynamical theory of semilinear parabolic equations (see [He], [Sm]). However, in the literature of travelling waves quoted above, the discrete character of the set of zeros of $f(u)$ plays an important role. It should be remarked that (H_f) allows $f(u)$ to exhibit infinitely many zeros in the results of travelling waves to the equation

$$\frac{\partial u}{\partial t} = u_{xx} + f(u) \tag{1.3}$$

contained in this paper (see Theorem 3).

(b) The condition (i) is usual in the literature. With regard to (iii), observe that it is a necessary condition for the existence of positive solutions to the problem (D) with $\mu = 0$ and Ω_ω starshaped relative to some point. This is a consequence of the Pohozaev's identity [P]. In the one-dimensional case of (D), (iii) is still necessary even when $\mu \neq 0$.

(c) The condition (ii) is responsible for an interesting phenomenon, namely, the fact that no positive solutions to (DW) exist in *any* bounded domain Ω, provided that the average μ is larger in *modulus* than a certain critical value μ_o. We study this phenomenon extensively in this paper. It is possible, however, to have positive solutions if $|\mu|$ is small enough and Ω is conveniently large. To state these facts precisely in the next result, it is suitable to introduce the following value, associated to a bounded domain $\Omega \subset \mathbf{R}^n$:

$$\mu_o(\Omega) = \sup\left\{ r > 0 / \begin{array}{l} \text{There exists a positive solution to (DW)} \\ \text{in } \Omega \text{ for some } \mu \in \mathbf{R}^n;\ 0 \leq |\mu| \leq r \end{array} \right\}.$$

THEOREM 1. *Let $b : \mathbf{R} \to \mathbf{R}^n$ be a continuous and T-periodic function with average μ. Let us write $b(t) = \mu + b_o(t)$ and choose $c(t)$ such that $c'(t) = -b_o(t)$. Let $f(u)$ be a C^1 function satisfying* (H_f) *and assume that Ω is a bounded domain of $\mathbf{R}^n$ so that Ω_ω is a smooth enough domain. Then,*

(a) *The problem* (DW) *admits nonnegative classical solutions $w(x,t) = u(x-c(t))$ which satisfy the estimate*

$$\|w(\cdot,t)\|_{\infty,\overline{\Omega}} < u_M, \quad \forall t \in [0,T].$$

(b) *$\exists\mu_1 > 0$, μ_1 not depending on Ω, such that $\mu_o(\Omega) < \mu_1$ for every bounded domain $\Omega \subset \mathbf{R}^n$. In other words,* (DW) *does not admit positive solutions in Ω provided that $|\mu| \geq \mu_1$. Moreover,*

$$\mu_1 = \sqrt{4\left(f'(0) + \sup_{[0,u_M]} \frac{g(u)}{u}\right)},$$

where $g(u) = f(u) - f'(0)u$.

(c) *Let us keep $b_o(t)$ fixed in the expression for $b(t)$ and consider $\mu \in \mathbf{R}^n$ as a parameter. For a fixed point $x_o \in \mathbf{R}^n$ consider also the family of domains $\Omega_\lambda = x_o + \lambda(\Omega - \{x_o\})$, where $\Omega - \{x_o\} = \{x - x_o / x \in \Omega\}$ and $\lambda \in \mathbf{R}^+$. Then there exists $\overline{\lambda} > 0$ such that*

$$0 < \mu_o(\Omega_\lambda) \quad \textit{for each} \quad \lambda \geq \overline{\lambda}.$$

Moreover, for every $\lambda \geq \overline{\lambda}$, there exists $r(\lambda) > 0$ such that (DW) *admits a positive solution in Ω_λ for each $\mu \in \mathbf{R}^n$, $0 \leq |\mu| \leq r(\lambda) \leq \mu_o(\Omega_\lambda)$, and $\lambda \geq \overline{\lambda}$.*

Remark 3.2. (a) Parts (b) and (c) of Theorem 1 motivate the introduction of an important parameter for the problem (DW). In Definition 1, $b_o(t)$ is kept fixed while $\mu \in \mathbf{R}^n$ is observed as a parameter. Notice that the domain Ω_ω does not depend on μ.

DEFINITION 1. *The largest existence value μ_o for positive solutions to* (DW) *is defined as*

$$\mu_o = \sup \mu_o(\Omega), \tag{3.2}$$

where the supremum is taken over all bounded domains $\Omega \in \mathbf{R}^n$.

The condition (ii) is crucial for part (b). For instance, the problem (DW) with $f(u) = u^m$, $m \geq 1$, admits a positive solution $w(x,t)$ in a certain domain $\Omega = (-a(\mu), a(\mu))$ for each positive μ, i.e., $\mu_o = +\infty$. On the other hand, Theorem 1 holds under slight variations of (H$_f$). For instance, keep (i) and (iii) in (H$_f$), and replace (ii) by,

$$\text{(ii)}' \quad \exists u_n > u_1 > 0 \quad \text{such that} \quad f(u_n) < 0.$$

Then the conclusions of Theorem 1 hold for the solutions to (DW) which satisfy in addition $0 < w(x,t) < u_n$, $x \in \Omega$, $t \in \mathbf{R}$.

(b) μ_1 is positive. Indeed, $\varphi(u) = \frac{g(u)}{u}$ for $u \neq 0$ and $\varphi(0) = 0$ is continuous in $u \geq 0$. Since $\varphi(u_M) + f'(0) = 0$ then $\mu_1 \geq 0$, but (iii) in (H$_f$) implies the existence of $u_p \in [0, u_M]$ with $\varphi(u_p) > 0$. Thus $\mu_1 > 0$. Observe that part (b) in Theorem 1 implies that $\mu_o \leq \mu_1$, while part (c) asserts that $\mu_o > 0$. It is possible to show the equality $\mu_o = \mu_1$ in the one-dimensional (DW) with $f(u) = f_1(u)$.

In the next two results a geometrical description of the constant μ_o in (3.2) is given. The following concept is key: Assume that $f(u)$ in equation (1.3) satisfies $f(0) = f(1) = 0$. A *monotone travelling wave* (MTW) to (1.3), which *connects* the zeros $u = 1$ and $u = 0$ of $f(u)$ with *propagation velocity* $c > 0$, is a solution $w(x,t) = u(x - ct) = u(y)$ to (1.3) which satisfies

(a) $0 < u(y) < 1$, $u'(y) < 0$, $\forall y \in \mathbf{R}$,

(b) $\lim_{y \to +\infty} u(y) = 0$; $\lim_{y \to -\infty} u(y) = 1$.

For $f = f_1(u)$ a well-known result due to Kolmogoroff et al. (see [F]) ensures the existence of MTWs with propagation velocities $c \geq c_1^* > 0$ (c_1^* is the minimum propagation velocity). As for $f = f_2(u)$, the existence of MTWs with a unique propagation velocity $c_2^* > 0$ is well known (see [F-M,77]). A first result, identifying μ_o with c_1^* and c_2^* is stated now.

THEOREM 2. *Let $b : \mathbf{R} \to \mathbf{R}$, $b(t) = \mu + b_o(t)$, be a continuous and T-periodic function with average μ. Define $M = \max_{[0,T]} \int_0^t b_o(s)\,ds$.*

(A) *The problem* (DW) *with the nonlinearity $f = f_2(u)$ has the following properties:*

(a) *The constant μ_o in* (3.2) *satisfies $\mu_o = c_2^*$. For $|\mu| \geq \mu_o$ the problem* (DW) *does not admit positive solutions in any domain $\Omega = (-a, a)$.*

(b) *Let us assume that $|\mu| < \mu_o$. Then, there exists $T_o > 0$ and $a(\mu) > 0$ for each $|\mu| < \mu_o$ such that*

(i) *for $a \geq a(\mu)$ there exists at least a positive solution $w(x,t) = u(x - c(t))$ to* (DW) *in $\Omega = (-a, a)$. Moreover,*

$$\|w(\cdot,\cdot)\|_{\infty,\Omega\times\mathbf{R}} > \alpha.$$

(ii) *If $M < T_o$ then $\inf_{[0,\mu_o]} a(\mu) > 0$. Furthermore,* (DW) *does not admit positive solutions in $\Omega = (-a, a)$ provided that $0 < a < a(\mu)$.*

(B) *The problem* (DW) *with $f = f_1(u)$ has the following properties:*

(a) *The constant $\mu_o = c_1^*$. For $|\mu| \geq \mu_o$* (DW) *does not admit positive solutions in any domain $\Omega = (-a, a)$.*

(b) *Assuming that $|\mu| < \mu_o$ there exist $T_o > 0$ and $a = a(\mu) > 0$ such that*

(i) *for $a \geq a(\mu)$ there exists a unique positive solution $w(x,t) = u(x - c(t))$ to* (DW) *in $\Omega = (-a, a)$. Moreover,*

$$\lim_{a \to a(\mu)+} \|w(\cdot,\cdot)\|_{\infty,\Omega\times\mathbf{R}} = 0.$$

(ii) *if $M < T_o$ then $\inf_{[0,\mu_o]} a(\mu) > 0$. Furthermore,* (DW) *does not admit positive solutions in $\Omega = (-a, a)$ for $0 < a < a(\mu)$.*

Remark 3.3. (a) In parts (A) and (B) two different kinds of bifurcation phenomena are shown regarding the domain $(-a, a)$. In the case of $f_2(u)$, the positive waves appear spontaneously when a crosses $a(\mu)$. As for $f_1(u)$ they bifurcate from 0 at the critical value $a(\mu)$ of a.

(b) The constant μ_o has been identified in Theorem 2 as a propagation velocity in the problem (D), provided the nonlinearities are $f(u) = f_i(u)$, $i = 1, 2$. The proof is based upon phase space analysis, in which the formation of MTWs to (1.3) in the range $0 < u < 1$ plays an important role (see §6). However, when $f(u)$ is so general as in (H$_f$) the possible existence of infinitely many zeros of $f(u)$ in the interval $(0, u_M)$ could cause the appearance of complex phenomena in the structure of the MTWs to (1.3). For instance, infinitely many MTWs, connecting infinitely many zeros of $f(u)$, with the same propagation velocity c, could be exhibited by the equation (1.3) (see Example 3.1 below and Example 5.1 in §5). Therefore, to identify the constant μ_o for general $f(u)$, a convenient extension of the concept of MTW is required. That is the objective of the next definition.

DEFINITION 2. *Let*

$$\left\{ \begin{aligned} \frac{du}{dy} &= v, \\ \frac{dv}{dy} &= -cv - f(u) \end{aligned} \right. \tag{3.3}$$

be the equation of the solutions $w(x,t) = u(x - ct) = u(y)$ of (1.3), *$x, t \in \mathbf{R}$, $c \in \mathbf{R}$, and let $0 \leq u_1 < u_2 \leq u_M$ be two zeros of $f(u)$. A* chain of monotone travelling waves (CW) *to* (1.3) *which connects u_1 to u_2 with propagation velocity c is defined as a set $C(c) \subset \mathbf{R}^2$ with the following properties:*

(i) $C(c) \subset [0,+\infty) \times (-\infty, 0]$ *and* $(u_i, 0) \in C(c)$, $i = 1, 2$.

(ii) $C(c)$ *is an invariant set regarding* (3.3) *with the form*

$$C(c) = \{(u,v)/v = g(u), a \leq u \leq b\}$$

for some numbers $0 \leq a < b$ *and some continuous real function* $g(u)$ *defined in the interval* $[a, b]$.

Remark 3.4. Notice that Definition 2 is given in terms of orbits rather than solutions of (3.3). Obviously, the orbit of a MTW to (1.3) together with its α- and ω-limit sets give the simpler example of a CW. On the other hand, it is implicit in the definition that a CW consists of the union of the orbits of (possibly) infinitely many MTWs, together with their limit sets. Observe that the only possible limit set of such an MTW into a CW is $\{(z, 0)\}$ with $f(z) = 0$. Finally, standard ordinary differential equation (ODE) arguments show that $u = u_-$ and $u = u_+$ defined as

$$u_-(\text{resp., } u_+) = \min \ (\text{resp., } \max) \ C(c) \cap ([0, +\infty) \times \{0\})$$

are the *first* and the *last* zeros of $f(u)$ connected by $C(c)$ with propagation velocity c. A more complex example of a CW follows.

Example 3.1. Consider $\varphi \in C_o^\infty(\mathbf{R})$, $0 \leq \varphi(x) \leq 1$, supp $\varphi = [-\frac{1}{2}, \frac{1}{2}]$ such that $\varphi(x) = 0$ for $|x| \geq \frac{1}{2}$ and $\varphi(x) = 1$ for $|x| \leq \frac{1}{4}$. Define $f_o(u) = u(1-u)\varphi(u - \frac{1}{2})$.

By taking $f(u) = f_o(u)$ in the equation (3.3), it is known (cf. [F]) that there exists $c_o^* \geq 2\sqrt{\sup_{[0,1]} f(u)/u} \geq 0$ and an MTW connecting $(u, v) = (1, 0)$ to $(0, 0)$ with velocity c, for each $c \geq c_o^*$. On the other hand, we claim that every MTW to equation (3.3) with $f = f_o$, which connects $(1, 0)$ to $(0, 0)$ with velocity c, generates an MTW to (3.3) with $f = \epsilon^2 f_o$ exhibiting the same properties but having velocity ϵc. In fact, observe that $u(y)$ solves

$$u''(y) + cu'(y) + f_o(u(y)) = 0, \qquad \left(' = \frac{d}{dy}\right),$$

provided that $U(z) = u(\epsilon z)$ solves

$$U''(z) + c^* U'(z) + \epsilon^2 f_o(U(z)) = 0, \qquad \left(' = \frac{d}{dz}\right),$$

with $c^* = \epsilon c$.

By using this remark, it is possible to build up nonlinearities $f(u)$ that exhibit interesting CW connections. In fact, let $\{\epsilon_n\}_{n \geq 1}$ be a positive decreasing sequence such that $\sum_{n=1}^\infty \epsilon_n < +\infty$. Let us choose $\epsilon_1 = 1$ and define $\{u_n\}_{n \geq 0}$ as follows: $u_o = 0$, $u_n = u_{n-1} + \epsilon_n$, $n \in \mathbf{N}$. Define $u_\infty = \sum_{n=1}^\infty \epsilon_n$. Observe that $\lim u_n = u_\infty$.

Let us define

$$f_n(u) = f_o\left(\frac{u - u_{n-1}}{u_n - u_{n-1}}\right), \quad n \in \mathbf{N}.$$

Observe that supp $f_n = [u_{n-1}, u_n]$. Finally, let us introduce

$$f(u) = \sum_{n=1}^\infty \epsilon_n^2 f_n(u).$$

First, observe that $f \in C_o^1(\mathbf{R})$. Its first zero with maximum energy is $u_M = u_\infty$ (see (3.1)). On the other hand, for each pair of points $(u_n, 0)$ and $(u_{n-1}, 0)$ the equation (3.3) with $f(u)$ exhibits MTWs connecting them with propagation velocities c, for each $c \geq \mu_n = \epsilon_n c_o^*$. On the other hand $\lim \epsilon_n = 0$. Therefore, for each fixed $c > 0$ there exists a CW, $C(c)$, connecting $u_+(c) = u_\infty$ to $u_-(c) = \inf\{u_m/c \geq \epsilon_m c_o^*;\ m \geq\ 1\}$. Notice that $C(c)$ also contains the orbits of all those MTWs quoted above, which exist provided $c \geq \epsilon_n c_o^*$. Finally, the zero $u_-(c)$ is nonincreasing regarding c. For $c \geq \max\{\epsilon_n c_o^*\}$ there exists a CW, $C(c)$, connecting $u_+(c) = u_\infty$ to $u_-(c) = 0$ and all the zeros of $f(u)$ in the interval $[0, u_\infty]$ (see Fig. 4).

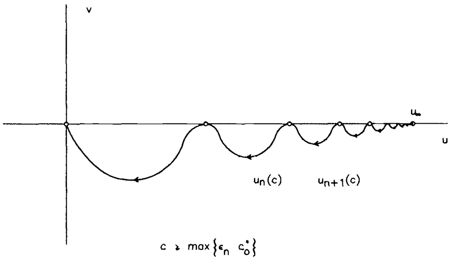


FIG. 4.

Another interesting example of CW is given in §5 (see Example 5.1).

With the concept of a CW in mind, it is already possible to furnish an interpretation of the constant μ_o associated to (DW) for general nonlinearities. In this sense, Theorem 2 is sharpened in the next result.

THEOREM 3. *Assume that $f(u)$ is a C^1 function satisfying* (H_f). *Consider the largest existence value μ_o associated to the problem* (DW). *Then the following alternatives hold:*

(i) *Either equation* (1.3) *exhibits a CW, $C(c)$, which connects a zero $u = u_+$ of $f(u)$, $u_+ \in\ [0, u_M]$, to the zero $u_- = 0$ with propagation velocity $c = \mu_o$, or*

(ii) *equation* (1.3) *exhibits a sequence $\{C_n(c_n)\}$ of CWs such that* (a) *$C_n(c_n)$ connects a pair of zeros of $f(u)$, $u_- = u_n^- < u_+ = u_n^+$; $u_n^-, u_n^+ \in\ [0, u_M]$,* (b) *the sequences $\{u_n^-\}$ and $\{u_n^+\}$ are decreasing and* $\lim u_n^- = 0$, *and* (c) *the propagation velocities sequence $\{c_n\}$ is increasing and* $\lim c_n = \mu_o$.

Some additional uniqueness and monotonicity properties of the CWs are given in the next result. The following convention will be used. It is said that two CWs to (1.3) satisfy $C \leq C'$ if $\forall (u, v') \in C'$ such that $\exists v \leq 0$ with $(u, v) \in C$ then $v \leq v'$. If $C \leq C'$ and $v < v'$ for some pair $(u, v) \in C$, $(u, v') \in C'$ we will put $C < C'$.

THEOREM 4. *Assume that the function $f(u)$ in equation* (3.3) *is of class C^1. Then,*

(i) *if $u_1 < u_2$ are two zeros of $f(u)$ in $[0, \infty)$ then, for every $c \geq 0$, there exists at most a unique CW, $C(c)$, connecting u_1 to u_2;*

(ii) *if $C(c)$ and $C'(c')$ are two CWs to equation* (1.3) *with velocities $c < c'$, then $C(c) < C'(c')$ provided that their last zeros satisfy $u_+(c') \leq\ u_+(c)$.*

Finally, let us introduce some results describing qualitative properties about the one-dimensional (DW) with zero average $\mu = 0$. The results are, to some extent, a partial continuation to general $f(u)$ of those contained in Theorem 2. However, the study of the same properties for μ positive is beyond the objectives of this work.

To state the next result, it is convenient to mention a few facts concerning the problem (DW) with $\mu = 0$. In this case, the equation in $(D)_1$,

$$u''(y) + f(u(y)) = 0,$$

is conservative. In fact, every solution $w(x, t) = u(x - c(t)) = u(y)$ to (DW) with $\mu = 0$ makes constant the energy $E(u, v)$ (see (3.1)), i.e., $E(u(\cdot), u'(\cdot)) \equiv \beta$. However, not all values of β are energy values of solutions to (DW). For instance, such values β

must belong to $[0, V_M)$ (see §4). On the other hand, a number $\alpha \in [0, V_M)$ could not be the energy of any solution to (DW). These are just the values introduced in the next definition.

DEFINITION 3. *A number $\alpha \in [0, V_M)$ is said to be a* critical value *for the energy $V(u) = \int_0^u f(s)\, ds$ if the number,*

$$u_\alpha = \min\{u/u > 0, V(u) = \alpha\}$$

is a zero of $f(u)$, i.e., $f(u_\alpha) = 0$. We will also say that u_α is the critical zero *associated to α.*

The set of critical values associated to certain f will be designated by Φ. Thus, the information about the solvability of (DW) is contained in $[0, V_M) - \Phi$. The structure of $[0, V_M) - \Phi$ and other features about (DW) are discussed in the next result.

THEOREM 5. *Assume that $f = f(u)$ satisfies* (H_f) *and $b : \mathbf{R} \to \mathbf{R}$ is continuous, T-periodic, and has average $\mu = 0$. Then,*

(i) *if $\alpha \in [0, V_M)$ is a critical value of $V(u)$ then there exist no solutions to* (DW) *in any domain $\Omega = (-a, a)$ satisfying $E \equiv \alpha$.*

(ii) *let $\Phi \subset [0, V_M)$ be the set of critical values of $V(u)$. Then $[0, V_M) - \Phi$ can be written as the union of a countable family of disjoint open intervals $I_n = (\alpha_n^-, \alpha_n^+)$, i.e.,*

$$[0, V_M) - \Phi = \bigcup_{n \in \mathbf{N}} I_n.$$

(iii) *assume that there are infinitely many intervals in statement (ii) and set $M = \max_{[0,T]} \int_0^u b(s)\, ds$. Then, there exist $T_o > 0$ and a sequence $\{a_n\}$, $a_n \geq 0$, such that*

(a) *the problem* (DW) *admits a positive solution in a domain $\Omega = (-a, a)$, $a > 0$, provided that there exists $n \in \mathbf{N}$ such that $a \geq a_n$.*

(b) *if $M < T_o$ then $a_n > 0$ for every $n \in \mathbf{N}$, and the number $a_o \doteq \inf\{a_n\}$ is positive. Moreover,* (DW) *does not admit positive solutions in a domain $\Omega = (-a, a)$ when $0 < a < a_o$.*

(c) $\forall N \in \mathbf{N}, \exists a(N) > 0$ *such that* (DW) *admits at least N positive solutions in a domain $\Omega = (-a, a)$ for every $a > a(N)$.*

4. The n-dimensional case: The proofs of Theorems 1 and 5.

Proof of Theorem 1. Let us begin with part (a). First, recall the equivalence between (DW) and (D). From part (ii) of $\overline{(\mathrm{H})_f}$, a direct computation implies that every positive solution $u \in C^2(\Omega_\omega) \cap C^o(\overline{\Omega_\omega})$ to (D) satisfies the estimate

$$0 \leq u(y) \leq u_o \quad \forall y \in \Omega_\omega. \tag{4.1}$$

Since $\varphi = 0$ and $\psi = u_o$ are, respectively, lower and upper solutions to (D), the existence of two nonnegative weak solutions $0 \leq \overline{u}(y) \leq \tilde{u}(y) \leq u_o$, $y \in \Omega_\omega$, follows from Theorem 9.4 in [Am]. Moreover, $\overline{u}$ and $\tilde{u}$ are, respectively, the minimal and the maximal nonnegative solutions to (D). From that inequality, the smoothness of the nonnegative solutions up to the boundary of Ω_ω follows from a well-known bootstrap argument (see [Am], [R]) provided $\partial\Omega_\omega$ is of class $C^{2+\alpha}$, $\alpha \in [0, 1)$. Since it could happen that $\overline{u}(y) = \tilde{u}(y) = 0$, $\forall y \in \Omega_\omega$, the existence of positive solutions will be studied later.

To prove the estimate by $u = u_M$, the first zero of $f(u)$ maximizing the energy

$V(u)$, consider first the one-dimensional case of (D)

$$(D)_1 \qquad \begin{cases} u_{yy} + \mu u_y + f(u) = 0, & y \in (-b,b), \\ u(\pm b) = 0. \end{cases}$$

If $u(y)$ is a positive solution to (D)$_1$ then $(u(y),v(y)) = (u(y),u_y(y))$ solves (3.3) and (i) $u(y) \geq 0, |y| \leq b$, (ii) $(u(\pm b),v(\pm b)) = (0,v_{\mp})$, $v_- < 0$, $v_+ > 0$. Assume, by contradiction, that $u_m \doteq \max_{|y|\leq b} u(y) \geq u_M$. Then $(u(y_o),v(y_o)) = (u_m,0)$ for a certain $y_o \in (-b,b)$. Since $(u_M,0)$ is a critical point to (3.3) then, by uniqueness, $u_M < u_m$. In addition, since $\frac{dE}{dy} = -\mu v^2(y)$, the function $E(y) = E(u(y),v(y))$ (see (3.1)) decreases. Thus, $u(y_1) = u_M$ and $v(y_1) < 0$ for a certain $y_1 \in (y_o,b)$. However, $E(y_o) = V(u_m) \leq V(u_M) = V_M < E(y_1)$, which is a contradiction. Therefore, $|u(y)| < u_M$ $\forall y \in (-b,b)$ for every positive solution to (D)$_1$.

For the proof of the estimate in the n-dimensional case, consider the special solutions to (2.2), $u_+(y) = u(k\cdot(y-y_o)) = u(\theta)$, $k \in \mathbf{R}^n$, $|k|=1$, defined in the ball $B_R(y_o)$. $u(\theta)$ solves

$$u_{\theta\theta} + f(u) = 0, \quad \theta \in (-R,R), \tag{4.2}$$

provided $k\cdot\mu = 0$. Take $y_o \in \mathbf{R}^N$ and $R_o > 0$ so large that $\Omega_\omega \subset B_R(y_o)$. As it will be seen in the proof of Theorem 5, there exists $R \geq R_o$ such that a positive solution $u_R(\theta)$ to (4.2) exists in $(-R,R)$ and satisfies $u_R(\pm R) = 0$. Therefore, $u_+(y) = u_R R.(y-y_o)$ is an upper solution to (D). If $u(y)$ is a positive solution to (D), the previous one-dimensional estimate for $u_R(\theta)$ gives $0 \leq u(y) \leq u_+(y) < u_M, \forall y \in \Omega_\omega$, as wished.

Let us now prove part (b). The problem (D) can be written in the form,

$$\begin{cases} -\Delta v + \left(\frac{|\mu|^2}{4} - f'(0) - g(y,v,\mu)\right) v = 0 & \text{in } \Omega_\omega, \\ v = 0 & \text{on } \partial\Omega_\omega, \end{cases} \tag{4.3}$$

where $v(y) = e^{\frac{\mu y}{2}} u(y)$, $g(v) = f(v) - f'(0)v$, so $g(v) = o(v)$ as $v \to 0$, and $g(y,v,\mu) = e^{\frac{\mu y}{2}} g(e^{-\frac{\mu y}{2}} v)/v$. Assume that $u(y)$ is a nonnegative solution to (D). Since the estimate in part (a) holds we obtain

$$\left(\frac{|\mu|^2}{4} - f'(0) - \sup_{[0,u_M]} \frac{g(v)}{v}\right) \leq \left(\frac{|\mu|^2}{4} - f'(0) - g(y,v,\mu)\right), \tag{4.4}$$

for each $y \in \Omega_\omega$. Hence, via the maximum principle and (4.4), the unique nonnegative solution to (4.3) is $v \equiv 0$, provided that the following inequality holds:

$$|\mu| \geq \mu_1 \doteq \sqrt{4\left(f'(0) + \sup_{[0,u_M]} \frac{g(v)}{v}\right)}.$$

Therefore, the same conclusion holds for $u(y)$.

Finally, let us show part (c). To begin with, consider the case $f'(0) > 0$. As usual, define $\lambda_1(Q)$ as the principal eigenvalue of $-\Delta$ subject to homogeneous Dirichlet conditions in the bounded domain $Q \subset \mathbf{R}^n$. Designate by $\varphi_1(\cdot,Q)$ the positive associated eigenfunction. By using $\epsilon\varphi_1(y,\Omega_\omega)$, $\epsilon > 0$ conveniently small, as a lower solution to (4.3) (see [Am]), it is seen that the inequality

$$0 < \lambda_1(\Omega_\omega) < \left(f'(0) - \frac{|\mu|^2}{4}\right) \tag{4.5}$$

provides a sufficient condition for the existence of positive solutions both to (4.3) and (D), when $|\mu| < \sqrt{4f'(0)}$. Since $\sup_{[0,u_M]}\frac{g(u)}{u} \geq 0$ observe that $\sqrt{4f'(0)} \leq \mu_1$.

Consider now the problem (D) in the domains $\Omega_\lambda = \lambda\Omega$, $\lambda > 0$, and set $(\Omega_\lambda)_\omega = \bigcup_{[0,T]} \Omega_\lambda - c(t)$. Then $\Omega_\lambda \subset (\Omega_\lambda)_\omega$ so $0 < \lambda_1((\Omega_\lambda)_\omega) < \lambda_1(\Omega_\lambda)$. Since $\lambda_1(\Omega_\lambda) = \frac{\lambda_1(\Omega)}{\lambda^2}$ then (DW) admits positive solutions in Ω_λ provided that

$$\lambda \geq \sqrt{4\frac{\lambda_1(\Omega)}{(4f'(0) - |\mu|^2)}}$$

and $0 < |\mu| < \sqrt{4f'(0)}$. To obtain the desired result it is sufficient to take $\bar{\lambda} = \sqrt{\lambda_1(\Omega)/f'(0)}$ and $r(\lambda) = \sqrt{4(f'(0) - \lambda_1(\Omega))/\lambda^2}$.

As for the case when $f'(0) \leq 0$ let us introduce the auxiliary problem

$$\text{(4.6)} \qquad \begin{cases} -\Delta u + \lambda^2 \frac{|\mu|^2}{4} u = \lambda^2 \tilde{f}(y, \mu, u) & \text{in } Q \\ u = 0 & \text{on } \partial Q, \end{cases}$$

where $Q \subset \mathbf{R}^n$ is certain domain, $\tilde{f}(y,\mu,u) = e^{\frac{\mu y}{2}} \tilde{f}(e^{-\frac{\mu y}{2}} u)$, and the C^1 function $\tilde{f} = f$ in $[0, u_M]$ while $\tilde{f} \equiv 0$ outside $(-\epsilon, u_M + \epsilon)$, $\epsilon > 0$, and $\tilde{f} \geq 0$ for $u \leq 0$. Note that from the estimate in part (a), (4.6) is equivalent to (4.3) and hence to the problem (D), provided $\lambda = 1$ and $Q = \Omega_\omega$.

We claim the existence of $\bar{\lambda} > 0$ and $0 < r(\lambda) < \mu_1$ for each $\lambda \geq \bar{\lambda}$ such that (4.6) admits a positive solution $u_o(y, \lambda, \mu)$ in Q provided that $|\mu| \leq r(\lambda)$. The conclusion in part (c) follows from this claim as it is shown now. In fact, assume without loss of generality that $0 \in \Omega$ and choose $Q = B_\delta(0) \subset \Omega$. Then $Q_\lambda = B_{\lambda\delta}(0) \subset \Omega_\lambda \subset (\Omega_\lambda)_\omega$ for each $\lambda \geq \bar{\lambda}$. On the other hand, observe that $\tilde{u}_o(y, \lambda, \mu) \doteq u_o(\frac{y}{\lambda}, \lambda, \mu)$ solves $(4.6)_{\lambda=1}$ in $Q_\lambda = B_{\lambda\delta}(0)$ and, from Hopf's maximum principle $\frac{\partial \tilde{u}_o}{\partial \rho} < 0$ at $\rho = \lambda\delta$. This fact and Theorem II.3 in [Be-L,80] allow us to assert that the continuous function

$$u_{o,-}(y) = \begin{cases} \tilde{u}_o(y, \lambda, \mu), & y \in B_{\lambda\delta}(0), \\ 0, & y \in (\Omega_\lambda)_\omega - B_{\lambda\delta}(0) \end{cases}$$

defines a nonnegative lower solution to $(4.6)_{\lambda=1}$ in the domain $(\Omega_\lambda)_\omega$. Therefore, (D) admits a positive solution in $(\Omega_\lambda)_\omega$ for each $\lambda \geq \bar{\lambda}$ and $|\mu| \leq r(\lambda)$, as wished.

Finally let us prove the claim. The essential ideas in the variational argument employed are similar to those in [R], therefore the details will be omitted. Consider the problem

$$\text{(4.7)} \qquad \operatorname{minimize}_{u \in W_o^{1,2}(Q)} \Phi(\lambda, \mu, u),$$

where

$$\begin{aligned} \Phi(\lambda,\mu,u) &= \frac{1}{2}\int_Q \left\{ |\nabla u|^2 + \lambda^2 \frac{|\mu|^2}{4} u^2 \right\} dy - \lambda^2 \int_Q \tilde{V}(y, \mu, u)\, dy \\ &\doteq \Phi_1(\lambda, \mu, u) - \lambda^2 \Phi_2(\mu, u), \quad \lambda > 0, \ |\mu| \leq \mu_1, \end{aligned}$$

and where $\tilde{V}(y,\mu,u) = \int_0^u \tilde{f}(y,\mu,s)\, ds$. Now, assuming that $Q \subset \mathbf{R}^n$, $n \geq 3$, then Φ_2 is weakly continuous in $W_o^{1,2}(Q)$, while Φ_1 is always weakly lower semicontinuous in $W_o^{1,2}(Q)$ (see Chapter 4 in [C-H]). On the other hand, positive constants K_2 and K_3 exist such that $\Phi_1(\lambda, \mu, u) \geq K_2 |u|^2_{W_o^{1,2}(Q)}$ and $|\Phi_2(\mu, u)| \leq |u|_{W_o^{1,2}(Q)}$ for each

$u \in W_o^{1,2}(Q)$. Therefore, (4.7) is equivalent to

$$\text{minimize}_{|u|_{W_o^{1,2}(Q)} \leq R}\ \Phi(\lambda, \mu, u), \tag{4.8}$$

for a certain $R > 0$ ($|\mu| \leq \mu_1$). Thus, classical variational arguments [R] imply the existence of a solution $u_o(y, \lambda, \mu)$ to (4.8), which is also a weak solution to (4.6). Actually, by a bootstrap argument and the estimate in part (a), $u_o(y, \lambda, \mu)$ is a classical solution to (4.7). The choice of $\tilde{f}(u)$ together with the maximum principle imply that $u_o(y, \lambda, \mu) \geq 0$ in Q. On the other hand, it is possible to construct a function $\varphi \in W_o^{1,2}(Q)$, $|u|_{W_o^{1,2}(Q)} \leq R$ such that $\Phi_2(0, \varphi) > 0$ (see Theorem 1.13 in [R]). Thus, there exist $\bar{\lambda} > 0$ and $0 < r(\lambda) < \mu_1$ for each $\lambda \geq \bar{\lambda}$, such that $\Phi(\lambda, \mu, \varphi) < 0$ for $|\mu| \leq r(\lambda)$. Therefore the solution $u_o(y, \lambda, \mu)$ to (4.7) is positive in this range of the parameters. Finally, see [Ab-R] for the details of the proof of the case $n = 2$. □

Proof of Theorem 5. First, let us prove (i). If $w(x,t) = u(x - c(t)) = u(y)$ solves (DW) and it is positive in $\Omega = (-a, a)$, then $(u, v) = (u(y), u_y(y))$ is a solution to equation (3.3) with $c = 0$. Moreover, $u(\pm a \pm M) = 0$, $0 < u(y) < u_M$ for $|y| < M + a$, and the orbit Γ of $(u(y), v(y))$ meets $\{u > 0\}$ in a unique point $(u_o, 0)$. Thus, $E(u(\cdot), v(\cdot)) \equiv V(u_o)$ which implies $u_o = \min\{u > 0 / V(u) = V(u_o)\}$ and

$$a + M = \int_0^{u_o} \frac{du}{\sqrt{2(V(u_o) - V(u))}}. \tag{4.9}$$

If $V(u_o) = \alpha$ were critical then $f(u_o) = 0$. Since f is C^1 this would imply that the integral in (4.9) diverges, which is not possible.

As for (ii), note that $\Phi \subset \{V(u)/V'(u) = 0\}$ and then Sard's theorem implies that the Lebesgue measure $|\Phi| = 0$. Thus, $[0, V_M] \backslash \Phi$ is nonempty. Moreover, as a consequence of the continuous dependence of solutions to $(3.3)_{c=0}$ on initial conditions and parameters, $(0, V_M) \backslash \Phi$ is open. Thus, well-known topological properties of the real line $\mathbf{R}$ imply that $(0, V_M) \backslash \Phi$ has a countable quantity of connected pieces $\{I_n\}$, $I_n = (\alpha_n^-, \alpha_n^+)$, i.e., $(0, V_M) \backslash \Phi = \cup\, I_n$. Moreover, Sard's theorem again implies that $\Phi = \overline{\{\alpha_n^\pm\}}$.

To prove point (iii) it is necessary to introduce $u_n^\pm$ as u_n^-(resp. u_n^+) = max (resp., min)$\{u/V(u) = \alpha_n^-(\text{r. } \alpha_n^+)\}$. For $\alpha \in (\alpha_n^-, \alpha_n^+)$ the solution to $(3.3)_{c=0}$ starting at $(u(0), v(0)) = (0, \sqrt{2\alpha})$ reaches, at first time, the negative semiaxis $\{v < 0\}$ in finite time $y = T(\alpha)$. Previously, its orbit Γ_α meets $\{u > 0\}$ at $(u, v) = (u(\alpha), 0)$ with $V(u(\alpha)) = \alpha$ (see Fig. 5). Thus

$$T(\alpha) = 2 \int_0^{u(\alpha)} \frac{ds}{\sqrt{2(\alpha - V(u))}}\ ds.$$

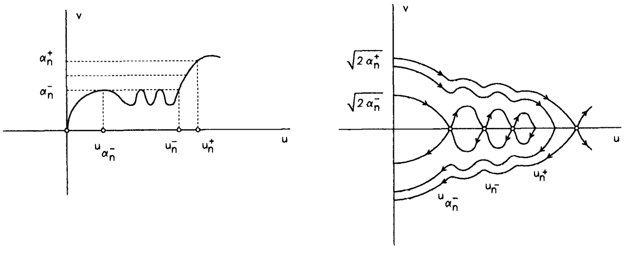


FIG. 5.

Therefore $w(x,t) = u(x - c(t))$ is a positive solution to (DW) in $\Omega = (-a,a)$ where $a = \frac{T(\alpha)}{2} - M$ and $M = \max \int_0^{\cdot} b(s)\,ds$. Since $T(\alpha) > 0$ for $\alpha \in (\alpha_n^-, \alpha_n^+)$ and $\lim_{\alpha \to \alpha_n^- +} T(\alpha) = \lim_{\alpha \to \alpha_n^+ -} T(\alpha) = +\infty$, the continuity of $T(\alpha)$ implies that $T_n \doteq \min_{\alpha \in (\alpha_n^-, \alpha_n^+)} T(\alpha) > 0$. This implies that (DW) has at least two positive solutions in $\Omega = (-a,a)$ provided $a > a_n \doteq \max\{0, \frac{T_n}{2} - M\}$.

We claim now that $T_o = \inf T_n > 0$. That fact will imply $\inf a_n > 0$ provided $T_o > M$, as desired. To prove the claim, write $T_n = T(V(\tilde{u}_n))$ for some $\tilde{u}_n \in (u_n^-, u_n^+)$ and set $\tilde{u} = \lim\ \inf \tilde{u}_n \geq 0$. Assume, by contradiction, that $T_o = 0$. We claim that $\tilde{u} = 0$. Otherwise,

$$\lim \int_0^{\tilde{u}_{n_1}} \frac{ds}{\sqrt{2(V(\tilde{u}_{n_1}) - V(s))}} = 0, \tag{4.10}$$

for a certain subsequence $\{\tilde{u}_{n_1}\}$ of $\{\tilde{u}_n\}$, with $\lim \tilde{u}_{n_1} = 0$. On the other hand,

$$\int_0^{\tilde{u}_n} \frac{ds}{\sqrt{2(V(\tilde{u}_n) - V(s))}} \geq \sqrt{\frac{\tilde{u}_n^2}{2V(\tilde{u}_n)}} > 0, \quad \forall n \in \mathbf{N}.$$

However, 0 accumulates zeros of $f(u)$ and then $\lim_{u \to 0+} \frac{V(u)}{u^2} = 0$. This contradicts (4.10) and the claim is proved. On the other hand, since $\tilde{u} > 0$ we arrive at

$$\int_0^{\tilde{u}} \frac{ds}{\sqrt{2(V(\tilde{u}) - V(s))}} = 0.$$

This fact together with $V(u) < V(\tilde{u})$ for $u \in (0, \tilde{u}]$ also lead to a contradiction.

Finally, the points (a)–(c) in part (iii) are now straightforward consequence of the discussion developed above. □

Remark 4.1. As a consequence of the nice results contained in [Sm-W,81] (see also [Sm]) it follows that the bistable $(\mathrm{DW})_{\mu=0}$, i.e., $f(u) = f_2(u)$, has exactly two positive solutions in $\Omega = (-a,a)$ provided that $a > a_1$ and $T_1 > 2M$.

5. Perturbation near critical zeros: Generation of chains. In the present section, the definitions and preliminary facts to prove the results concerning the one-dimensional (DW) are introduced. We should recall that solving (DW) in $\Omega = (-a,a)$ is equivalent to finding positive solutions to $(\mathrm{D})_1$ in $\Omega_\omega = (-a - M, a + M)$ (see §2). Thus, a solution $u(y)$ to $(\mathrm{D})_1$ defines a solution $(u(y), v(y)) = (u(y), u_y(y))$ to $(3.3)_{c=\mu}$ whose orbit Γ meets $\{u > 0, v = 0\}$ at a unique point $(u,v) = (u_1, 0)$. On the other hand, the orbit Γ can be described in terms of the orbital equation associated with (3.3) (see the equation (5.1) below). Finally, recall that it is sufficient to study (DW) for nonnegative μ (see Remark 2.3). For later use, some of those facts are summarized in the next definition.

DEFINITION 4. *A point $u_1 > 0$ is said to satisfy the condition $(\mathrm{S})_\mu$, $\mu \geq 0$, if there exists a solution $v(u)$, $v \in C^1([0,u_1)) \cap C^o([0,u_1])$, to the equation*

$$\frac{dv}{du} = -\mu - \frac{f(u)}{v} \tag{5.1}$$

such that $v(u_1) = 0$ and $v < 0$ in $[0, u_1)$.

Remark 5.1. (a) By using ODE arguments we can show that if u_1 satisfies $(\mathrm{S})_\mu$, then the unique orbit Γ to $(3.3)_{c=\mu}$ passing through $(u_1, 0)$ is such that the intersection $\Gamma^+ = \Gamma \cap \{u > 0\}$ can be written as $\Gamma^+ = \{(\tilde{u}(y), \tilde{v}(y))/y \in (-a,a)\}$, where $\tilde{u}(y)$

solves $(\mathrm{D})_1$ in $\Omega = (-a, a)$. Moreover, $\tilde{v}(y) = v(\tilde{u}(y))$ for $y_1 \leq\ y \leq\ a$ and certain $y_1 \in \Omega$ with $\tilde{u}(y_1) = u_1$.

(b) A certain u_1 satisfies $(\mathrm{S})_{\mu=0}$ provided it satisfies $(\mathrm{S})_\mu$, $\mu > 0$. In fact, since $E(u, v)$ decreases along solutions to $(3.3)_\mu$, then $v_o(u) = -\sqrt{2(V(u_1) - V(u))} < v(u) < 0$ for $u \in [0, u_1]$. A more elaborate argument proves that u_1 also satisfies $(\mathrm{S})_{\bar{\mu}}$ for each $\bar{\mu} \in [0, \mu]$.

In the proof of the existence of an MTW to (1.3) in the logistic ($f = f_1(u)$) and bistable ($f = f_2(u)$) cases, an outstanding role is played by the zero $u = 1$. It generates a saddle critical point $(u, v) = (1, 0)$ to equation (3.3) whose unstable manifold (see [C-H]) generates the MTW when c increases from $c = 0$ to $c = c_1^*$ or $c = c_2^*$ (see [F], [F-M,77]). In the next definition, the concept of saddle point is slightly extended to accomplish the objectives of the present work (compare with [C-H]).

DEFINITION 5. *Let $u_\alpha \in [0, u_M]$ be a critical zero of $f(u)$. The unstable (resp., stable) manifold $W_u^o(u_\alpha)$ (resp., $W_s^o(u_\alpha)$) associated to the critical point $(u, v) = (u_\alpha, 0)$, with regard to equation $(3.3)_{c=0}$, is defined as*

$$\begin{aligned} &W_u^o(u_\alpha)(\mathit{resp.},\ W_s^o(u_\alpha)) \\ &\quad = \{(u, v) \in \mathbf{R}^2 / 0 \leq\ u \leq\ u_\alpha, v = -(\mathit{resp.},\ +)\sqrt{2(V(u_\alpha) - V(u))}\}. \end{aligned}$$

To describe the process of CW generation, a suitable notion of perturbable manifolds is needed. Indeed, complex CWs arise in some cases when the unstable manifold $W_u^o(u_\alpha)$, associated to a critical zero $u = u_\alpha$, which accumulates zeros of $f(u)$ from the left, is perturbed from $c = 0$ to $c > 0$.

DEFINITION 6. *Let $u_\alpha \in [0, u_M]$ be as in Definition 5. The manifold $W_s^o(u_\alpha)$ (resp., $W_u^o(u_\alpha)$) is said to be perturbable to $\mu > 0$ if $\exists \epsilon > 0, \delta > 0$ and a unique solution $v = v^+(u, \mu)$ (resp., $v^-(u, \mu)$), with $0 \leq\ u(\mu) \leq\ u \leq\ u_\alpha$ and $\mu \in (0, \epsilon)$. In addition, $v^+(u, \mu)$ (resp., $v^-(u, \mu)$) must satisfy*

(i) $v^+(u, \mu)$ (resp., $v^-(u, \mu)$) $>$ (resp., $<$) $0, u \in [u(\mu), u_\alpha)$,

(ii) $\lim_{u \to u_\alpha(-)}\ v^+(u, \mu)$ (resp., $v^-(u, \mu)) = 0$.

If $W_s^o(u_\alpha)$ (resp., $W_u^o(u_\alpha)$) is perturbable, the stable (resp., unstable) manifold associated to $(u, v) = (u_\alpha, 0)$ regarding the equation $(3.3)_\mu$, $\mu \in\ (0, \epsilon)$, is defined as

$$\begin{aligned} &W_s^\mu(u_\alpha)(\mathit{resp.},\ W_u^\mu(u_\alpha)) \\ &\quad = \{(u, v) \in \mathbf{R}^2 / u(\mu) \leq\ u \leq\ u_\alpha,\ v = v^+(u, \mu)\ (\mathit{resp.},\ v^-(u, \mu))\}. \end{aligned}$$

Remark 5.2. (a) The manifolds $W_s^o(u_\alpha)$ and $W_u^o(u_\alpha)$ are perturbable provided that the critical zero $u = u_\alpha$ is isolated from the left (see Chapter 4 in [F]). Moreover, a result due to Kannel (see [F]) establishes that both $W_s^\mu(u_\alpha)$ and $W_u^\mu(u_\alpha)$ are monotonic with respect to μ.

(b) It will be shown later that $W_s^o(u_\alpha)$ is always perturbable (see Proposition 1). However, if the critical zero u_α accumulates zeros of $f(u)$ from the left, $W_u^o(u_\alpha)$ may not be perturbable. More precisely, a monotone chain $C(\mu)$ is generated when $W_u^o(u_\alpha)$ is not perturbable. Let us now give an example describing that situation.

Example 5.1. Let us proceed as in Example 3.1 to construct a suitable $f(u)$ in the equation (3.3). To accomplish this, let us choose similar sequences $\{\epsilon_n\}$, $\{u_n\}$, $\{f_n\}$ but now taking $f_o(u) = -u(\alpha - u)(1 - u)\varphi(u - \frac{1}{2})$, $0 < \alpha < \frac{1}{2}$, $\varphi(u)$ being the same function as in Example 3.1. Now, in addition to $\{u_n\}$, $f(u)$ has another sequence $\{w_n\}$ of zeros, where $w_1 = \alpha$ and $w_n = w_{n-1} + \epsilon_n$ for $n \geq\ 2$.

If we consider $u_\alpha\ =\ u_\infty\ =\ \lim\ u_n$ as the critical zero, then $W_u^o(u_\alpha)$ is not perturbable. In fact, if $W_u^o(u_\alpha)$ were perturbable, then for each $\mu > 0$, $\mu \to 0$, there would exist a negative solution $v(u,\mu)$ to (5.1), $u_\infty - \delta \leq u < u_\infty$, such that $\lim_{u\to u_\infty(-)} v(u,\mu) = 0$. Now, let $v(u,\mu,\epsilon)$ be the solution $v(u)$ to (5.1) such that $v(u_\infty) = -\epsilon$. Then, $v(u,\mu) > v(u,\mu,\epsilon)$ $\forall\ \epsilon > 0$ and $u \in [u_\infty - \delta, u_\infty]$ (see Fig. 6).

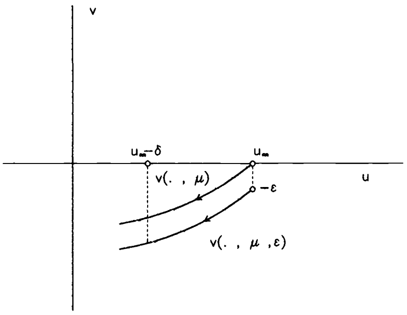


FIG. 6.

However, that behaviour of the family $\{v(\cdot,\mu,\epsilon)\}_{\epsilon>0}$ is not possible. In fact, arguing as in Example 3.1, it can be shown that the behaviour of equation $(5.1)_\mu$ in the strip $u_{n-1} \leq u \leq u_n$ is the same as that of $(5.1)_{\frac{\mu}{\epsilon_n}}$ in the strip $0 \leq u \leq 1$, but with the nonlinearity $f = f_o(u)$. Observe that (5.1) (or equivalently (3.3)) with $f = f_o(u)$ is an equation of bistable type. Therefore, there exists a unique $c_o^* > 0$ such that the equation (1.3), with $f = f_o(u)$, admits an MTW connecting $u = 1$ to $u = 0$ with velocity $c = c_o^*$ (see [F-M,77]). This fact implies that (1.3) with $f(u)$ as constructed here exhibits an MTW $C(\mu_n)$, connecting u_n to u_{n-1} with velocity $\mu_n = \epsilon_n c_o^*$. On the other hand, a careful analysis of the equation $(5.1)_\mu$ in the strip $u_{n-1} \leq u \leq u_n$ reveals the existence, for μ arbitrarily small, of a family of solutions $v_n(u)$ of $(5.1)_\mu$ with the following properties (see Fig. 7):

(i) $\exists n(\mu) \in \mathbf{N}$ such that $\forall\ n \geq n(\mu), v_n(u) < 0$ for $u_n < u \leq u_\infty$,

(ii) $v_n(u_n) = 0$ and $\lim v_n(u_\infty) = 0$.

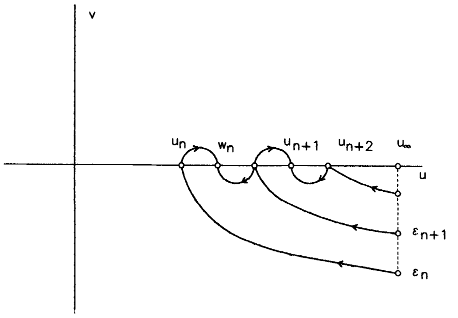


FIG. 7.

The existence of $\{v_n(u)\}$ with properties (i) and (ii) is not compatible with that of $\{v(u,\mu,\epsilon)\}$. Therefore $W_u^o(u_\alpha)$ is not perturbable.

Finally, observe that Example 5.1 shows an explicit case where option (ii) in Theorem 3 holds. In fact, observe that $\mu_n = c_o^*\epsilon_n$ is decreasing, while $C(\mu_n)$ moves toward the left when μ takes the values μ_n in the increasing sense. Observe that the largest existence value μ_o for this $f(u)$ is $\mu_o = c_o^*\epsilon_1$ (see Fig. 8).

5.1. Perturbability of W_s^o. The next result asserts that, under quite general conditions, $W_s^o(u_\alpha)$ is always perturbable.

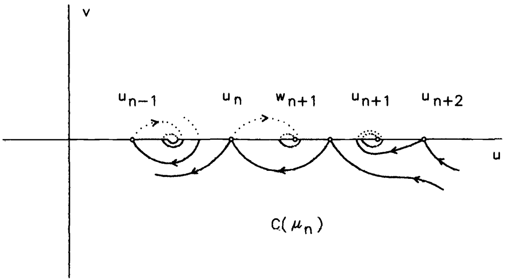


FIG. 8.

PROPOSITION 1. *Let $f(u)$ be a C^1 function with critical zero $u = u_\alpha > 0$. Assume that*

$$(5.2) \qquad f(u) = O((u - u_\alpha)^2), \quad as \quad u \to u_\alpha(-).$$

Then, $W_s^o(u_\alpha)$ is perturbable to every $\mu > 0$. Moreover, $v = v^+(\cdot, \mu)$ is defined in the interval $[0, u_\alpha]$ and is increasing with respect to μ.

Remark 5.3. (a) Condition (5.2) holds provided that $\frac{d^2f}{du^2}$ exists at $u = u_\alpha$.

(b) Proposition 1 furnishes new information when u_α accumulates zeros of $f(u)$ from the left and $f(u)$ exhibits, infinitely many times, both signs in $u < u_\alpha$. Under those conditions, our result sharpens a well-known result due to Kannel (see Lemma 4.14 in [F]), which arrives at the same assertion under the more restrictive requirement

$$f(u)(u - u_\alpha) \leq 0, \quad u \in (u_\alpha - \epsilon, u_\alpha).$$

Proof of Proposition 1. We will assume that $u = u_\alpha$ accumulates zeros from the left. First, let us prove the existence for each $\mu > 0$, of a solution $v(u)$ to $(5.1)_\mu$ such that

$$(5.3) \quad v(u) \in C^o([0, u_\alpha]) \cap C^1([0, u_\alpha)), v(u) > 0 \text{ for } 0 \leq u < u_\alpha, \text{ and } v(u_\alpha) = 0.$$

However, that is equivalent to proving the existence of a solution $(u(y), v(y))$ to $(3.3)_\mu$ such that $(u(0), v(0)) = (u_1, v_1)$, $u_1, v_1 > 0$, and $\lim_{y \to +\infty}(u(y), v(y)) = (u_\alpha, 0)$ with $v(y) > 0$ for $0 \leq y$. In fact, $(u(y), v(y))$ will meet backward $\{u = 0, v > 0\}$ in some point $(0, v^+)$ at a certain negative time y_o. Thus, the orbit Γ of $(u(y), v(y))$ in $\{u > 0\}$ will give the desired solution $v(u)$.

To find out the initial position (u_1, v_1), let us define the set

$$W = \{(u, v) \in \mathbf{R}^2 / v_o(u) \leq v \leq v_1(u), u \in [0, u_\alpha]\},$$

where $v_o(u) = \sqrt{2(\alpha - V(u))}$, $V(u_\alpha) = \alpha$, and $v_1(u) = \sqrt{2(\gamma - V(u))}$ with $0 < \alpha < \gamma$. Then W is a Wazewsky set (see [Co]) regarding $(3.3)_\mu$. Since the strict exit points set W^- of W is $W^- = E_1 \cup E_2 = \{(u, v) \in W / v = v_o(u)\} \cup \{(u_\alpha, v)/0 < v \leq v_1(u_\alpha)\}$ (see Fig. 9), W^- is not connected. Thus, Wazewsky's theorem (see [Co]) gives the existence of the desired point (u_1, v_1).

Now, we claim that if $v_i(u)$ are solutions to $(5.1)_{\mu_i}$, $i = 1, 2$, which satisfy (5.3), and $0 \leq \mu_1 < \mu_2$, then $0 < v_1(u) < v_2(u)$ for $0 \leq u < u_\alpha$. Indeed, the uniqueness, monotonicity, and perturbability properties of $W_s^o(u_\alpha)$ follow from that fact.

To prove the claim we first show that a solution $v(u)$ to (5.1), which satisfies (5.3), is also C^1 in the interval $[0, u_\alpha]$. In fact, $v_o(u) = \sqrt{2(\alpha - V(u))}$, i.e., $W_s^o(u_\alpha)$, is C^1 in that interval and $v'(u_\alpha) = 0$ (recall that u_α accumulates zeros of $f(u)$). The

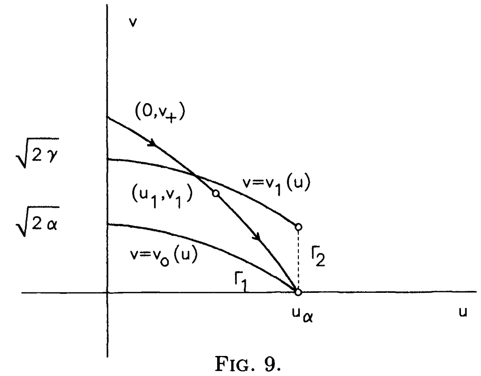


FIG. 9.

last assertion follows from the limit

$$\lim_{u\to u_\alpha(-)} \frac{f(u)}{\sqrt{\int_u^{u_\alpha} f(s)\,ds}} = 0. \tag{5.4}$$

However, since u_α accumulates zeros of $f(u)$, (5.4) is not obvious and its proof is delayed until Lemma 3 below. Finally observe that $0 < v_o(u) < v(u)$ in $[0, u_\alpha)$. Hence,

$$|v'(u) + \mu| = \left|\frac{f(u)}{v(u)}\right| < \left|\frac{f(u)}{v_o(u)}\right|, \quad u \in [0, u_\alpha).$$

Thus $v'(u_\alpha) = -\mu$ and $v(u)$ is C^1 up to $u = u_\alpha$.

On the other hand, set $w(u) = v_2(u) - v_1(u)$ and $\Delta\mu = \mu_2 - \mu_1$. From (5.1) we arrive at

$$\frac{dw}{du} = -\Delta\mu + \frac{f}{v_1 v_2}, \quad u \in [0, u_\alpha),$$

which implies that the function $g(u) = w(u)\exp\left\{-\int_{u_1}^{u} \frac{f(s)}{v_1(s)v_2(s)}\,ds\right\}$, $0 < u_1 < u_\alpha$, is decreasing. Since

$$\frac{f(u)}{v_1(u)v_2(u)} = \left\{\frac{1}{\mu_1\mu_2 + o(1)}\right\}\frac{f(u)}{(u-u_\alpha)^2}, \quad \text{as } u \to u_\alpha(-),$$

then the integral $\int_{u_1}^{u} \frac{f(s)}{v_1(s)v_2(s)}\,ds$ converges as $u \to u_\alpha$. Thus, $g(u)$ is bounded and $\lim_{u\to u_\alpha(-)} g(u) = 0$. Therefore, $w(u) > 0$ in $0 \le u < u_\alpha$. □

LEMMA 3. *Under the assumptions of Proposition* 1 *assume that* $u = u_\alpha$ *accumulates zeros of* $f(u)$ *from the left. Then, the limit in* (5.4) *holds true.*

Proof. Since u_α is a critical zero for $f(u)$, then $\int_u^{u_\alpha} f(s)\,ds = V(u_\alpha) - V(u) > 0$ for $u \in (0, u_\alpha)$. From the fact that u_α accumulates zeros of $f(u)$ from the left, it follows that 0 is the unique possible value for the limit in (5.4).

Let us define $G(u) = \frac{f^2(u)}{F(u)}$ where $F(u) = \int_u^{u_\alpha} f(s)\,ds$. Then, (5.4) is equivalent to $\lim_{u\to u_\alpha} G(u) = 0$. Note that the classical L'Hôpital rule cannot be used in this case. To show the existence of the last limit consider the sets

$$U^+ \text{ (resp., } U^-,\ U^o) = \{u/0 < u < u_\alpha, f(u) > \text{(resp., } <,\ =) \ 0\}.$$

Then, it is sufficient to show that

$$\lim_{u\in U^+, u\to u_\alpha} G(u) = 0 \quad \text{and} \quad \lim_{u\in U^-, u\to u_\alpha} G(u) = 0. \tag{5.5}$$

To prove the first limit, define $w(u_o) = \max\{u' \in U^o / u' < u_o\}$ for every $u_o \in U^+$. Since $f(u) > 0$ in the interval $(w(u_o), u_o]$ then $F(u_o) - F(w(u_o)) < F(u_o)$. Thus, Cauchy's mean value theorem gives the existence of $\xi(u_o) \in (w(u_o), u_o)$ such that

$$G(u_o) = \frac{f^2(u_o)}{F(u_o)} \leq \frac{f^2(u_o) - f^2(w(u_o))}{F(u_o) - F(w(u_o))} = \frac{2f(\xi(u_o))f'(\xi(u_o))}{-f(\xi(u_o))} = -2f'(\xi(u_o)).$$

Since $f(u)$ is C^1, that means that the first limit in (5.5) holds true. The second one is proven in a similar way. □

5.2. Perturbability of W_u^o: Chain of waves formation. The relation between the perturbability of the unstable manifold $W_u^s(u_\alpha)$ and the generation of CWs to equation (1.3), will now be studied in detail. Let u_α be a critical zero for $f(u)$. It will be said, for short, that $u = u_\alpha$ is perturbable when $W_u^o(u_\alpha)$ be perturbable.

Our first objective will be to characterize the perturbability property for critical zeros $u = u_\alpha$.

PROPOSITION 2. *Assume that $f(u)$ satisfies $(\mathrm{H})_f$ and let $u = u_\alpha$ be a critical zero of $f(u)$. Then, u_α is perturbable if and only if the following condition holds:*

$$\text{(5.6)} \qquad \begin{aligned} &\exists \mu_\epsilon > 0 \textit{ and } \{u_n\} \subset (0, u_\alpha),\ \lim u_n = u_\alpha, \textit{such} \\ &\textit{that } \forall n \in \mathbf{N}, u = u_n \textit{ satisfies condition } (\mathrm{S})_{\mu=\mu_\epsilon} \end{aligned}$$

Proof. (a) *Necessity of* (5.6). Assume that (5.6) does not hold for $\mu \in (0, \epsilon)$, $\epsilon > 0$. We will arrive at a contradiciton.

First, observe that the set $U = f^{-1}((0, +\infty)) \cap (0, u_\alpha)$ is open. Thus it has, at most, countable many connected pieces $\{I_n\}$, $I_n = (u_n^-, u_n^+)$. Since $u = u_\alpha$ is critical then $\lim u_n^- = \lim u_n^+ = u_\alpha$. Moreover, $\forall n \in \mathbf{N}, f(u_n^\pm) = 0$.

Fix $\mu \in (0, \epsilon)$ and take $n \in \mathbf{N}$ large enough so that no point in $I_n = (u_n^-, u_n^+)$ satisfies $(\mathrm{S})_\mu$. Then, $\forall w \in I_n$ the solution $(u(y), v(y))$ to $(3.3)_\mu$ starting at $(u(0), v(0)) = (w, 0)$ reaches $\{u > 0, v = 0\}$ at $(u, v) = (w^-, 0)$ in finite time or $\lim_{y \to +\infty}(u(y), v(y)) = (w^-, 0)$. In both cases $0 \leq w^- \leq u_n^- < w$.

Next, consider an increasing sequence $\{w_m\} \subset I_n$ with $\lim w_m = u_n^+$. Designate by $v = g_m(u)$ that piece of the orbit to $(3.3)_\mu$ passing through $(u, v) = (w_m, 0)$, which is contained in $\{v \leq 0\}$. Observe that $g_m(u)$ is defined in the interval $[w_m^-, w_m]$. Moreover, $\{w_m^-\}$ is decreasing and $\lim w_m^- = w_n$, with $w_n \leq u_n^-$ (see Fig. 10). Note also that $\forall m \in \mathbf{N}$, $[w_m^-, w_m] \subset [w_{m+1}^-, w_{m+1}]$, and $g_{m+1}(u) < g_m(u)$, for every $u \in [w_m^-, w_m]$. Finally, let us introduce the sequence of functions $\tilde{g}_m(u)$, $u \in [w_n, u_n^+]$, as follows:

$$\tilde{g}_m(u) = \begin{cases} g_m(u) & \text{if } w_m^- \leq u \leq w_m, \\ 0 & \text{if } u \in [w_n, u_n^+] \backslash [w_m^-, w_m]. \end{cases}$$

Since u_α is perturbable, the solution $v^-(u, \mu)$ quoted in Definition 6 exists and $v^-(u, \mu) < \tilde{g}_m(u)$ for every $u \in [w_n, u_n^+]$. Therefore, the limit $h_n(u) = \lim \tilde{g}_m(u)$ exists for every $u \in [w_n, u_n^+]$.

On the other hand we claim that $h_n(\cdot) \in C^o([w_n, u_n^+]) \cap C^1((w_n, u_n^+))$. Moreover, the graph of $v = h_n(u)$ in the interval (w_n, u_n^+) is a piece of an orbit of the equation $(3.3)_\mu$. In fact, by using the equation $(5.1)_\mu$ it follows that $\tilde{g}_m \longrightarrow h_n$ in the topology of $C^1([\alpha, \beta])$, for every $[\alpha, \beta] \subset (w_n, u_n^+)$. Thus, the orbital character of $v = h_n(u)$ follows from applying the continuous dependence on initial data and parameters (see [Cp], especially Theorem 3) to equation $(5.1)_\mu$. In addition, since $h_n(w_n) = h_n(u_n^+) = 0$ it is easily seen that $h_n(u)$ is continuous at $u = w_n$ and $u = u_n^+$ (note that $(5.1)_\mu$ is singular at $v = 0$). Then, Dini's theorem and the continuity of $h_n(u)$ in $[w_n, u_n^+]$ imply that the limit $\tilde{g}_m \to h_n$ is uniform in that interval. Thus, $v = h_n(u)$, $u \in [w_n, u_n^+]$, is an orbit, or a piece of an orbit, of the equation $(3.3)_\mu$. Moreover, $v^-(u, \mu) < h_n(u)$ in $[w_n, u_n^+]$ (see Fig. 10).

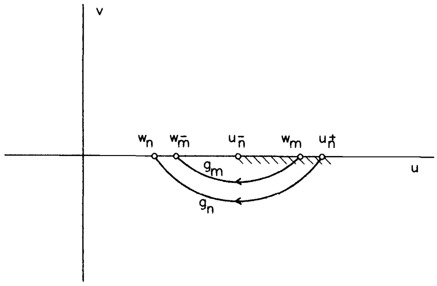


FIG. 10.

Finally, choose $\delta > 0$ and $n_o \in \mathbf{N}$ such that $v^-(\cdot, \mu)$ is defined in $[u_\alpha - \delta, u_\alpha]$, and $I_n \subset [u_\alpha - \delta, u_\alpha]$ for each $n \geq n_o$. Let us define the set Δ_δ which consists of the points (u, v) with $u \in [u_\alpha - \delta, u_\alpha]$ such that $v^-(u, \mu) \leq v \leq h_n(u)$ provided that $u \in I_n$ for some $n \geq n_o$, or $v^-(u, \mu) \leq v \leq 0$ if $u \notin I_n$ for all $n \in \mathbf{N}$. A careful analysis of the boundary $\partial\Delta_\delta$ of Δ_δ reveals that Δ_δ is strictly negatively invariant regarding $(3.3)_\mu$ (see [Co]). That is, $\Delta_\delta \subset \Phi_y(\Delta_\delta)$ and $\Delta_\delta \neq \Phi_y(\Delta_\delta)$ for small $y \geq 0$, where Φ_y stands for the flow of the equation $(3.3)_\mu$. In addition, since div $(v, -\mu v - f(u)) = -\mu < 0$ then $(3.3)_\mu$ is dissipative, which contradicts the fact $|\Delta_\delta| < |\Phi_y(\Delta_\delta)|$ for $y \geq 0$ small ($|A|$ = the Lebesgue measure of A). Thus, the necessity of (5.6) is proven.

(b) *Sufficency of* (5.6). Assume that (5.6) holds and, for each $n \in \mathbf{N}$, let $v_n(u)$ be the negative solution to $(5.1)_{\mu_\epsilon}$ associated to u_n (see Definition 4). Let us now introduce the decreasing sequence of functions $g_n(u) \in C^o([0, u_\alpha])$ defined as

$$g_n(u) = \begin{cases} v_n(u), & 0 \leq u \leq u_n, \\ 0, & u_n < u \leq u_\alpha. \end{cases}$$

If $v(u, \delta)$, $\delta > 0$, designates the solution $v(u)$ to $(5.1)_{\mu_\epsilon}$ such that $v(u_\alpha) = -\delta$ then $g_n(u) > v(u, \delta)$ in $0 \leq u \leq u_\alpha$, for each $n \in \mathbf{N}$ and $\delta > 0$. Thus, the function $g(u) = \lim g_n(u) = \inf g_n(u)$ satisfies the last inequality for $u \in [0, u_\alpha)$, with $g(u_\alpha) = 0$. Moreover, arguing as in part (a) we determine that $g(u) \in C^1([0, u_\alpha))$ and solves $(5.1)_{\mu_\epsilon}$ in $[0, u_\alpha)$ (cf. Theorem 3 in [Cp]). Since $g(u) > v(u, \delta)$ for each $\delta > 0$, then $g(u) \in C^o([0, u_\alpha])$ and $g_n \to g$ uniformly in $[0, u_\alpha]$.

Define now $v(u, \mu_\epsilon) = \sup_{\delta \to 0^+} v(u, \delta)$ in $[0, u_\alpha]$. By using again the arguments of part (a) it follows that $v(u, \mu_\epsilon)$ solves $(5.1)_{\mu_\epsilon}$ in $[0, u_\alpha]$, with $v(u_\alpha, \mu_\epsilon) = 0$. Moreover, $v(u, \mu_\epsilon) \leq g(u) < 0$ for $0 \leq u < u_\alpha$. Thus, every solution $v(u)$ to $(5.1)_{\mu_\epsilon}$ which satifisfies (i) and (ii) in Definition 6, must satisfy in addition the inequality

$$v(u, \mu_\epsilon) \leq v(u) \leq g(u), \quad u \in [0, u_\alpha].$$

We claim that $v(u, \mu_\epsilon) = v(u) = g(u)$ for $u \in [0, u_\alpha]$. Otherwise the set

$$\Delta_\epsilon = \{u/u_\alpha - \epsilon \leq u \leq u_\alpha, \ v(u, \mu_\epsilon) \leq v \leq g(u)\},$$

for $\epsilon > 0$ small enough, has $|\Delta_\epsilon| \neq 0$. That is not compatible with the dissipative character of $(3.3)_{\mu_\epsilon}$ (see part (a)). Therefore, there exists a unique solution $v^-(u, \mu_\epsilon) = v(u, \mu_\epsilon)$ to $(5.1)_{\mu_\epsilon}$ which satisfies (i) and (ii) in Definition 6. Now, by using Remark

5.1(b), the argument given here for the existence of $v^-(u,\mu_\epsilon)$ can be repeated to prove the existence of $v^-(u,\mu)$ for $0<\mu\le\ \mu_\epsilon$ and u_α is perturbable, as wished. □

The next result describes the behaviour near nonperturbable zeros u_α. It is useful to explain how the perturbation of that kind of zero generates a CW of (1.3) (see also Proposition 4 below).

PROPOSITION 3. *Assume that $f(u)$ satisfies* (H)$_f$ *and let $u=u_\alpha$ be a nonperturbable critical zero of $f(u)$. Then $\exists\ \epsilon_o>0$ such that $\forall\mu\in(0,\epsilon_o)$ there exists a perturbable critical zero $u=u(\mu)$ of $f(u)$, $0<u(\mu)<u_\alpha$, such that $\lim_{\mu\to0+}u(\mu)=u_\alpha$. Moreover, no point in the interval $u(\mu)<u<u_\alpha$ satisfies the condition* (S)$_{\tilde\mu}$ *for $\mu\le\ \tilde\mu$.*

Proof. Since $u=u_\alpha$ is a nonperturbable critical zero for $f(u)$ then it accumulates zeros of $f(u)$ from the left. Otherwise, u_α would be isolated from the left and Kannel's result (see Lemma 4.14 in [F]) would imply the perturbabilty of u_α. On the other hand (see the proof of Theorem 5 for the notation) since α is a critical value for $V(u)$, a subsequence $\{\alpha_{n_1}^+\}$ of $\{\alpha_n^+\}$ exists such that $\lim\alpha_{n_1}^+=\alpha$ with $\alpha_{n_1}<\alpha$ for all $n_1\in\mathbf{N}$. Rename $\{\alpha_{n_1}^+\}$ as $\{\alpha_n^+\}$ and let u_n^+ be the critical zero associated to α_n^+. Following Definition 3 we have $\lim u_n^+=u_\alpha$. Recall that, from the proof of Theorem 5, $f(u)>0$ in $u_n^-<u<u_n^+$. Thus, direct computations show that u_n^+ is perturbable for each n. Therefore, $\forall n\in\mathbf{N}$, $\exists\mu_n>0$ such that (5.6) holds for $0<\mu<\mu_n$. For $\mu>0$ small enough, let us define the set

$$\mathcal{S}_\mu(u_\alpha)=\{u/0<u<u_\alpha,\ \text{and } u \text{ satisfies condition } (\mathrm{S})_\mu\ \}. \tag{5.7}$$

$\mathcal{S}_\mu$ (u_α will be omitted for brevity) is a nonempty open set which is bounded from above. If we denote $u(\mu)=\sup\ \mathcal{S}_\mu$ then it is easily seen that $f(u(\mu))=0$. Thus $u(\mu)$ is a zero for $f(u)$ which is also critical (see Definition 3). Finally, since $u(\mu)$ is perturbable (see Proposition 2) then it holds that $u(\mu)<u_\alpha$. Observe that no point $u\in(u(\mu),u_\alpha)$ satisfies condition (S)$_{\tilde\mu}$ for all $\tilde\mu\ge\mu$. □

Remark 5.4. Let us analyze two situations of nonperturbability with different behaviours regarding the CW formation. In both Examples 3.1 and 5.1 $u_\alpha=u_\infty$ is nonperturbable. In Example 3.1, $u(\mu)=u_{n-1}$ for $0<\mu<\mu_{n-1}$ (recall that $\mu_n=c_o^*\epsilon_n$), and no point in the interval (u_{n-1},u_∞) satisfies (S)$_\mu$ due to the existence of a CW, $C(\mu)$, which connects u_∞ to u_{n-1} with velocity $0<\mu<\mu_{n-1}$ (see Fig. 11).

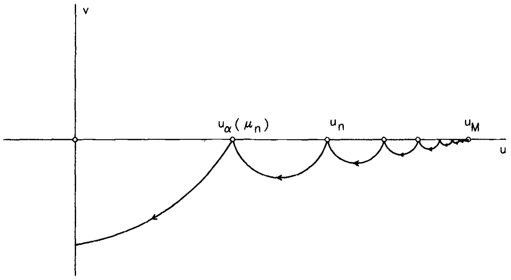


FIG. 11.

In Example 5.1 we have again $u(\mu)=u_{n-1}$ for $0<\mu<\mu_{n-1}$. However, condition (S)$_\mu$ now fails in $(u(\mu_{n-1}),u_\infty)$ for other reasons. In fact, CWs to (1.3) only occur with velocities $0<\mu_m=c_o^*\epsilon_m\le\mu_{m-1}$ and connect u_m to u_{m-1}, where $u_{n-1}\le u_{m-1}<u_m<u_\infty$.

The arising of CWs to (1.3) due to perturbation of $W_u^\mu(u_\alpha)$ at perturbable critical zeros u_α is described in the next result. Recall that (see Definition 6)

$$W_u^\mu(u_\alpha) = \{(u,v)/0 \leq u \leq u_\alpha, v = v^-(u,\mu)\},$$

provided that $\mu > 0$ is small enough, say $0 < \mu < \epsilon$. However, to handle the perturbation of W_u^μ for $\mu > 0$ large, it is convenient to extend the function $v^-(\cdot,\mu)$ for μ not small. To accomplish that, let $v(u,\mu,\delta)$ be the solution $v(u)$ to $(5.1)_\mu$ such that $v(u_\alpha) = -\delta$, $\delta > 0$. Then,

$$v^-(u,\mu) = \sup_{\delta \to 0+} v(u,\mu,\delta), \quad \mu \in (0,\epsilon).$$

However, even when μ is not small, we can still define for $u \in [0,u_\alpha]$ the function

$$g(u,\mu) = \min\{0, \sup_{\delta \to 0+} v(u,\mu,\delta)\},$$

provided that the supremum is defined at $u \in [0,u_\alpha]$ (that is just the case when μ is small). Otherwise we will set $g(u,\mu) = 0$.

PROPOSITION 4. *Let $u = u_\alpha$ be a perturbable critical zero for $f(u)$. Define*

$$\mu(u_\alpha) = \sup\ \{\mu > 0/g(u,\mu) < 0, \textit{for all } 0 \leq u < u_\alpha\},$$

and $u(u_\alpha) = \inf\ \{u \geq 0/g(u,\mu(u_\alpha)) = 0\}$. Then,

$$C = \{(u,v)/u(u_\alpha) \leq u \leq u_\alpha, v = g(u,\mu(u_\alpha))\}$$

is a CW to (1.3), *which connects the zeros $u = u_\alpha$ and $u = u(u_\alpha)$ with propagation velocity $c = \mu(u_\alpha)$. Moreover, the unstable manifold $W_u^\mu(u_\alpha)$ is strictly increasing and depends continuously on μ.*

However, the perturbation process of W_u^μ for $\mu > 0$ exhibits special features when the value $\mu = \mu_o$, i.e., the largest existence value for the problem (DW) (see (3.2)), is reached. That behaviour is precisely stated in the next result. Recall the definition of the set $\mathcal{S}_\mu$ in (5.7).

PROPOSITION 5. *Let u_α be a perturbable critical zero for $f(u)$, and let μ_o be the largest existence value for the problem* (DW). *Then,*

(i) *$\mu(u_\alpha) < \mu_o$ implies that $u = u(u_\alpha)$ is a perturbable critical zero with $u(u_\alpha) > 0$. Moreover, $\mathcal{S}_\mu \subset (0, u(u_\alpha))$ for $\mu(u_\alpha) \leq \mu < \mu_o$.*

(ii) *$\mu(u_\alpha) = \mu_o$ implies that $u(u_\alpha) = 0$.*

The proofs of Propositions 4 and 5 are a direct consequence of the ideas contained in the present section and do not involve special arguments to be explicitly required for the continuity of the exposition. Therefore, they are omitted.

6. The one-dimensional case: The proofs of Theorems 2–4.

Proof of Theorem 2. The proofs of parts (A) and (B) are similar. Thus, it is sufficient to give that of (A) to show the main ideas and techniques involved. Recall that there is no loss of generality if it is assumed that $\mu \geq 0$ (see Remark 2.3).

To prove (a), let us first describe the problem $(\mathrm{DW})_{\mu=0}$. Following the notation of the proof of Theorem 5, observe that a unique interval I_n exists. Namely, $I_1 = (u_1^-, u_1^+)$, where $u_1^- = (2(1+\alpha) - \sqrt{4(1+\alpha)^2 - 18\alpha})/3$, $u_1^+ = 1$. Indeed, only a unique critical zero exists, $u_M = u_1^+ = 1$ with energy $V_M = \frac{1}{6}(\frac{1}{2} - \alpha)$ (recall that $0 < \alpha < \frac{1}{2}$). For $\mu = 0$, the set $\mathcal{S}_\mu = (u_1^-, u_1^+)$ (see (5.7) and Fig. 12(a)).

For $\mu \geq 0$ and $u_1 \in \mathcal{S}_\mu$, designate by $T(u_1,\mu)$ the time $2a$ employed by a solution of $(3.3)_\mu$ in reaching $\{u = 0, v < 0\}$ after starting at $\{u = 0, v > 0\}$, travelling into $\{u \geq 0\}$, and meeting $\{u > 0, v = 0\}$ at $(u,v) = (u_1, 0)$ (see Remark 5.1(a)). Observe that

$$T(u_1,0)=\int_0^{u_1}\frac{ds}{\sqrt{2(V(u_1)-V(s))}}.$$

Now, let us study how the set $\mathcal{S}_\mu$ is perturbed when $\mu\in(0,\epsilon)$ and $\epsilon>0$ is small enough. Observe that $(0,0)$ and $(1,0)$ are saddle critical points for $(3.3)_\mu$. Let $v^-(u,\mu)$, $0\le u\le 1$, be the stable manifold associated to $(1,0)$ and let $v_o^-(u,\mu)$, $0\le u\le u_1^-(\mu)$, be that piece of the stable manifold of $(0,0)$ (see [C-H]), that is contained in the semiplane $\{v\le 0\}$ (see Fig. 12(b)).

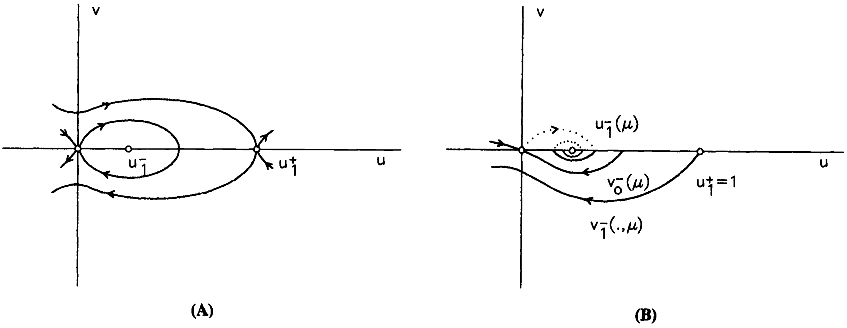


FIG. 12.

From Proposition 4 we have that $v_1^-(\cdot,\mu)$ is increasing with respect to μ. From an argument similar to that given in the proof of Proposition 1, it follows that $v_o^-(\cdot,\mu)$ is decreasing in the interval $[0,u_1^-(\mu)]$. Notice that $u_1^-(0)=u_1^-$. Since $f'(\alpha)>0$, then the behaviour of the orbits to $(3.3)_\mu$ near $(\alpha,0)$ together with Proposition 4 imply that $u(u_M)=0$. Thus, Proposition 5 asserts that $\mu(u_M)=c_2^*=\mu_o$, which implies the existence of a TW which connects $u_M=1$ to $u=0$ with velocity c_2^*. Moreover, from the uniqueness of the stable manifold associated to $(0,0)$ it follows that $v_o^-(u,\mu_o)=v_1^-(u,\mu_o)$ for $0\le u\le 1$. Finally, the set $\mathcal{S}_\mu=(u_1^-(\mu),1)$ for $\mu\in(0,\mu_o)$.

To prove b) observe that $T(\cdot,\mu)$ is continuous and positive in $\mathcal{S}_\mu$. Standard ODE results ensure that $\lim_{u\to u_1^-(\mu)+}T(u,\mu)=\lim_{u\to 1-}T(u,\mu)=+\infty$. Therefore, $T_\mu\doteq\inf_{\mathcal{S}_\mu}T(\cdot,\mu)>0$ for each $\mu\in(0,\mu_o)$. We now claim that $T_\mu>\frac{T_o}{2}>0$ for $0\le\mu<\mu_o$. In fact, following the notation of Remark 5.1(a) define $L(u_1,\mu)=a-y_1$ for every $u_1\in\mathcal{S}_\mu$. Then,

$$L(u_1,\mu)=-\int_0^{u_1}\frac{du}{v(u)}.$$

Now, fix $u_1\in(u_1^-,u_1^+)$. Since $u_1^-(\mu)$ is increasing in μ then $u_1\in\mathcal{S}_\mu$ if $0\le\mu<\tilde\mu(u_1)$, for a unique $\tilde\mu(u_1)$. Moreover, $u_1=u_1^-(\tilde\mu(u_1))$. From the above expression for L we have that $L(u_1,\mu)$ is increasing for $0\le\mu<\tilde\mu(u_1)$. Standard ODE results imply again that $\lim_{\mu\to\tilde\mu(u_1)-}L(u_1,\mu)=+\infty$ (see Fig. 13). Since $L(u_1,0)=\frac{T(u_1,0)}{2}$ for every $u_1^-<u_1<u_1^+$, we conclude that

$$T_\mu=\inf_{\mathcal{S}_\mu}T(\cdot,\mu)>\frac{T_o}{2}>0\quad\forall\mu\in(0,\mu_o).$$

By choosing $a(\mu)=T_\mu-M$ we arrive at the conclusions of part (b). □

Remark 6.1. For $f(u)=f_2(u)$ it was proven in [Sm-W,81] that $T(\cdot,0)$ has a unique local extreme at a certain $\tilde u$. Therefore, (DW) has at least two solutions in

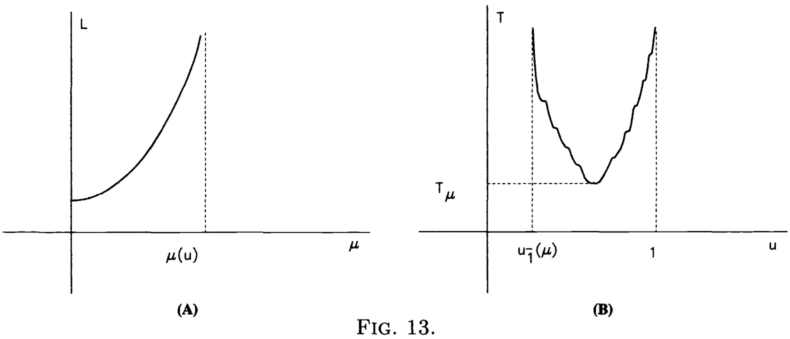


FIG. 13.

$\Omega = (-a, a)$ for $a > a(\mu)$ and $\mu > 0$, while the number of solutions is exactly two when $\mu = 0$ and $a > a(0)$.

Proof of Theorem 3. First, it can be assumed that u_M is perturbable. Otherwise, it is sufficient to work with the pertubable critical zero $u_M(\mu)$, $0 < \mu < \epsilon_o$, whose existence is given by Proposition 3. In the last case recall that $\mathcal{S}_{\tilde{\mu}} = \{u > 0/u < u_M$ and u satisfies $(\mathrm{S})_{\tilde{\mu}}\} \subset (0, u_M(\mu))$, for $\mu \leq \tilde{\mu} < \mu_o$.

The strategy of the proof consists in applying recursively Proposition 4 to generate two sequences $\{u_n\}$ and $\{\mu_n\}$, which take the zero $u = u_M$ as starting point. More precisely, we set $u_1 = u_M$, $\mu_1 = \mu(u_M)$ and $u_n = u(u_{n-1})$, $\mu_n = \mu(u_n)$ for each $n \geq 2$. Thus, $\{u_n\}$ defines a decreasing sequence of peturbable zeros of $f(u)$, while μ_n is increasing. Moreover, $\forall n \in \mathbf{N}$ there exists a CW, $C(\mu_n)$, which connects u_n to u_{n+1} with velocity μ_n.

Observe that both sequences $\{u_n\}$ and $\{\mu_n\}$ could possibly be finite. Indeed, the conclusion of part (i) is obtained when there exists $n_o \in \mathbf{N}$ such that $u_{n_o} = 0$. In this case (see Proposition 5) $\mu_{n_o - 1} = \mu_o$. Therefore, there exists a CW to (1.3) , $C(\mu_o)$, which connects the zero $u_+ = \max\, C(\mu_o) \cap \{(u, 0)/0 \leq\ u \leq\ u_M\}$ to $u = 0$ with velocity μ_o.

Assume now that $\{u_n\}$ is nonfinite. In that case, $u_n > 0$ for every n and Proposition 5 implies that $\mu_n < \mu_{n+1} < \mu_o$, $\forall n \in \mathbf{N}$. Since $\{u_n\}$ is decreasing then the limits $u_\infty = \lim u_n$, $\mu_\infty = \lim \mu_n$ exist and satisfy $u_\infty \geq 0$, $\mu_\infty \leq\ \mu_o$.

We claim that $u_\infty = 0$ implies $\mu_\infty = \mu_o$. In fact, the inequality $\mu_\infty < \mu_o$ would give the existence of some $\tilde{u}$ satisfying condition $(\mathrm{S})_\mu$ for some $\mu \in (\mu_\infty, \mu_o)$. Thus, $\forall n \in \mathbf{N}$ we would get $0 < \tilde{u} < u_n$, which contradicts the fact $\lim u_n = 0$. Therefore μ_∞ must coincide with μ_o. Observe that option (ii) in Theorem 3 is obtained under that asumption.

Next, consider the case where $u_\infty > 0$. First, we will prove that $V(u_\infty) > 0$ (observe that $V(u_\infty) \geq\ 0$). If fact, designate by $v^-(u, \mu_n)$ the unstable manifold associated to $u = u_n$ and designate also $g_n(u)$, the restriction of $v^-(u, \mu_n)$ to the interval $[0, u_\infty]$. If we assume that $V(u_\infty) = 0$ then we arrive to the inequalities

$$-\sqrt{2(V(u_n) - V(u))} < g_n(u) < -\sqrt{-2V(u)}, \quad 0 \leq\ u \leq\ u_\infty,$$

for each $n \in \mathbf{N}$. Observe that $g(u) = \lim g_n(u)$ satisfies $g(u) = -\sqrt{-2V(u)}$. However, that is not possible. In fact, the set $I = \{(u, v)/v = g(u)\}$ should be an invariant set for the equation $(3.3)_{\mu=\mu_\infty}$. On the other hand, the energy $E(u, v) = 0$ for $(u, v) \in I$. This contradicts the fact $\frac{dE}{dy} = -\mu_\infty v^2(y)$, where the derivative is computed over a solution $(u(y), v(y))$ of $(3.3)_{\mu=\mu_\infty}$. Therefore, $V(u_\infty) > 0$.

Let us study the case where $u_\infty > 0$ and $\mu_\infty = \mu_o$. If we keep the notation of the preceding paragraph for the functions $g_n(u)$, it will be shown that $C = \{(u,v)/0 \leq u \leq u_\infty, v = g(u)\}$ where $g(u) = \lim g_n(u)$ is a CW which connects $u = u_\infty$ to $u = 0$ with velocity μ_o. Thus, we arrive to the conclusion of part (i) in this case. In fact, observe that C is an invariant set for the equation $(3.3)_{\mu_\infty}$. On the other hand, the inequality

$$-\sqrt{2(V(u_n) - V(u_\infty))} < g_n(u_\infty) < 0, \quad \forall n \in \mathbf{N}$$

implies that $g(u_\infty) = 0$. In addition, $g(0) = 0$. Otherwise, $g(0) < 0$ and $\tilde{u} = \min\{u/u > 0, g(u) = 0\}$ would be a perturbable critical zero, with unstable manifold $v = v^-(u, \mu_\infty) = g(u)$ defined in $[0, \tilde{u}]$. Thus, due to the fact $\mu_\infty = \mu_o$, Proposition 5 prohibits the existence of such a perturbable zero $\tilde{u}$. Therefore, $g(0) = 0$.

Finally, let us analyze the case $u_\infty > 0$ and $\mu_\infty < \mu_o$. First, let us show that $u = u_\infty$ is a critical zero for $f(u)$ (see Definition 3). In fact, $u = u_n$ is a critical zero for every n. In addition, it is clear that $V(u_\infty) = \inf_{n \in \mathbf{N}} V(u_n)$. Thus $V(u) \leq V(u_\infty)$ for $0 \leq u \leq u_\infty$. We claim that $V(u) < V(u_\infty)$ for $0 \leq u < u_\infty$. Otherwise, a point $\tilde{u} \in (0, u_\infty)$ with $V(\tilde{u}) = V(u_\infty)$ would exist. If, as in the preceding paragraphs, we consider the sequence $g_n(u)_{|[\tilde{u}, u_\infty]} = v^-(u, \mu_n)_{|[\tilde{u}, u_\infty]}$, i.e., restricted to the interval $[\tilde{u}, u_\infty]$, and argue in a similar way we arrive again at a contradiction. Thus, $V(u) < V(u_\infty)$ for $0 \leq u < u_\infty$, i.e., u_∞ is a critical zero. Now, consider again the sequence $g_n(u)$ but defined in the whole interval $[0, u_\infty]$, and set again $g(u) = \lim g_n(u)$ in that interval. We claim that $g(0) < 0$. Otherwise, $C(\mu_\infty) = \{(u,v)/0 \leq u \leq u_\infty, v = g(u)\}$ would define a CW which connects the zeros $u = u_\infty$ to $u = 0$ with velocity μ_∞. Thus, no point $u \in (0, u_\infty)$ could satisfy the condition $(\mathrm{S})_\mu$ for $\mu \geq \mu_\infty$, i.e., $\mathcal{S}_\mu(u_\infty) = \emptyset$ for $\mu \geq \mu_\infty$. However, from the construction of the sequences $\{u_n\}$ and $\{\mu_n\}$ and Proposition 5 it follows that the condition $(\mathrm{S})_\mu$, with $\mu \geq \mu_\infty$ can only be satisfied by points $u \in (0, u_\infty)$. Therefore, no point in the interval $(0, u_M)$ can satisfy $(\mathrm{S})_\mu$ for $\mu \geq \mu_\infty$, which means that $\mu_\infty \geq \mu_o$ (see (3.2)), which is a contradiction. Thus, $g(0) < 0$.

Since $g(0) < 0$ then $\tilde{u}_1 = \min\{u > 0/g(u) = 0\}$ is a positive and perturbable critical zero for $f(u)$ with unstable manifold $v^-(u, \mu_\infty) = g(u)$, $0 \leq u \leq \tilde{u}_1$. Moreover, $C(\mu_\infty) = \{(u,v)/\tilde{u}_1 \leq u \leq u_\infty, v = g(u)\}$ defines a CW which connects u_∞ to $\tilde{u}_1$ with velocity $c = \mu_\infty$.

On the other hand, since $\tilde{u}_1$ is perturbable then the argument developed at the beginning of the proof can be repeated again, taking now $u = \tilde{u}_1$ as the starting point instead of u_M. Thus, a decreasing sequence of perturbable critical zeros $\{u_n^1\}$ is generated again, together with an increasing sequence of values of μ, $\{\mu_n^1\}$, with $\mu_\infty < \mu_n^1$ for each $n \in \mathbf{N}$. In addition, there exists a sequence of CWs, $\{C(\mu_n^1)\}$, which connect $u = u_n^1$ to $u = u_{n+1}^1$ with velocity $c = \mu_n^1$. If $\{u_n^1\}$ is not finite, i.e., $u_n^1 > 0$ for every $n \in \mathbf{N}$ (so option (i) in Theorem 3 does not hold in the first approximation), then we arrive at a new value $\mu_\infty^1 = \lim \mu_n^1$, $\mu_\infty^1 \leq \mu_o$, and a new zero $u_\infty^1 = \lim u_n^1$, $0 \leq u_\infty^1 < \tilde{u}_1$.

In this stage of the proof we can obtain the conclusions of Theorem 3 in two ways, namely, if $u_\infty^1 = 0$ or provided that $u_\infty^1 > 0$ and $\mu_\infty^1 = \mu_o$. If neither of those situations occurs, i.e., $u_\infty^1 > 0$ and $\mu_\infty^1 < \mu_o$, a positive perturbable zero $0 < \tilde{u}_2 < u_\infty^1$ is obtained. The proceeding can be started again at $\tilde{u}_2$ to obtain new sequences of positive perturbable critical zeros $\{u_n^2\}$, μ values $\{\mu_n^2\}$, and CWs $\{C(\mu_n^2)\}$. If neither point (i) nor (ii) holds we again get a new perturbable critical zero $0 < \tilde{u}_3 < u_\infty^2$.

Finally, assume that the proceeding is repeated infinitely many times without achieving the points (i) and (ii) in Theorem 3. That means that two sequences $\{u_\infty^n\}$,

$\{\mu_\infty^n\}$ will have been generated and $0 < u_\infty^n$ and $\mu_\infty^n < \mu_o$ for each $n \in \mathbf{N}$. However, it is implicit in the discussion developed above that $\lim u_\infty^n = 0$. Therefore, by using Proposition 5 as above (see the case $u_\infty = 0$ implies $\mu_\infty = \mu_o$) we get $\lim \mu_\infty^n = \mu_o$. Therefore, we can construct a decreasing sequence of perturbable zeros $\{u_{n_k}^k\}$, $\lim u_{n_k}^k = 0$, with associated μ values $\{\mu_{n_k}^k\}$, $\lim \mu_{n_k}^k = \mu_o$, and a sequence of CWs $\{C(\mu_{n_k}^k)\}$, such that $C(\mu_{n_k}^k)$ connects the zero $u_{n_k}^k$ to the zero $u_{n_k+1}^k$ with velocity $c = \mu_{n_k}^k$. This implies that point (ii) in Theorem 3 holds. Thus, the proof of Theorem 3 is completed. □

Remark 6.2. Observe that the unique possible option in Theorem 3 is (i) when $u = 0$ is an isolated zero.

Proof of Theorem 4. To prove the point (i) let us consider two CWs, C_1 and C_2 of equation (1.3), which connect the zeros $0 < u_1 < u_2 \leq u_M$ of $f(u)$ with velocity c. Then, there exist nonpositive functions $v = g_i(u)$, $g_i \in C([u_1, u_2])$ such that $C_i = \{(u,v)/u_1 \leq u \leq u_2\}$, $i = 1, 2$ (see Definition 2).

On the other hand, observe that the functions $g_i(u)$ are C^1 and solve the equation $(5.1)_c$ in $\{u/g_i(u) < 0\}$ (see the proof of Proposition 5). If D designates the set $D = \{u/g_1(u) \neq g_2(u)\}$ and D is nonempty, let (a, b) be some arbitrary connected piece of D. By the uniqueness of solutions to $(5.1)_c$ we arrive at $g_i(a) = g_i(b) = 0$, $i = 1, 2$. Assume, for instance, that $g_1(u) < g_2(u)$ for $u \in (a, b)$. Then,

$$\Delta = \{(u,v)/g_1(u) \leq v \leq g_2(u)\}$$

is invariant for $(3.3)_c$. That is not compatible with the dissipativeness of $(3.3)_c$ when $c > 0$. If $c = 0$ it is straightforward to show the equality $C_1 = C_2$. Thus, $C_1 = C_2$ in any case.

As for the point (ii) let us write $C(c) = \{(u,v)/u_-(c) \leq u \leq u_+(c), v = g_1(u)\}$ and $C(c') = \{(u,v)/u_-(c') \leq u \leq u_+(c'), v = g_2(u)\}$. The assertion of point (ii) is equivalent to showing that the set $\{u/g_2(u) < g_1(u), u_-(c) < u < u_+(c')\}$ is empty. First, observe that $\frac{dg_1}{du}(u_o) < \frac{dg_2}{du}(u_o)$ provided that C and C' meet at (u_o, v_o). Let (a, b) be a connected piece of that set. Since the function $g(u, \mu)$ (see Definition 1) is increasing in μ, the equality $g_1(b) = g_2(b) = 0$ is excluded. On the other hand, if $g_1(b) = g_2(b) < 0$ then the preceding observation implies $\frac{dg_1}{du}(b) > \frac{dg_2}{du}(b)$, which is not possible. Therefore, that set is empty and the proof is concluded. □

Remark 6.3. Consider the equation (3.3) with the nonlinearity $f(u) = \epsilon f_o(u) + f_o(u-1)$, where $f_o(u)$ is as in Example 5.1 and $0 < \epsilon < 1$. A careful analysis of this example reveals that the conclusion of point (ii) in Theorem 4 is false when $u_+(c) < u_+(c')$.

REFERENCES

[A-A] J. C. ALEXANDER AND G. AUCHMUTY, *Bifurcation analysis of reaction-diffusion systems equations* IV, SIAM J. Math. Anal., 19 (1988), pp. 100–109.

[Ab-R] A. AMBROSETTI AND P. H. RABINOWITZ, *Dual variational methods in critical point theory and applications*, J. Funct. Anal., 14 (1973), pp. 349–381.

[Am] H. AMANN, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.

[Ar-W] D. ARONSON AND H. F. WEIMBERGER, *Nonlinear diffusion in population genetics, combustion and nerve propagation*, in Partial Differential Equations and Related Topics, J. A. Goldstein, ed., Lecture Notes in Math. 446, Springer-Verlag, Berlin, 1975.

[Ar-W] ———, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. in

Math., 30 (1978), pp. 33–76.

[B] J. BANKS, *Modelling and control in the biomedical sciences*, in Lecture Notes in Biomath. 6, Springer-Verlag, Berlin, 1975.

[Ba-P] L. BAO-PING AND C. V. PAO, *Almost periodic plane wave solutions for reaction-diffusion systems*, J. Math Anal. Appl., 105 (1985), pp. 231–349.

[Be] H. BERESTYCKI, *Le nombre de solutions de certains problèmes semi-linéaires elliptiques*, J. Funct. Anal., 40 (1981), pp. 1–29.

[Be-L,80] H. BERESTYCKI AND P. L. LIONS, *Some applications of the method of super and subsolutions*, in Bifurcation and Nonlinear Eigenvalue Problems, Lecture Notes in Physics 782, Springer-Verlag, Berlin, 1980, pp. 16–41.

[Be-L,83] ———, *Nonlinear scalar field equations, I. Existence of a ground state*, Arch. Rational Mech. Anal., 82 (1983), pp. 313–345.

[C-H] S. N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory*, Springer-Verlag, Berlin, 1982.

[Ca-M-S] A. CALSINA, X. MORA, AND J. SOLÀ -MORALES, *The dynamical approach to elliptic problems in cylindrical domains, and a study of their parabolic singular limit*, J. Differential Equations, 102 (1993), pp. 244–304.

[Co] C. CONLEY, *Isolated Invariant Sets and the Morse Index*, CBMS Regional Conf. Series in Math., vol. 38, American Mathematical Society, Providence, RI, 1978.

[Co-H-M] D. S. COHEN, F. C. HOPPENSTEADT, R. M. MIURA, *Slowly modulated oscillations in nonlinear diffusion processes*, SIAM J. Appl. Math., 33 (1977), pp. 217–229.

[Cp] W. A. COPPEL, *Stability and Asymptotic Behavior of Differential Equations*, D. C. Heath and Company, Boston, 1965.

[Do] M. DO CARMO, *Differential Geometry of Curves and Surfaces*, Prentice–Hall, Englewood Cliffs, NJ, 1976.

[Du] S. DUNBAR, *Travelling waves in diffusive predator–prey equations: Periodic orbits and point-to-periodic heteroclinic orbits*, SIAM J. Appl. Math., 46 (1986), pp. 1057–1078.

[F] P. C. FIFE, *Mathematical aspects of reacting and diffusing systems*, in Lecture Notes in Biomaths. 28, Springer-Verlag, Berlin, 1979.

[F-M,77] P. C. FIFE AND J. B. MCLEOD, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 334–361.

[F-M,81] ———, *A phase plane discussion of convergence to travelling waves for nonlinear diffusion*, Arch. Rational Mech. Anal., 75 (1981), pp. 281–314.

[Fr-S,84] J. FRAILE AND J. SABINA, *Boundary value conditions for wave fronts in reaction-diffusion systems*, Proc. Roy. Soc. Edinburgh Sect. A, 99 (1984), pp. 127–136.

[Fr-S,89] ———, *General conditions for the existence of a critical point-periodic wave front connection for reaction-diffusion systems*, Nonlinear Anal., 13 (1989), pp. 767–786.

[G] R. GARDNER, *Existence of multidimensional travelling wave solutions of an initial-boundary value problem*, J. Differential Equations, 52 (1986), pp. 291–328.

[He] D. HENRY, *Geometric theory of semilinear parabolic equations*, in Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1981.

[Kl] W. KLINGENBERG, *Eine Vorlesung Über Differentialgeometrie*, Springer-Verlag, Berlin, Heidelberg, 1973.

[Ko-H,73] N. KOPELL AND L. N. HOWARD, *Plane wave solutions to reaction-diffusion equations*, Stud. Appl. Math., 52 (1973), pp. 291–328.

[Ko-H,75] ———, *Bifurcations and trajectories joining critical points*, Adv. in Math., 18 (1975), pp. 306–358.

[L] P. L. LIONS, *On the existence of positive solutions of semilinear elliptic equations*, SIAM Rev., 24 (1982), pp. 441–467.

[Li] LIÑÁN A., Personal communication, 1991.

[M] J. D. MURRAY, *Lectures on Nonlinear Differential Equation Models in Biology*, Clarendon Press, Oxford, 1977.

[N-N-B] C. NITSCHE, J. NITSCHE, AND H. BRENNER, *Existence, uniqueness and regularity of a time-periodic probability density distribution, arising in a sedimentation-diffusion problem*, SIAM J. Math. Anal., 19 (1988), pp. 153–166.

[P] S. I. POHOZAEV, *Eigenfunctions of the equation $\Delta u + \lambda f(u) = 0$*, Soviet Math. Dokl., 5 (1965), pp. 1408–1411.

[R] P. H. RABINOWITZ, *Pairs of positive solutions of nonlinear elliptic partial differential* *equations*, Indiana Univ. Math. J., 23 (1973/74), pp. 173–186.

[S] J. SABINA, *Directional wave fronts of reaction-diffusion systems*, Bull. Australian Math. Soc., 33 (1986), pp. 1–20.

[S-F,87] J. SABINA AND J. FRAILE, *Directional wave fronts in reaction-diffusion systems: existence and asymptotic behaviour*, in Contributions in Nonlinear Partial Differential Equations, vol. II, P. L. Lions and J. I. Díaz, eds., Pitman Res. Notes in Math. Ser. 155, Longman, London, 1987, pp. 232–248.

[S-F,89] ———, *Qualitative properties of a nonlinear diffusion equation with periodic convection*, Istituto per le Applicazioni del Calcolo "Mauro Picone," Quaderno no. 27/1989, Rome, 1989, pp. 163–178.

[Sh-A-S] M. SHARAN, A. AMINATACI, AND M. SINGH, *A numerical study of nonsteady transport of gases in the pulmonary capillarities*, J. Math. Biol., 25 (1987), pp. 433–452.

[Sm] J. SMOLLER, *Shock Waves and Reaction-Diffusion Systems*, Springer-Verlag, Berlin, 1983.

[Sm-W,81] J. SMOLLER AND A. WASSERMAN, *Global bifurcations of steady-state solutions*, J. Differential Equations, 39 (1981), pp. 269–290.

[Sm-W,86] ———, *An existence theorem for positive solutions of semilinear elliptic equations*, Arch. Rational Mech. Anal., 95 (1986), pp. 211–216.

[Sm-W,87] ———, *Existence of positive solutions for semilinear elliptic equations in general domains*, Arch. Rational Mech. Anal., 98 (1987), pp. 229–250.

[St] W. STRAUSS, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys., 55 (1977), pp. 149–162.

[V] VEGA DE PRADA J. M., I. E. PARRA, *Multiple solutions of some semilinear elliptic equations in slender cylindrical domains*, J. Differential Equations, 100 (1992), pp. 225–256.

[W] J. WLOKA, *Partial Differential Equations*, Cambridge University Press, Cambridge, 1987.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 204–220, January 1996

011

ANALYSIS OF THE DOMAIN INTEGRAL OPERATOR FOR ANISOTROPIC DIELECTRIC WAVEGUIDES*

H. P. URBACH†

Abstract. The domain integral equation for guided electromagnetic waves in anisotropic inhomogeneous guides is formulated. The spectral properties of the noncompact integral operator are analysed and the existence of guided modes is proved for lossless nonmagnetic media.

Key words. dielectric waveguide, anisotropic, guided wave, domain integral operator, semi-Fredholm

AMS subject classifications. 78A50, 47G10, 47A75, 47A53, 45C05

1. Introduction.

The determination of the propagation constants and field distributions of guided electromagnetic waves in open cylindrical dielectric waveguides is important in telecommunications technology and in integrated optics. In this paper we consider guides consisting of lossless, nonmagnetic, generally anisotropic and inhomogeneous materials.

Several equivalent mathematical formulations of the guided wave problem can be derived from Maxwell's equations by eliminating different components of the electromagnetic field. Neither of these formulations is an eigenvalue problem of standard type. For instance, by eliminating the axial field components one obtains an eigenvalue problem with the propagation constant as eigenvalue, but the operator in this formulation is not normal. On the other hand, in the case of lossless materials formulations exist in which the operators are self-adjoint, but then the propagation constant is not an eigenvalue but enters the problem in a more complicated way.

In spite of the fact that there exists a vast amount of literature on computational methods for dielectric waveguides, the existence of guided modes in waveguides of arbitrary cross-section has been proved only recently. By analysing the partial differential equations for the magnetic field, Bamberger and Bonnet [4] proved that for lossless isotropic media there exist at least two linearly independent guided modes.

In this paper we study the so-called domain integral equation for the electric field. This equation is obtained by applying the method of Green functions to the vector Helmholtz equation for the electric field. The propagation constant appears as parameter in the kernel of the integral operator. The determination of guided waves is then equivalent to tuning the propagation constant such that -1 is eigenvalue of the integral operator.

Guided waves are often computed by applying the finite element method directly to the partial differential equations (see, e.g., the reviews of Saad [9] and Rahman [8]). However, the domain integral equation is in many cases more suitable for numerical computations (Bagby [2], Baken [3], Pichot [7]) and therefore it is appropriate to study the integral operator mathematically.

The operator occurring in the domain integral formulation of the waveguide problem is an interesting example of an integral operator which is symmetric and bounded but not compact. By using a characterisation of semi-Fredholm operators proved in Schechter [10], we show that the nonpositive part of the spectrum of this operator

* Received by the editors December 2, 1992; accepted for publication (in revised form) April 25, 1994.

† Philips Research Laboratory, Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands (urbach@natlab.research.philips.nl).

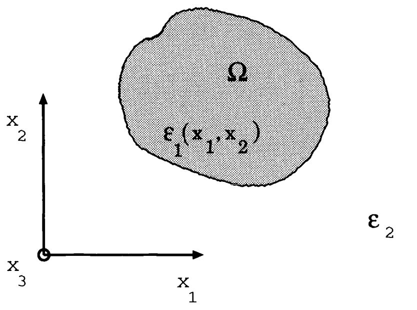


FIG. 2.1. *Cross-section* Ω *of a waveguide in a plane* $x_3 =$ *constant.*

consists of a countable set of negative eigenvalues of finite multiplicity with 0 as unique accumulation point. By applying the mini–max principle to the negative eigenvalues the existence of at least two linearly independent guided modes that propagate in the same axial direction is obtained.

2. The domain integral formulation. Let Ω be the bounded open cross-section of a cylindrical waveguide that is parallel to the x_3-axis of a Cartesian coordinate system (x_1, x_2, x_3) (see Fig. 2.1). Ω may be disconnected, in which case several waveguides with parallel axes are present. The boundary $\partial\Omega$ is assumed Lipschitz.

All materials are nonmagnetic with magnetic permeability μ_0. The guide consists of a lossless, in general inhomogeneous and anisotropic dielectric. For given frequency ω the electric permittivity in point (x_1, x_2) of Ω is given by a positive definite hermitian tensor $\epsilon_1(x_1, x_2)$ of rank 2, of which the components are bounded measurable functions on Ω. We introduce the numbers $\epsilon_{1,\max}$ and $\epsilon_{1,\min}$ by

$$\epsilon_{1,\max} = \sup_{(x_1,x_2)\in\Omega} \max_{\mathbf{V}\in\mathbf{C}^3\setminus\{0\}} \frac{\epsilon_1(x_1,x_2)\mathbf{V}\cdot\bar{\mathbf{V}}}{\mathbf{V}\cdot\bar{\mathbf{V}}}, \tag{2.1}$$

$$\epsilon_{1,\min} = \inf_{(x_1,x_2)\in\Omega} \min_{\mathbf{V}\in\mathbf{C}^3\setminus\{0\}} \frac{\epsilon_1(x_1,x_2)\mathbf{V}\cdot\bar{\mathbf{V}}}{\mathbf{V}\cdot\bar{\mathbf{V}}}. \tag{2.2}$$

The bar denotes complex conjugation and bold letters are used for vectors in $\mathbf{C}^3$.

The exterior of the guide consists of a single homogenous isotropic dielectric with permittivity ϵ_2. It is assumed that

$$\epsilon_{1,\min} \geq \epsilon_2, \tag{2.3}$$

and that for some open set $\tilde{\Omega} \subset \Omega$

$$\inf_{(x_1,x_2)\in\tilde{\Omega}} \min_{\mathbf{V}\in\mathbf{C}^3\setminus\{0\}} \frac{\epsilon_1(x_1,x_2)\mathbf{V}\cdot\bar{\mathbf{V}}}{\mathbf{V}\cdot\bar{\mathbf{V}}} > \epsilon_2. \tag{2.4}$$

In some cases (2.4) holds with $\tilde{\Omega} = \Omega$ and then (2.3) and (2.4) are equivalent to $\epsilon_{1,\min} > \epsilon_2$. However, in other applications the permittivity in the guide varies continuously from a value strictly larger than ϵ_2 to the value ϵ_2 on the boundary of the guide. This is an example for which $\epsilon_{1,\min} = \epsilon_2$ and, to include it in the analysis, we use the assumptions (2.3) and (2.4) instead of $\epsilon_{1,\min} > \epsilon_2$.

Let the permittivity tensor ϵ on $\mathbf{R}^2$ be defined by

$$\epsilon(x_1,x_2)=\begin{cases}\epsilon_1(x_1,x_2) \text{ for } (x_1,x_2)\in\Omega,\\ \\ \epsilon_2 I \text{ for } (x_1,x_2)\in\mathbf{R}^2\backslash\Omega.\end{cases} \tag{2.5}$$

In what follows we will use the symbol ϵ_2 for both the number introduced above and the tensor $\epsilon_2 I$. The meaning will be clear from the context.

For given time frequency ω we seek time harmonic solutions of the source-free Maxwell equations of the form

$$\begin{aligned}\mathcal{E}(x_1,x_2,x_3,t)&=\mathrm{Re}\{\mathbf{E}(x_1,x_2)e^{i(\beta x_3-\omega t)}\},\\ \mathcal{H}(x_1,x_2,x_3,t)&=\mathrm{Re}\{\mathbf{H}(x_1,x_2)e^{i(\beta x_3-\omega t)}\},\end{aligned} \tag{2.6}$$

for some β, where $\mathcal{E}$ and $\mathcal{H}$ are the electric and the magnetic fields, respectively. Substitution of (2.6) into Maxwell's equations yields, after elimination of the magnetic field,

$$\omega^2\mu_0\epsilon\mathbf{E}-\ \mathrm{curl}_\beta\ \mathrm{curl}_\beta\mathbf{E}=\mathbf{0}, \tag{2.7}$$

where curl_β is the differential operator obtained from the curl by replacing $\partial/\partial x_3$ by multiplication by $i\beta$. Hence, with respect to the Cartesian coordinates (x_1,x_2,x_3)

$$\mathrm{curl}_\beta\mathbf{E}=\begin{pmatrix}\frac{\partial E_3}{\partial x_2}-i\beta E_2\\ -\frac{\partial E_3}{\partial x_1}+i\beta E_1\\ \frac{\partial E_2}{\partial x_1}-\frac{\partial E_1}{\partial x_2}\end{pmatrix}. \tag{2.8}$$

The operator in the left-hand side of (2.7) will be denoted by A_β^ϵ:

$$A_\beta^\epsilon(\mathbf{E})=\omega^2\mu_0\epsilon\mathbf{E}-\ \mathrm{curl}_\beta\ \mathrm{curl}_\beta\mathbf{E}. \tag{2.9}$$

It will be considered as operator in $L^2(\mathbf{R}^2)^3$ with domain

$$D(A_\beta^\epsilon)=\{\mathbf{E}\in L^2(\mathbf{R}^2)^3;\ \mathrm{curl}_\beta\ \mathrm{curl}_\beta\mathbf{E}\in L^2(\mathbf{R}^2)^3\}. \tag{2.10}$$

It is easy to see that A_β^ϵ is a densely defined closed operator with $A_{\bar\beta}^{\bar\epsilon}$ as adjoint. In particular, for real β the operator A_β^ϵ is self-adjoint.

We conclude from (2.7) that the problem of determining all time harmonic electromagnetic waves of the form (2.6) is equivalent to determining β such that 0 is in the spectrum of A_β^ϵ. For isotropic guides it is well known [6] that for real β with $-\omega(\mu_0\epsilon_2)^{1/2}\le\beta\le\omega(\mu_0\epsilon_2)^{1/2}$, there exist nontrivial solutions of (2.7) (see Fig. 2.2). The fields $\mathbf{E}(x_1,x_2)$ and $\mathbf{H}(x_1,x_2)$ of these so-called propagating radiation modes are not in $L^2(\mathbf{R}^2)^3$. In addition there exist for all purely imaginary β solutions of (2.7) of which the fields also are not in $L^2(\mathbf{R}^2)^3$. These waves are called evanescent radiation modes. Finally, for a finite number of real β with $\omega^2\mu_0\epsilon_2<\beta^2<\omega^2\mu_0\epsilon_{1,\max}$ there exist solutions of (2.7) that are in $D(A_\beta^\epsilon)$. These solutions propagate a finite amount of energy in the x_3-direction and this energy is confined to the waveguide and its immediate surrounding. Therefore, these modes are called guided modes. The existence of guided modes was proved for isotropic guides in [4]. For $\beta^2>\omega^2\mu_0\epsilon_{1,\max}$ the operator A_β^ϵ can be shown to have a bounded inverse, hence for these β, (2.7) does not have nonzero solutions.

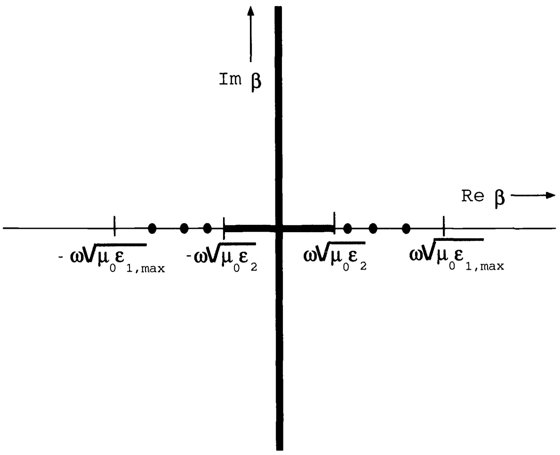


FIG. 2.2. *The set of β's corresponding to propagating radiation modes, evanescent radiation modes, and guided modes.*

It is anticipated (Marcuse [6]) that the propagating radiation modes, evanescent radiation modes, and guided modes together form a complete system. For slab waveguides the completeness of the modes is trivial because the waveguide problem can be formulated as an eigenvalue problem for a self-adjoint operator. In the case of waveguides of bounded cross-section, however, this is not possible and it seems that completeness has not been proved rigorously.

In this paper we study the guided modes and we therefore always assume that β is real and satisfies

$$\beta^2 > \omega^2\mu_0\epsilon_2. \tag{2.11}$$

The other inequality satisfied by propagation constants of guided modes, $\beta^2 < \omega^2\mu_0\epsilon_{1,\max}$, will follow from the analysis.

We use the domain integral formulation of the problem. To derive this we first write (2.7) as

$$A_\beta^{\epsilon_2}(\mathbf{E}) = -\omega^2\mu_0(\epsilon - \epsilon_2)\mathbf{E}. \tag{2.12}$$

The operator $A_\beta^{\epsilon_2}$ has constant coefficients and can thus be analysed using the Fourier transform. Let $\mathcal{F}(\mathbf{E})$ be defined by

$$\mathcal{F}(\mathbf{E})(\xi_1,\xi_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp[-2\pi i(\xi_1 x_1 + \xi_2 x_2)]\mathbf{E}(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2. \tag{2.13}$$

By applying the Fourier transform to $A_\beta^{\epsilon_2}(\mathbf{E}) = \mathbf{G}$, one obtains the matrix multiplication

$$\widehat{A_\beta^{\epsilon_2}}(\xi_1,\xi_2)\mathcal{F}(\mathbf{E})(\xi_1,\xi_2) = \mathcal{F}(\mathbf{G})(\xi_1,\xi_2), \tag{2.14}$$

where $\widehat{A_\beta^{\epsilon_2}}(\xi_1,\xi_2)$ is the matrix

$$\widehat{A_\beta^{\epsilon_2}}(\xi_1,\xi_2) = \begin{pmatrix} \omega^2\mu_0\epsilon_2-4\pi^2\xi_2^2-\beta^2 & 4\pi^2\xi_1\xi_2 & 2\pi\xi_1\beta \\ 4\pi^2\xi_1\xi_2 & \omega^2\mu_0\epsilon_2-4\pi^2\xi_1^2-\beta^2 & 2\pi\xi_2\beta \\ 2\pi\xi_1\beta & 2\pi\xi_2\beta & \omega^2\mu_0\epsilon_2-4\pi^2(\xi_1^2+\xi_2^2) \end{pmatrix}. \tag{2.15}$$

Its determinant is given by

$$D(\xi_1,\xi_2) = \omega^2\mu_0\epsilon_2\{4\pi^2(\xi_1^2+\xi_2^2)+\beta^2-\omega^2\mu_0\epsilon_2\}^2, \tag{2.16}$$

and since $\beta^2 > \omega^2\mu_0\epsilon_2$ it follows that the matrices $\widehat{A_\beta^{\epsilon_2}}(\xi_1,\xi_2)$ are invertible for all (ξ_1,ξ_2). The inverse is

$$\widehat{A_\beta^{\epsilon_2}}(\xi_1,\xi_2)^{-1} = \frac{1}{[\omega^2\mu_0\epsilon_2\, D(\xi_1,\xi_2)]^{1/2}} \begin{pmatrix} 4\pi^2\xi_1^2-\omega^2\mu_0\epsilon_2 & 4\pi^2\xi_1\xi_2 & 2\pi\xi_1\beta \\ 4\pi^2\xi_1\xi_2 & 4\pi^2\xi_2^2-\omega^2\mu_0\epsilon_2 & 2\pi\xi_2\beta \\ 2\pi\xi_1\beta & 2\pi\xi_2\beta & \beta^2-\omega^2\mu_0\epsilon_2 \end{pmatrix}. \tag{2.17}$$

The inverse of the operator $A_\beta^{\epsilon_2}$ is then

$$\left(A_\beta^{\epsilon_2}\right)^{-1}(\mathbf{G}) = \mathcal{F}^{-1}\circ\left(\widehat{A_\beta^{\epsilon_2}}\right)^{-1}\circ\mathcal{F}(\mathbf{G}). \tag{2.18}$$

This is a bounded linear operator $L^2(\mathbf{R}^2)^3 \mapsto L^2(\mathbf{R}^2)^3$ because it is in the Fourier domain given by multiplication by the bounded matrix-valued function $(\xi_1,\xi_2)\mapsto \widehat{A_\beta^{\epsilon_2}}(\xi_1,\xi_2)^{-1}$. Using (2.18) it follows that (2.12) is equivalent to

$$-\mathbf{E} = \omega^2\mu_0\left(A_\beta^{\epsilon_2}\right)^{-1}[(\epsilon-\epsilon_2)\mathbf{E}]. \tag{2.19}$$

Now by assumption (2.3), the tensor $\epsilon(x_1,x_2)-\epsilon_2$ is positive semidefinite for all (x_1,x_2), hence its square root is well defined. We define the field

$$\mathbf{F} = (\epsilon-\epsilon_2)^{1/2}\mathbf{E}, \tag{2.20}$$

which has support contained in $\bar{\Omega}$, and we introduce the operator $T_\beta^\epsilon : L^2(\Omega)^3 \mapsto L^2(\Omega)^3$ by

$$T_\beta^\epsilon(\mathbf{F}) = \omega^2\mu_0(\epsilon-\epsilon_2)^{1/2}\circ\left(A_\beta^{\epsilon_2}\right)^{-1}\circ(\epsilon-\epsilon_2)^{1/2}\mathbf{F}, \tag{2.21}$$

where $\mathbf{F}$ is identified with the field obtained by extension of $\mathbf{F}$ to $\mathbf{R}^2$ by setting it equal to zero in the complement of Ω. Then (2.19) implies

$$-\mathbf{F} = T_\beta^\epsilon(\mathbf{F}), \quad \text{on } \Omega. \tag{2.22}$$

Hence, if $\mathbf{E}$ is the electric field of a guided mode with propagation constant β, then $\mathbf{F}$ defined by (2.20) is eigenfield of the operator T_β^ϵ with eigenvalue -1. Conversely, if $\mathbf{F}\in L^2(\Omega)^3$ is eigenfield of T_β^ϵ with eigenvalue -1, then $\mathbf{E}$ given by

$$\mathbf{E} = -\omega^2\mu_0\left(A_\beta^{\epsilon_2}\right)^{-1}\left[(\epsilon-\epsilon_2)^{1/2}\mathbf{F}\right] \tag{2.23}$$

is the electric field of a guided mode with propagation constant β. Note that for real β, T_β^ϵ is a symmetric operator. The operator in the right-hand side of (2.19) is not symmetric and this is the reason for introducing T_β^ϵ.

By computing the Fourier transforms (2.18), $(A_\beta^{\epsilon_2})^{-1}$ can be written as a matrix of convolution operators in $L^2(\Omega)$ with highly singular kernels that are derivatives with respect to x_1 and x_2 of the function $K_0\left((\beta^2-\omega^2\mu_0\epsilon_2)^{1/2}(x_1^2+x_2^2)^{1/2}\right)$, where K_0 is the modified Bessel function of order 0. We will use expression (2.18) in terms of the Fourier transforms, however.

We conclude that the determination of guided modes is equivalent to finding values for β such that -1 is eigenvalue of the operator T_β^ϵ. The fact that the domain integral equation is an equation for the field $\mathbf{F}$ with support in $\bar{\Omega}$ and is an equation to be satisfied on Ω makes it very suitable for numerical computations. In §§3 and 4 we shall derive properties of the operator T_β^ϵ and its spectrum.

3. Lack of compactness. T_β^ϵ is a bounded symmetric operator in $L^2(\Omega)^3$ but it is not compact. If T_β^ϵ were compact then the quadratic functional

$$\mathbf{F} \mapsto \left(T_\beta^\epsilon(\mathbf{F}), \mathbf{F}\right) \tag{3.1}$$

would be continuous for weakly converging sequences in $L^2(\Omega)^3$. We shall show that the functional does not have this property. The argument leading to this conclusion applies to all permittivity tensors and all cross-sections that were introduced in §2. Hence, the operator T_β^ϵ is never compact, not even when the waveguide has smooth permittivity tensor and smooth boundary.

Let

$$\operatorname{div}_\beta \mathbf{F} = \frac{\partial F_1}{\partial x_1} + \frac{\partial F_2}{\partial x_2} + i\beta F_3. \tag{3.2}$$

Its Fourier transform satisfies

$$\mathcal{F}\left(\operatorname{div}_\beta \mathbf{F}\right) = 2\pi i \xi_1 \mathcal{F}(F_1) + 2\pi i \xi_2 \mathcal{F}(F_2) + i\beta \mathcal{F}(F_3). \tag{3.3}$$

By using (2.17), (2.21), and (3.3) one can derive

$$\begin{aligned} \left(T_\beta^\epsilon(\mathbf{F}), \mathbf{F}\right) &= \omega^2 \mu_0 \left(\left(\widehat{A_\beta^{\epsilon_2}}\right)^{-1} \circ \mathcal{F}\left[(\epsilon - \epsilon_2)^{1/2} \mathbf{F}\right], \mathcal{F}\left[(\epsilon - \epsilon_2)^{1/2}\mathbf{F}\right] \right) \\ &= p_1(\mathbf{F}) - p_2(\mathbf{F}), \end{aligned} \tag{3.4}$$

where p_1 and p_2 are nonnegative continuous quadratic forms on $L^2(\Omega)^3$ defined by

$$p_1(\mathbf{F}) = \frac{1}{\epsilon_2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left| \mathcal{F}\left[\operatorname{div}_\beta \left((\epsilon - \epsilon_2)^{1/2} \mathbf{F} \right) \right] \right|^2 s(|\xi|)^{-1}\, d\xi_1\, d\xi_2 \tag{3.5}$$

and

$$p_2(\mathbf{F}) = \omega^2 \mu_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left| \mathcal{F}\left[(\epsilon - \epsilon_2)^{1/2} \mathbf{F} \right] \right|^2 s(|\xi|)^{-1}\, d\xi_1\, d\xi_2 \tag{3.6}$$

with

$$s(|\xi|) = 4\pi^2 |\xi|^2 + \beta^2 - \omega^2 \mu_0 \epsilon_2, \qquad |\xi| = (\xi_1^2 + \xi_2^2)^{1/2}. \tag{3.7}$$

When considered as function of $\mathbf{G} \equiv (\epsilon - \epsilon_2)^{1/2}\mathbf{F}$, the right-hand side of (3.6) is the square of a norm which is equivalent to the norm on the space $H^{-1}(\mathbf{R}^2)^3$ (or $W^{-1,2}(\mathbf{R}^2)^3$). Because Ω is a bounded set with Lipschitz boundary, a well-known imbedding theorem (Adams [1]) implies that the inclusion $L^2(\Omega)^3 \hookrightarrow H^{-1}(\mathbf{R}^2)^3$ is compact. Hence, p_2 is continuous for sequences that converge weakly in $L^2(\Omega)^3$.

However, p_1 is not continuous for weakly converging sequences. To show this we assume that (0,0) is in the interior of the set $\tilde{\Omega}$ of property (2.4). Choose R such that the disc with center $(0,0)$ and radius R is contained in $\tilde{\Omega}$. Let (r, φ) be polar coordinates. It follows from assumption (2.4) that for $(x_1^2 + x_2^2)^{1/2} < R$ the matrices $\epsilon(x_1, x_2) - \epsilon_2$ are invertible. For integer n, let $f_n(r)$ be functions with support in $(0, R)$ such that

$$\int_0^R |f_n(r)|^2 r\, dr = 1, \tag{3.8}$$

and

$$\lim_{n\to\infty} f_n = 0 \tag{3.9}$$

weakly in L^2. Let the vector fields $\mathbf{F}_n$ be given by

$$\mathbf{F}_n = (\epsilon - \epsilon_2)^{-1/2} f_n(r)\,\hat{\mathbf{r}}, \tag{3.10}$$

where $\hat{\mathbf{r}}$ is the unit radial vector. Then $\mathbf{F}_n \in L^2(\Omega)^3$ and

$$\lim_{n\to\infty} \mathbf{F}_n = \mathbf{0}, \tag{3.11}$$

weakly in $L^2(\Omega)^3$.

We remark that for an arbitrary function $h(r)$ and integer ν the Fourier transform of the function $h(r)\exp(i\nu\varphi)$ can be written as

$$\mathcal{F}[h(r)\exp(i\nu\varphi)](\varrho,\psi) = 2\pi(-i)^\nu \int_0^\infty J_\nu(2\pi\varrho r)h(r)r\,dr\exp(i\nu\psi), \tag{3.12}$$

where ϱ, ψ are polar coordinates in the Fourier plane and J_ν is the Bessel function of order ν. By using the identity of Parseval it follows from (3.12) that

$$\int_0^\infty \left| \int_0^\infty J_\nu(2\pi\varrho r)h(r)r\,dr \right|^2 \varrho\,d\varrho = \frac{1}{4\pi^2}\int_0^\infty |h(r)|^2 r\,dr. \tag{3.13}$$

Now,

$$\operatorname{div}_\beta(f_n\hat{\mathbf{r}}) = \frac{1}{r}\frac{d}{dr}\left[r f_n(r)\right], \tag{3.14}$$

and by using (3.12) one finds after a partial integration

$$\mathcal{F}\left[\operatorname{div}_\beta(f_n\hat{\mathbf{r}})\right](\varrho,\psi) = 4\pi^2\varrho \int_0^\infty J_1(2\pi\varrho r)f_n(r)r\,dr. \tag{3.15}$$

Hence,

$$\begin{aligned} p_1(\mathbf{F}_n) &= \frac{1}{\epsilon_2}\int_0^\infty\int_0^{2\pi} |\mathcal{F}[\operatorname{div}_\beta(f_n\hat{\mathbf{r}})](\varrho,\psi)|^2 \frac{\varrho}{s(\varrho)}\,d\psi\,d\varrho \\ &= \frac{8\pi^3}{\epsilon_2}\int_0^\infty \left|\int_0^\infty J_1(2\pi\varrho r)f_n(r)r\,dr\right|^2 \frac{4\pi^2\varrho^2}{s(\varrho)}\varrho\,d\varrho \\ &= \frac{8\pi^3}{\epsilon_2}\int_0^\infty \left|\int_0^\infty J_1(2\pi\varrho r)f_n(r)r\,dr\right|^2 \varrho\,d\varrho - q(f_n), \end{aligned} \tag{3.16}$$

where

$$\begin{aligned} q(f_n) &= \frac{8\pi^3}{\epsilon_2}(\beta^2 - \omega^2\mu_0\epsilon_2)\int_0^\infty \left|\int_0^\infty J_1(2\pi\varrho r)f_n(r)r\,dr\right|^2 \frac{\varrho}{s(\varrho)}\,d\varrho \\ &= \frac{(\beta^2-\omega^2\mu_0\epsilon_2)}{\epsilon_2}\int_0^{2\pi}\int_0^\infty |\mathcal{F}[f_n\exp(i\varphi)](\varrho,\psi)|^2 \frac{\varrho}{s(\varrho)}\,d\varrho\,d\psi. \end{aligned} \tag{3.17}$$

In the last equality we used (3.12). Now, as functional of $f_n(r)\exp(i\varphi)$, the right-hand side of (3.17) is the square of a norm that is equivalent to the norm on $H^{-1}(\mathbf{R}^2)$.

Since $\lim_{n\to\infty} f_n(r)\exp(i\varphi) = 0$ weakly in $L^2(\mathbf{R}^2)$ and the supports of these functions are contained in a bounded set, it follows that

$$\lim_{n\to\infty} q(f_n) = 0. \tag{3.18}$$

On the other hand, by using (3.13) with $\nu = 1$, the first term on the right-hand side of (3.16) becomes

$$\frac{8\pi^3}{\epsilon_2}\int_0^\infty \left|\int_0^\infty J_1(2\pi\varrho r) f_n(r) r\,dr\right|^2 \varrho\,d\varrho = \frac{2\pi}{\epsilon_2}\int_0^\infty |f_n(r)|^2 r\,dr = \frac{2\pi}{\epsilon_2}. \tag{3.19}$$

Hence,

$$\lim_{n\to\infty} p_1(\mathbf{F}_n) = \frac{2\pi}{\epsilon_2}. \tag{3.20}$$

This shows that p_1 is not continuous for sequences that converge weakly in $L^2(\Omega)^3$. Hence, T^ϵ_β is not compact.

4. Spectral analysis of the integral operator. We first determine the spectrum of $A^{\epsilon_2}_\beta$ for $\beta^2 > \omega^2\mu_0\epsilon_2$. As mentioned in §2, $A^{\epsilon_2}_\beta$ with domain (2.10) is a self-adjoint operator in $L^2(\mathbf{R}^2)^3$. Let λ be a real number. The operator $A^{\epsilon_2}_\beta - \lambda I$ has constant coefficients and reduces in the Fourier domain to a matrix multiplication with matrices (2.15) in which the number $\omega^2\mu_0\epsilon_2$ is replaced by $\omega^2\mu_0\epsilon_2 - \lambda$. It follows from (2.16) that the determinants of these matrices are

$$(\omega^2\mu_0\epsilon_2 - \lambda)\left[4\pi^2(\xi_1^2+\xi_2^2) + \beta^2 - \omega^2\mu_0\epsilon_2 + \lambda\right]^2. \tag{4.1}$$

The spectrum $\sigma(A^{\epsilon_2}_\beta)$ is the set of all λ such that for some value of $\xi_1^2 + \xi_2^2$ the polynomial (4.1) vanishes. Hence,

$$\sigma(A^{\epsilon_2}_\beta) = (-\infty, \omega^2\mu_0\epsilon_2 - \beta^2] \cup \{\omega^2\mu_0\epsilon_2\}. \tag{4.2}$$

If $\lambda = \omega^2\mu_0\epsilon_2$ then $(A^{\epsilon_2}_\beta - \lambda I)(\mathbf{E}) = \mathbf{0}$ is equivalent to $\mathrm{curl}_\beta\,\mathrm{curl}_\beta\mathbf{E} = \mathbf{0}$. Hence $\lambda = \omega^2\mu_0\epsilon_2$ is eigenvalue with infinite-dimensional eigenspace

$$\mathcal{H}_{\mathrm{grad}_\beta} = \left\{\mathrm{grad}_\beta\psi\ ;\ \psi \in H^1(\mathbf{R}^2)\right\}, \tag{4.3}$$

with grad_β the operator obtained from the gradient by replacing differentiation with respect to x_3 by the multiplication by $i\beta$. On the other hand, for $\lambda \neq \omega^2\mu_0\epsilon_2$, $(A^{\epsilon_2}_\beta - \lambda I)(\mathbf{E}) = \mathbf{0}$ implies that $\mathrm{div}_\beta\mathbf{E} = \mathbf{0}$. Then every component of $\mathbf{E}$ satisfies the scalar Helmholtz equation with wavenumber $k = (\omega^2\mu_0\epsilon_2 - \beta^2 - \lambda)^{1/2}$ in two-dimensional space and the eigenfields are divergence-free vector plane waves. Therefore, the eigenspaces corresponding to the negative part of the spectrum (4.2) together span the space of vector fields that have vanishing divergence

$$\mathcal{H}_{\mathrm{div}_\beta} = \left\{\mathbf{F} \in L^2(\mathbf{R}^2)^3\ ;\ \mathrm{div}_\beta\mathbf{F} = 0\right\}. \tag{4.4}$$

One has the orthogonal decomposition

$$L^2(\mathbf{R}^2)^3 = \mathcal{H}_{\mathrm{grad}_\beta} \oplus \mathcal{H}_{\mathrm{div}_\beta}. \tag{4.5}$$

The spectrum of the inverse of $A_\beta^{\epsilon_2}$ is given by

$$\sigma\left((A_\beta^{\epsilon_2})^{-1}\right) = \left[-\frac{1}{\beta^2 - \omega^2\mu_0\epsilon_2}, 0\right] \cup \left\{\frac{1}{\omega^2\mu_0\epsilon_2}\right\}. \tag{4.6}$$

Hence, the numerical range of $\left(A_\beta^{\epsilon_2}\right)^{-1}$ is the interval

$$\left[-\frac{1}{\beta^2 - \omega^2\mu_0\epsilon_2}, \frac{1}{\omega^2\mu_0\epsilon_2}\right]. \tag{4.7}$$

Since

$$\left(T_\beta^\epsilon(\mathbf{F}), \mathbf{F}\right) = \omega^2\mu_0\left(\left(A_\beta^{\epsilon_2}\right)^{-1}\left[(\epsilon - \epsilon_2)^{1/2}\mathbf{F}\right], (\epsilon - \epsilon_2)^{1/2}\mathbf{F}\right), \tag{4.8}$$

it follows that the numerical range of T_β^ϵ is contained in the interval

$$\left[-\frac{\omega^2\mu_0\epsilon_{1,\max} - \omega^2\mu_0\epsilon_2}{\beta^2 - \omega^2\mu_0\epsilon_2}, \frac{\epsilon_{1,\max} - \epsilon_2}{\epsilon_2}\right]. \tag{4.9}$$

Hence we have

$$\sigma\left(T_\beta^\epsilon\right) \subset \left[-\frac{\omega^2\mu_0\epsilon_{1,\max} - \omega^2\mu_0\epsilon_2}{\beta^2 - \omega^2\mu_0\epsilon_2}, \frac{\epsilon_{1,\max} - \epsilon_2}{\epsilon_2}\right]. \tag{4.10}$$

To proceed with the spectral analysis of T_β^ϵ we shall apply a characterisation of semi-Fredholm operators proved in Schechter [10]. Recall that a bounded operator T is called semi-Fredholm if it has closed range and its kernel has finite dimension. If in addition its cokernel is finite dimensional, then T is called a Fredholm operator. It is clear that for self-adjoint operators there is no distinction between Fredholm and semi-Fredholm operators. For a compact symmetric operator T the operators $T - \lambda I$ are for all nonzero λ Fredholm, but for operators that are not compact this is not true. However, T_β^ϵ has the property that for all negative λ the operator $T_\beta^\epsilon - \lambda I$ is Fredholm. In the proof of this statement we shall use a characterisation of semi-Fredholm operators which we formulate next. We call a seminorm $|.|$ compact with respect to the norm of a Hilbert space $\mathcal{H}$ if every bounded sequence $\{x_n\}$ in $\mathcal{H}$ has a subsequence $\{x_{n_k}\}$ such that $\lim_{k\to\infty}\lim_{l\to\infty}|x_{n_k} - x_{n_l}| = 0$. Then we have the following result (Schechter [10]).

THEOREM 4.1. *Let T be a bounded linear operator in a Hilbert space $\mathcal{H}$. T is semi-Fredholm if and only if there exists a constant $C > 0$ and a seminorm $|.|$ which is compact with respect to the norm on $\mathcal{H}$, such that for all $x \in \mathcal{H}$*

$$\|x\| \le C\|T(x)\| + |x|. \tag{4.11}$$

We apply this theorem in the proof of the following proposition.

PROPOSITION 4.2. *$T_\beta^\epsilon - \lambda I$ is Fredholm for all $\lambda < 0$.*

Proof. We have

$$\begin{aligned}\|T_\beta^\epsilon(\mathbf{F}) - \lambda\mathbf{F}\|^2 &\ge \lambda^2\|\mathbf{F}\|^2 - 2\lambda\left(T_\beta^\epsilon(\mathbf{F}), \mathbf{F}\right)\\ &= \lambda^2\|\mathbf{F}\|^2 - 2\lambda p_1(\mathbf{F}) + 2\lambda p_2(\mathbf{F})\\ &\ge \lambda^2\|\mathbf{F}\|^2 + 2\lambda p_2(\mathbf{F}),\end{aligned} \tag{4.12}$$

where (3.4) and the fact that $p_1 \geq 0$ and $\lambda < 0$ have been used. Hence,

$$\|\mathbf{F}\| \leq \frac{1}{|\lambda|}\|T^{\epsilon}_{\beta}(\mathbf{F}) - \lambda\,\mathbf{F}\| + \left[\frac{2}{|\lambda|}p_2(\mathbf{F})\right]^{1/2}. \tag{4.13}$$

As remarked in §3, p_2 is a quadratic form that is continuous for weakly converging sequences in $L^2(\Omega)^3$.. This implies that the second term on the right-hand side of (4.13) is a seminorm which is compact with respect to the norm of $L^2(\Omega)^3$. Hence, by Theorem 4.1, $T^{\epsilon}_{\beta} - \lambda I$ is Fredholm. □

We conclude therefore that the negative part of the spectrum of T^{ϵ}_{β}, if not empty, consists of eigenvalues with finite-dimensional eigenspaces. In fact there exist infinitely many negative eigenvalues.

PROPOSITION 4.3. *T^{ϵ}_{β} has a countable set of negative eigenvalues.*

Proof. Proposition 4.2 implies that every negative element of $\sigma(T^{\epsilon}_{\beta})$ is eigenvalue with finite-dimensional eigenspace. Using

$$\begin{aligned}\left(T^{\epsilon}_{\beta}(\mathbf{F}),\mathbf{F}\right) &= \omega^2\mu_0\left(\left(\widehat{A^{\epsilon_2}_{\beta}}\right)^{-1}\circ\mathcal{F}\left[(\epsilon-\epsilon_2)^{1/2}\mathbf{F}\right],\mathcal{F}\left[(\epsilon-\epsilon_2)^{1/2}\mathbf{F}\right]\right)\\ &= p_1(\mathbf{F}) - p_2(\mathbf{F}),\end{aligned} \tag{4.14}$$

(3.5), and (3.6) we see that $(T^{\epsilon}_{\beta}(\mathbf{F}),\mathbf{F}) < 0$ for all $\mathbf{F}$ in the set

$$X(\Omega) = \left\{\mathbf{F}\in L^2(\Omega)^3\ ;\ \mathrm{div}_{\beta}[(\epsilon-\epsilon_2)^{1/2}\mathbf{F}] = 0 \text{ on } \Omega,\ (\epsilon-\epsilon_2)^{1/2}\mathbf{F}\neq\mathbf{0}\right\}. \tag{4.15}$$

Using property (2.4) we see that this set has infinite-dimensional span and therefore there must exist a countable set of negative eigenvalues. □

We can now show that the lower bound of the interval (4.10) is not in the spectrum of T^{ϵ}_{β}. If it would be in the spectrum, then it would be an eigenvalue. Let $\mathbf{F}$ be a normalised eigenfield corresponding to this eigenvalue and put

$$\mathbf{G} = \frac{(\epsilon-\epsilon_2)^{1/2}}{(\epsilon_{1,\max}-\epsilon_2)^{1/2}}\mathbf{F}. \tag{4.16}$$

Then $\|\mathbf{G}\| \leq 1$ and

$$\begin{aligned}\left(\left(A^{\epsilon_2}_{\beta}\right)^{-1}(\mathbf{G}),\mathbf{G}\right) &= \frac{1}{\omega^2\mu_0\epsilon_{1,\max}-\omega^2\mu_0\epsilon_2}\left(T^{\epsilon}_{\beta}(\mathbf{F}),\mathbf{F}\right)\\ &= -\frac{1}{\beta^2-\omega^2\mu_0\epsilon_2}.\end{aligned} \tag{4.17}$$

Because $-1/(\beta^2-\omega^2\mu_0\epsilon_2)$ is the lower bound of the numerical range of the operator $(A^{\epsilon_2}_{\beta})^{-1}$ (see (4.7)) and because $\mathbf{G}$ is in $L^2(\mathbf{R}^2)^3$, we conclude that $-1/(\beta^2-\omega^2\mu_0\epsilon_2)$ is eigenvalue of $(A^{\epsilon_2}_{\beta})^{-1}$ with eigenfield $\mathbf{G}$. This contradicts the fact that eigenfields of $(A^{\epsilon_2}_{\beta})^{-1}$ corresponding to the negative part of the spectrum are plane waves, i.e., they are not in $L^2(\mathbf{R}^2)^3$. Hence the lower bound of the interval (4.10) is not in the spectrum of T^{ϵ}_{β}.

PROPOSITION 4.4. *The negative eigenvalues are isolated and* 0 *is the unique nonpositive accumulation point of the spectrum.*

Proof. To prove this we assume that $\lambda < 0$ is accumulation point of $\sigma(T^{\epsilon}_{\beta})$. Let $\{\lambda_n\}$ be a sequence of negative eigenvalues which converge to λ and let $\{\mathbf{F}_n\}$ be a corresponding sequence of normalised eigenfields. Then by using estimate (4.13)

$$
\begin{aligned}
\|\mathbf{F}_n - \mathbf{F}_m\| &\leq \frac{1}{|\lambda|} \| T^{\epsilon}_{\beta}(\mathbf{F}_n - \mathbf{F}_m) - \lambda\,(\mathbf{F}_n - \mathbf{F}_m)\| \\
&\quad + \left[\frac{2}{|\lambda|} p_2(\mathbf{F}_n - \mathbf{F}_m)\right]^{1/2} \\
&\leq \left(|\lambda_n - \lambda| + |\lambda_m - \lambda|\right)/|\lambda| + \left[\frac{2}{|\lambda|} p_2(\mathbf{F}_n - \mathbf{F}_m)\right]^{1/2}. \qquad (4.18)
\end{aligned}
$$

There exists a subsequence, which we assume to be equal to the original sequence, such that $\lim_{n\to\infty} \lim_{m\to\infty} p_2(\mathbf{F}_n - \mathbf{F}_m) = 0$. Then (4.18) implies that $\mathbf{F}_n$ is a Cauchy sequence in $L^2(\Omega)^3$. But this contradicts the fact that $\mathbf{F}_n$ is an orthonormal system in $L^2(\Omega)^3$. This completes the proof. □

We summarise the spectral properties of the operator T^{ϵ}_{β}, with $\beta^2 > \omega^2 \mu_0 \epsilon_2$, in the following theorem.

THEOREM 4.5. *The spectrum of T^{ϵ}_{β} satisfies*

$$
\sigma\left(T^{\epsilon}_{\beta}\right) \subset \left(-\frac{\omega^2 \mu_0 \epsilon_{1,\max} - \omega^2 \mu_0 \epsilon_2}{\beta^2 - \omega^2 \mu_0 \epsilon_2}, \frac{\epsilon_{1,\max} - \epsilon_2}{\epsilon_2}\right]. \qquad (4.19)
$$

The negative part of the spectrum consists of eigenvalues λ_n with finite-dimensional eigenspaces, such that

$$
-\frac{\omega^2 \mu_0 \epsilon_{1,\max} - \omega^2 \mu_0 \epsilon_2}{\beta^2 - \omega^2 \mu_0 \epsilon_2} < \lambda_1 < \lambda_2 < \cdots < \lambda_n < \cdots < 0, \qquad (4.20)
$$

with 0 as unique accumulation point

$$
\lim_{n\to\infty} \lambda_n = 0. \qquad (4.21)
$$

The positive part of the spectrum of T^{ϵ}_{β} is more difficult to characterise than the negative part. The positive spectrum may be (partly) continuous with electric fields that are not in the domain of A^{ϵ}_{β}. In any case, the positive spectrum is not empty and is also not a countable set of eigenvalues with finite-dimensional eigenspaces and with 0 as only possible accumulation point, because then Theorem 4.5 would imply that T^{ϵ}_{β} would be a compact operator. The characterisation of the spectrum of T^{ϵ}_{β} as given in Theorem 4.5 suffices for the determination of the guided modes in the next section.

5. Guided modes. By combining the results of §§2 and 4 we see that the determination of guided modes is equivalent to determining β with $\beta^2 > \omega^2 \mu_0 \epsilon_2$, such that one of the eigenvalues of the operator T^{ϵ}_{β} is -1. In the following the eigenvalues are always numbered in increasing order and counting their multiplicities. To emphasize the dependence on β and on ϵ we denote the nth eigenvalue by $\lambda^{\epsilon}_n(\beta)$.

According to the mini–max principle we have

$$
\lambda^{\epsilon}_n(\beta) = \min_{W_n(\Omega)} \max_{\mathbf{F}\in W_n(\Omega)\setminus\{0\}} R(\mathbf{F}), \qquad (5.1)
$$

where the minimum is taken over all n-dimensional subspaces of $L^2(\Omega)^3$ and where R is the Rayleigh quotient

$$
R(\mathbf{F}) = \left(T^{\epsilon}_{\beta}(\mathbf{F}), \mathbf{F}\right) / (\mathbf{F}, \mathbf{F}). \qquad (5.2)
$$

First we recall the following result.

LEMMA 5.1. *Let A and B be positive definite matrices and let $A - B$ be positive semidefinite. Then $B^{-1} - A^{-1}$ is positive semidefinite.*

Proof. $B^{-1} - A^{-1}$ is positive semidefinite if and only if $I - B^{1/2}A^{-1}B^{1/2}$ is positive semidefinite. Therefore, it is sufficient to show that all eigenvalues of $B^{-1/2}AB^{-1/2}$ are ≥ 1. Let λ and $\mathbf{V} \neq 0$ be such that $B^{-1/2}AB^{-1/2}\mathbf{V} = \lambda\mathbf{V}$. Then $A\mathbf{W} = \lambda B\mathbf{W}$, where $\mathbf{W} = B^{-1/2}\mathbf{V}$. Since $A\mathbf{W}.\bar{\mathbf{W}} \geq B\mathbf{W}.\bar{\mathbf{W}}$ we see that $\lambda \geq 1$. □

Next we prove the following result.

LEMMA 5.2. *Let Ω' be the cross-section of a waveguide that is contained in Ω and let ϵ' be its positive definite permittivity tensor. We assume that ϵ' is identical to ϵ_2 in the exterior of Ω' and that ϵ' has the properties* (2.3) *and* (2.4) *on Ω'. Furthermore, we assume*

$$\epsilon'(x_1, x_2)\mathbf{V}.\bar{\mathbf{V}} \leq \epsilon(x_1, x_2)\mathbf{V}.\bar{\mathbf{V}} \tag{5.3}$$

for all $\mathbf{V} \in \mathbf{C}^3$ and all (x_1, x_2). Then

$$\lambda_n^{\epsilon'}(\beta) \geq \lambda_n^{\epsilon}(\beta) \tag{5.4}$$

for all n and all β with $\beta^2 > \omega^2\mu_0\epsilon_2$.

Proof. We first assume additionally that $\epsilon' - \epsilon_2$ is positive definite on Ω', i.e., that we have $\epsilon'_{1,\min} > \epsilon_2$. Then (5.3) implies that $\epsilon - \epsilon_2$ is also positive definite on Ω' and by Lemma 5.1 we have

$$(\epsilon'(x_1, x_2) - \epsilon_2)^{-1}\mathbf{V}.\bar{\mathbf{V}} \geq (\epsilon(x_1, x_2) - \epsilon_2)^{-1}\mathbf{V}.\bar{\mathbf{V}}, \tag{5.5}$$

for all $(x_1, x_2) \in \Omega'$ and all $\mathbf{V} \in \mathbf{C}^3$. Furthermore, both

$$(\epsilon'(x_1, x_2) - \epsilon_2)^{-1/2} \text{ and } (\epsilon(x_1, x_2) - \epsilon_2)^{-1/2} \tag{5.6}$$

exist and are measurable and bounded on Ω' and hence the mappings

$$\mathbf{G} \mapsto (\epsilon' - \epsilon_2)^{-1/2}\mathbf{G} \text{ and } \mathbf{G} \mapsto (\epsilon - \epsilon_2)^{-1/2}\mathbf{G} \tag{5.7}$$

are isomorphisms of $L^2(\Omega')^3$ onto itself and from n-dimensional subspaces of $L^2(\Omega')^3$ onto n-dimensional subspaces of $L^2(\Omega')^3$. Hence,

$$\begin{aligned}
\lambda_n^{\epsilon'}(\beta) &= \min_{W_n(\Omega')} \max_{\mathbf{F}\in W_n(\Omega')} \frac{\left(T_\beta^{\epsilon'}(\mathbf{F}),\, \mathbf{F}\right)}{(\mathbf{F},\, \mathbf{F})} \\
&= \min_{W_n(\Omega')} \max_{\mathbf{G}\in W_n(\Omega')} \omega^2\mu_0 \frac{\left((A_\beta^{\epsilon_2})^{-1}(\mathbf{G}),\, \mathbf{G}\right)}{\left((\epsilon' - \epsilon_2)^{-1/2}\mathbf{G},\, (\epsilon' - \epsilon_2)^{-1/2}\mathbf{G}\right)} \\
&\geq \min_{W_n(\Omega')} \max_{\mathbf{G}\in W_n(\Omega')} \omega^2\mu_0 \frac{\left((A_\beta^{\epsilon_2})^{-1}(\mathbf{G}),\, \mathbf{G}\right)}{\left((\epsilon - \epsilon_2)^{-1/2}\mathbf{G},\, (\epsilon - \epsilon_2)^{-1/2}\mathbf{G}\right)} \\
&= \min_{W_n(\Omega')} \max_{\mathbf{F}\in W_n(\Omega')} \frac{\left(T_\beta^{\epsilon}(\mathbf{F}),\, \mathbf{F}\right)}{(\mathbf{F},\, \mathbf{F})} \\
&\geq \min_{W_n(\Omega)} \max_{\mathbf{F}\in W_n(\Omega)} \frac{\left(T_\beta^{\epsilon}(\mathbf{F}),\, \mathbf{F}\right)}{(\mathbf{F},\, \mathbf{F})} \\
&= \lambda_n^{\epsilon}(\beta), \qquad (5.8)
\end{aligned}$$

where the first inequality follows from (5.5) and the fact that we may assume that $((A_\beta^{\epsilon_2})^{-1}(\mathbf{G}), \mathbf{G}) < 0$, and where the last inequality is a consequence of $\Omega' \subset \Omega$. Hence (5.4) is proved under the additional assumption that $\epsilon'_{1,\min} > \epsilon_2$.

The general case can be reduced to the previous one by replacing the tensors ϵ'_1 and ϵ_1 on Ω' and Ω by

$$\epsilon'_1 + \delta I \quad \text{and} \quad \epsilon_1 + \delta I, \tag{5.9}$$

respectively, where δ is a positive number. By the proof just given, the eigenvalues corresponding to the modified tensors satisfy (5.4) for all positive δ. Hence, by passing to the limit $\delta \to 0$, Lemma 5.2 follows for the general case. □

With this lemma we can prove Theorem 5.3.

THEOREM 5.3. *There always exist two guided modes that propagate in the same axial direction and that are linearly independent as electromagnetic fields. The propagation constants satisfy* $\beta^2 < \omega^2 \mu_0 \epsilon_{1,\max}$.

Proof. If $\mathbf{E}$ is the electric field corresponding to propagation constant β, then the field $(E_1, E_2, -E_3)^T$ corresponds to $-\beta$. Hence, it is sufficient to consider guided modes corresponding to positive β, i.e., the modes (2.6) travel in the positive x_3-direction. Because of assumption (2.4) there exists a disc Ω_R with radius R that is contained in Ω and such that

$$\epsilon_{R,1} := \inf_{(x_1,x_2)\in\Omega_R} \min_{\mathbf{V}\in\mathbf{C}^3} \frac{\epsilon(x_1,x_2)\mathbf{V}.\bar{\mathbf{V}}}{\mathbf{V}.\bar{\mathbf{V}}} > \epsilon_2. \tag{5.10}$$

We define the piecewise constant permittivity function $\epsilon_R : \mathbf{R}^2 \mapsto [0,\infty)$ by

$$\epsilon_R(x_1,x_2) = \begin{cases} \epsilon_{R,1} \text{ for } (x_1,x_2) \in \Omega_R, \\ \epsilon_2 \text{ for } (x_1,x_2) \in \mathbf{R}^2 \backslash \Omega_R. \end{cases} \tag{5.11}$$

For the isotropic, homogeneous circular waveguide Ω_R with permittivity $\epsilon_{R,1}$, one can derive transcendental equations linking the propagation constant $\beta > \omega(\mu_0\epsilon_2)^{1/2}$ to the eigenvalues $\lambda_n^{\epsilon_R}(\beta)$. These equations are listed in the appendix. We only need the property that the smallest two eigenvalues coincide,

$$\lambda_1^{\epsilon_R}(\beta) = \lambda_2^{\epsilon_R}(\beta), \tag{5.12}$$

and that

$$\lim_{\beta \downarrow \omega(\mu_0\epsilon_2)^{1/2}} \lambda_1^{\epsilon_R}(\beta) = -\infty. \tag{5.13}$$

By applying Lemma 5.2 with Ω_R instead of Ω' and with ϵ_R instead of ϵ', we obtain the inequalitites

$$\lambda_n^{\epsilon}(\beta) \le \lambda_n^{\epsilon_R}(\beta), \tag{5.14}$$

for all n and all $\beta > \omega(\mu_0\epsilon_2)^{1/2}$. Hence

$$\liminf_{\beta \downarrow \omega(\mu_0\epsilon_2)^{1/2}} \lambda_n^{\epsilon}(\beta) = -\infty, \quad \text{for } n = 1, 2. \tag{5.15}$$

Furthermore, by Theorem 4.5 we have for all n and $\beta > \omega(\mu_0\epsilon_2)^{1/2}$

$$\tau(\beta) < \lambda_n^{\epsilon}(\beta), \tag{5.16}$$

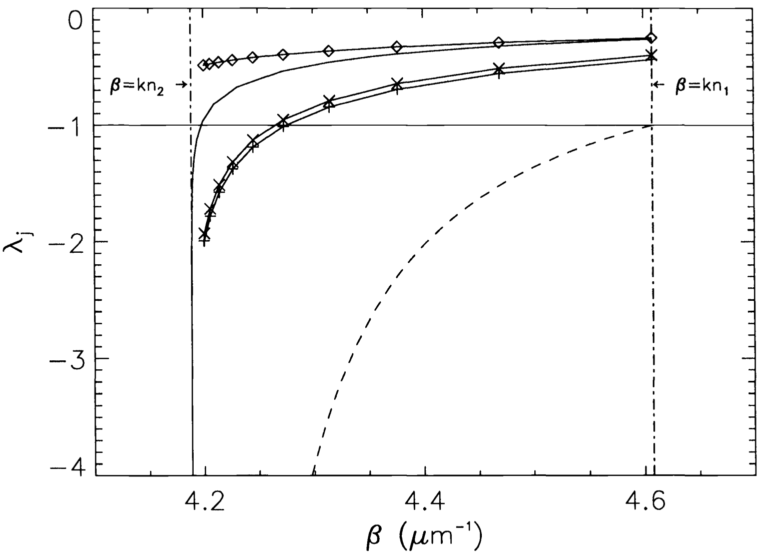


FIG. 5.1. *The smallest three eigenvalues $\lambda_1(\beta)$ (marker $+$), $\lambda_2(\beta)$ (marker $\times$), and $\lambda_3(\beta)$ (marker $\diamond$) for a homogeneous isotropic rectangular waveguide with dimensions $2\mu m \times 1\mu m$ and with refractive index $n_1 = 1.1$. The exterior region has refractive index $n_2 = 1.0$ and the wavenumber in vacuo $k = \omega(\mu_0\epsilon_0)^{1/2} = 4.18880\ (\mu m)^{-1}$. The dashed curve is the function $\tau(\beta)$ and the continuous curve is the smallest eigenvalue $\lambda_1(\beta) = \lambda_2(\beta)$ (counting muliplicity) of the largest circular waveguide contained in the rectangular guide, with refractive index n_1. It is seen that the third mode of the rectangular waveguide does not intersect the line $\lambda = -1$ and hence it is not guided. The eigenvalues of the rectangular guide were computed by applying the Galerkin method to the domain integral operator.*

where

$$\tau(\beta) = -\frac{\omega^2\mu_0\epsilon_{1,\max} - \omega^2\mu_0\epsilon_2}{\beta^2 - \omega^2\mu_0\epsilon_2}. \tag{5.17}$$

Hence

$$\lambda_n^\epsilon\left(\beta = \omega(\mu_0\epsilon_{1,\max})^{1/2}\right) > -1, \tag{5.18}$$

for all n. Because the operator T_β^ϵ depends continuously on $\beta > \omega(\mu_0\epsilon_2)^{1/2}$, so do its eigenvalues and therefore (5.15) and (5.18) imply that $\lambda_1^\epsilon(\beta)$ and $\lambda_2^\epsilon(\beta)$ both intersect the line $\lambda = -1$ at least once (see Fig. 5.1). It is clear that for positive β, points of intersection only occur for $\omega(\mu_0\epsilon_2)^{1/2} < \beta < \omega(\mu_0\epsilon_{1,\max})^{1/2}$. Hence to complete the proof it only remains to show that the modes are linearly independent as electromagnetic fields.

In case the values of β in the points of intersection are identical, the corresponding two fields are eigenfields of the same operator T_β^ϵ and hence they are linearly independent. In case the β are different, the fields are eigenfields of different operators and then the linear independence requires a proof. Now, as has already been pointed out in the introduction, the axial components E_3 and H_3 of the electromagnetic field can be eliminated from Maxwell's equations to obtain an eigenvalue problem for the remaining four components E_1, E_2, H_1, and H_2 with β eigenvalue

$$(\beta I - B^\epsilon)\begin{pmatrix} E_1 \\ E_2 \\ H_1 \\ H_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{5.19}$$

Here B^ϵ is the operator

$$B^\epsilon \begin{pmatrix} E_1 \\ E_2 \\ H_1 \\ H_2 \end{pmatrix} = \omega \begin{pmatrix} \mu_0 H_2 \\ -\mu_0 H_1 \\ -\epsilon E_2 \\ \epsilon E_1 \end{pmatrix} + \frac{1}{\omega} \begin{pmatrix} \frac{\partial}{\partial x_1}\left[\frac{1}{\epsilon}\left(\frac{\partial H_2}{\partial x_1} - \frac{\partial H_1}{\partial x_2}\right)\right] \\ \frac{\partial}{\partial x_2}\left[\frac{1}{\epsilon}\left(\frac{\partial H_2}{\partial x_1} - \frac{\partial H_1}{\partial x_2}\right)\right] \\ -\frac{1}{\mu_0}\frac{\partial}{\partial x_1}\left(\frac{\partial E_2}{\partial x_1} - \frac{\partial E_1}{\partial x_2}\right) \\ -\frac{1}{\mu_0}\frac{\partial}{\partial x_2}\left(\frac{\partial E_2}{\partial x_1} - \frac{\partial E_1}{\partial x_2}\right) \end{pmatrix}. \tag{5.20}$$

Because the electromagnetic fields of guided modes satisfy (5.19), we conclude that when the propagation constants differ, the electromagnetic fields are also always linearly independent. Hence Theorem 5.3 is proved. □

Although formulation (5.19) may at first sight seem relatively simple because it is in terms of an ordinary eigenvalue problem with the propagation constant as eigenvalue, it is actually difficult to study because the differential operator B^ϵ is not normal.

It should be noted that for $n = 1$ and for $n = 2$ only we have

$$\liminf_{\beta \downarrow \omega(\mu_0 \epsilon_2)^{1/2}} \lambda_n^\epsilon(\beta) = -\infty. \tag{5.21}$$

For $n \geq 3$ this lim inf is a finite negative number which, depending on the values of the permittivity and on the geometry of the cross-section, can be smaller or larger than -1. This can be verified explicitly for the case of the circular waveguide. Then, by choosing a circular waveguide that encloses the waveguide with cross-section Ω and has homogeneous isotropic permittivity $\epsilon_{R,1}$ with $\epsilon_{R,1} > \epsilon_{1,\max}$, the eigenvalues $\lambda_n^\epsilon(\beta)$ can be estimated from below by the $\lambda_n^{\epsilon_R}(\beta)$ and thus the lim inf in (5.21) is finite for $n \geq 3$ also in the general case. If the lim inf is smaller than -1 for a certain n, then for all $m \leq n$ the curves $\lambda_m^\epsilon(\beta)$ intersect $\lambda = -1$ at least once. Note that the proof of Theorem 5.3 implies that whenever two eigenvalue curves $\lambda_n^\epsilon(\beta)$ and $\lambda_m^\epsilon(\beta)$, with $n \neq m$, intersect $\lambda = -1$, the corresponding guided modes are always linearly independent as electromagnetic fields, whether the propagation constants are equal or not.

Numerical calculations suggest that the eigenvalues are increasing functions of $\beta > \omega(\mu_0\epsilon_2)^{1/2}$, but this has not been proved. It would be interesting to have a general proof because then we could conclude that each curve of eigenvalues $\lambda_n^\epsilon(\beta)$ yields at most one guided mode with positive β. Furthermore, the test whether the lim inf in (5.21) is smaller than -1 would not merely be a sufficient but also a necessary condition for that curve to yield a guided mode.

In addition to the mini–max principle (5.1) the following characterisation of the eigenvalues is useful:

$$\lambda_n^\epsilon(\beta) = \min_{\mathbf{F} \in \mathcal{H}_{n-1}^\perp,\, \|\mathbf{F}\| \leq 1} \left(T_\beta^\epsilon(\mathbf{F}),\, \mathbf{F}\right), \tag{5.22}$$

where $\mathcal{H}_{n-1}^\perp$ is the orthogonal complement in $L^2(\Omega)^3$ of the $(n-1)$-dimensional space of eigenfields corresponding to the smallest $n-1$ eigenvalues (counting multiplicities, of course). It is interesting to note that the existence of a solution of minimization problem (5.22) can be inferred directly from expression (3.4) for the quadratic form $(T^{\epsilon}_{\beta}(\mathbf{F}),\, \mathbf{F})$, without using a priori knowledge about the spectrum of T^{ϵ}_{β}. In fact, by (3.4)

$$\Big(T^{\epsilon}_{\beta}(\mathbf{F}),\, \mathbf{F}\Big) = p_1(\mathbf{F}) - p_2(\mathbf{F}), \tag{5.23}$$

where $p_1, p_2 : L^2(\Omega)^3 \mapsto [0, \infty)$ are continuous convex quadratic forms. The functional p_2 is in addition continuous for sequences that converge weakly in $L^2(\Omega)^3$. Hence both p_1 and $-p_2$ and therefore also the functional on the left-hand side of (5.23) are lower semicontinuous for sequences that converge weakly in $L^2(\Omega)^3$. The set in (5.22) in which the minimum is sought is compact with respect to the weak topology. Hence, the existence of a solution to minimization problem (5.22), and thus of an eigenfield of T^{ϵ}_{β}, follows from the general theorem that a lower semicontinuous functional attains its infimum on a compact set.

Characterisation (5.1) is particularly useful to deduce properties of numerical approximations. When for every fixed β an eigenvalue $\lambda^{\epsilon}_{n}(\beta)$ and the corresponding eigenfields are approximated by applying the Galerkin method to the domain integral equation, then (5.1) holds also for the discretised problem, provided the minimization over the n-dimensional spaces is restricted to subspaces of the finite-dimensional space used in the Galerkin approximation. It is thus evident that in the limit of increasing Galerkin base spaces, the computed eigenvalues approximate the exact eigenvalues from above. The convergence of the corresponding normalised eigenfields can be deduced from decomposition (5.23) of the quadratic form $(T^{\epsilon}_{\beta}(\mathbf{F}),\, \mathbf{F})$, using the fact that p_1 is a nonnegative bounded quadratic form and that p_2 is compact. By employing an appropriate iteration scheme to solve $\lambda^{\epsilon}_{n}(\beta) = -1$, the guided modes are determined. Hence, the Galerkin method applied to the domain integral equation always yields a converging algorithm. This is in contrast to the finite-element method applied directly to the partial differential equations. In the latter case the condition that the field has vanishing divergence has to be taken care of in the discretised equations because otherwise it can happen that the algorithm does not converge [8]. More details on computational aspects of the Galerkin method can be found in [12].

Appendix. Eigenvalues for the isotropic circular waveguide. Let Ω_R be a disc with radius R and let $\epsilon : \mathbf{R}^2 \mapsto [0, \infty)$ be the piecewise constant permittivity given by

$$\epsilon(x_1, x_2) = \begin{cases} \epsilon_1 & \text{for } (x_1, x_2) \in \Omega_R, \\ \epsilon_2 & \text{for } (x_1, x_2) \in \mathbf{R}^2 \backslash \Omega_R, \end{cases} \tag{A.1}$$

where ϵ_1 and ϵ_2 are positive constants with $\epsilon_1 > \epsilon_2$. Let β satisfy (2.11). Let $\lambda < 0$ be an eigenvalue of T^{ϵ}_{β} for the circular waveguide and let $\mathbf{F}$ be the eigenfield

$$\lambda \mathbf{F} = T^{\epsilon}_{\beta}(\mathbf{F}). \tag{A.2}$$

Using definition (2.21) of T^{ϵ}_{β}, it follows that

$$-\mathbf{F} = T^{\epsilon_{|\lambda|}}_{\beta}(\mathbf{F}), \tag{A.3}$$

where $\epsilon_{|\lambda|}$ is given by

$$\epsilon_{|\lambda|} = \epsilon_2 + \frac{\epsilon - \epsilon_2}{|\lambda|}. \tag{A.4}$$

Hence, $\mathbf{F}$ is eigenfield of the guided mode with propagation constant β of the circular waveguide with permittivity $\epsilon_{|\lambda|,1} = \epsilon_2 + (\epsilon_1 - \epsilon_2)/|\lambda|$ and with a cladding having permittivity ϵ_2. The guided modes of the circular waveguide are computed in many textbooks (e.g., [5], [11]). We give the equations that link the propagation constants β to the dielectric constant of the guide $\epsilon_{|\lambda|,1}$ and hence to the eigenvalues λ. Let $u = R\,(\omega^2\mu_0\epsilon_{|\lambda|,1} - \beta^2)^{1/2}$ and $w = R\,(\beta^2 - \omega^2\mu_0\epsilon_2)^{1/2}$. Then for some $m = 0, 1, 2, \dots$

$$\text{(A.5)}\quad \left\{\left[1 + \frac{u^2 + w^2}{R^2\,\omega^2\mu_0\epsilon_2}\right]\mathcal{J}_m(u) + \mathcal{K}_m(w)\right\}\{\mathcal{J}_m(u) + \mathcal{K}_m(w)\} - m^2\,\mathcal{D}(u,w) = 0,$$

where

$$\mathcal{J}_m(u) = \frac{J_m'(u)}{u\,J_m(u)},\quad \mathcal{K}_m(w) = \frac{K_m'(w)}{w\,K_m(w)},$$

$$\text{(A.6)}\qquad \mathcal{D}(u,w) = \left(1 + \frac{w^2}{R^2\,\omega^2\mu_0\epsilon_2}\right)\left(\frac{1}{u^2} + \frac{1}{w^2}\right)^2,$$

with J_m and K_m Bessel and modified Bessel functions, respectively. The eigenfields corresponding to the solutions of (A.5) for given m depend on the angle variable φ through the factor $\exp(i\,m\,\varphi)$ or $\exp(-i\,m\,\varphi)$. For $m \geq 1$ this means that there always exist two linearly independent fields corresponding to the same solution pair β, λ of (A.5). By using an approximation of $\mathcal{K}_1(w)$ for $w \approx 0$, it can be shown that for $m = 1$ there exists for every β with $\omega^2\mu_0\epsilon_2 < \beta^2 < \omega^2\mu_0\epsilon_1$ a solution λ of (A.5), such that if $\beta^2 \downarrow \omega^2\mu_0\epsilon_2$ then $\lambda \to -\infty$ [5], [11]. Furthermore, this solution is the smallest solution for all β. Hence it is equal to the lowest two eigenvalues $\lambda_1(\beta) = \lambda_2(\beta)$ when counting multiplicities.

Acknowledgment. I thank Prof. Ricardo Weder of the Universidad Nacional Autónoma de México for very valuable discussions.

REFERENCES

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

[2] J. S. Bagby, D. N. Nyguist, and B. C. Drachman, *Integral formulation for analysis of integrated dielectric waveguides,* IEEE Trans. Microwave Theory Tech., 33 (1985), pp. 906–915.

[3] N. H. G. Baken, *Computational modeling of integrated-optical waveguides*, Thesis, PTT Neher Laboratorium, Leidschendam, 1990.

[4] A. Bamberger and A. S. Bonnet, *Mathematical analysis of the guided modes of an optical fiber*, SIAM J. Math. Anal., 21 (1990), pp. 1487–1510.

[5] D. Marcuse, *Light Transmission Optics*, Van Nostrand, New York, 1972.

[6] ———, *Theory of Dielectric Optical Waveguides*, Academic Press, New York, 1974.

[7] C. Pichot, *Exact numerical solution for the diffused channel waveguide*, Optics Comm., 41 (1982), pp. 169–173.

[8] B. M. A. Rahman, F. A. Fernandez, and J. B. Davies, *Review of finite element methods for microwave and optical waveguides*, Proc. IEEE, 79 (1991), pp. 1442–1448.

[9] S. M. Saad, *Review of numerical methods for the analysis of arbitrarily-shaped microwave and optical dielectric waveguides*, IEEE Trans. Microwave Theory Tech., 33 (1985), pp. 894–899.

[10] M. Schechter, *Principles of Functional Analysis*, Academic Press, New York, 1973.

[11] M. S. Sodha and A. K. Ghatak, *Inhomogeneous Optical Waveguides*, Plenum Press, New York, 1977.

[12] H. P. Urbach and E. S. A. M. Lepelaars, *On the domain integral equation method for anisotropic inhomogeneous waveguides*, IEEE Trans. Microwave Theory Tech., 42 (1994), pp. 118–126.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 221–240, January 1996

WHEN THE LONG-TIME BEHAVIOR IS INDEPENDENT OF THE INITIAL DENSITY*

ANDRZEJ LASOTA† AND JAMES A. YORKE‡

Abstract. This paper investigates dynamical processes for which the state at time t is described by a density function, and specifically dynamical processes for which the shape of the density becomes largely independent of the initial density as time increases. A sufficient condition (weak ergodic theorem) is given for this "asymptotic similarity" of densities. The processes investigated are in general time dependent, that is, nonhomogeneous in time. Our condition is applied to processes generated by expanding mappings on manifolds, piecewise convex transformations of the unit interval, and integro-differential equations.

Key words. asymptotic behavior, Frobenius–Perron operator, invariant density, expanding mapping, integro-differential equation

AMS subject classifications. Primary 47A35; Secondary 58F11, 46D05

Introduction. A central feature in much of ergodic theory is that there is an invariant density or an invariant measure to which other initial densities are attracted as time evolves. The classical example is the Boltzmann equation and the celebrated H-theorem, which says that the process evolves to the state described by the density which maximalize the entropy. However, in some cases where there may be no stationary density, the long-time behavior is nonetheless independent of the initial distribution. For example, processes that vary in time do not in general have a stationary distribution. Results concerning the asymptotic behavior of such processes are often called weak ergodic theorems. One of the main problems of the weak ergodic theory can be summarized as follows. Consider a sequence of operators $P_1, P_2, \dots$ on a function space $L(X)$ and the ratio

$$q_n(x) = \frac{P_n P_{n-1} \cdots P_1 f(x)}{P_n P_{n-1} \cdots P g(x)} \quad \text{for } f, g \in L(X).$$

Find conditions under which $\{q_n(x)\}$ converges to a constant. In the case of linear operators P_i, answers to this question were given by many authors, see for example [G–K–R] and [C]. A systematic treatment in the special case where P_i are nonnegative matrices can be found in [Se]. The nonlinear case was considered in [F–K] and [N]. These results were obtained in general by quite sophisticated methods in which some topological, geometrical, and analytical properties of the space X and generators P_i played an important role. A particularly useful technique is related with Hilbert projection metric and its extensions [B].

A different method for studying the asymptotic behavior of the quotient $q_n(x)$ was proposed by D. Ruelle in his study of ergodic properties of the lattice gas [Rl]. In

* Received by the editors April 8, 1993; accepted for publication (in revised form) April 8, 1994. This research was supported in part by the National Science Foundation and the Department of Energy (Scientific Computing Staff Office of Energy Research).

† Institute of Mathematics, Silesian University, 40-007 Katowice, Poland. This research was done in part at the Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, MN 55455.

‡ Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742.

some places, his arguments are close to ours. In particular, he uses as in our proof of Theorem 1 the consecutive evaluation of $q_n(x)$ by a sequence of constants. However, as a whole, the Ruelle proof is based on the specific, delicate properties of the Gibbs measures. Further, in his case, the invariant measure always exists, which allows him to obtain a result corresponding to our formula (1.4).

Our approach is based on a relatively simple lower-bound technique. This method has two advantages. First, we do not need to assume any compactness conditions. Second, in applications a lower bound can be explicitly constructed due to the special properties of systems under consideration.

The paper is organized as follows. In this introductory section we show two examples which illustrate the notion of asymptotic similarity. The precise definition of this notion and the formulation of our main result—a convergence theorem—are given in §1. The next section is devoted to the proof. Section 3 contains an application to processes defined by expanding mappings on manifolds. In §4, we study processes defined by piecewise convex transformations on the unit interval. Finally, in §5, we show an application to integro-differential equations.

We now proceed with the examples.

Example 1. Assume that our state space is the interval [0, 1] and that the system under consideration evolves as follows. We have a sequence of transformations

$$F_i : [a_i, b_i] \xrightarrow{\text{onto}} [0,1], \qquad i = 0, 1, 2, \dots,$$

where $[a_i, b_i]$ are subintervals of [0, 1]. (This is basically a time-dependent process in which F_i is applied at time i.) We assume that F_i are C^1 functions with derivatives that satisfy

$$(0.1) \qquad |F_i'(x)| \geq \lambda > 1 \quad \text{for } x \in [a_i, b_i]; \qquad i = 0, 1, 2, \dots,$$

where λ is a constant independent of i. Given an initial point $x_0 \in [0,1]$, we define

$$(0.2) \qquad x_{n+1} = F_n(x_n) \quad \text{if } x_n \in [a_n, b_n].$$

If $x_n \notin [a_n, b_n]$, the process terminates and x_j is not defined for $j > n$. Now assume that x_0 is not fixed but is a random variable with a density distribution function f_0. Then the densities f_n of x_n are defined by the formula

$$(0.3) \qquad \int_A f_{n+1}(x)\,dx = \int_{F_n^{-1}(A)} f_n(x)\,dx \quad \text{for } A \subset [0,1], \quad A \text{ measurable.}$$

This formula becomes obvious if we observe that x_{n+1} is defined and belongs to A if and only if $x_n \in F_n^{-1}(A)$. Setting $G_i = F_i^{-1}$, we immediately obtain from (0.3)

$$(0.4) \qquad f_{n+1}(x) = |G_n'(x)| f_n(G_n(x)), \qquad n = 0, 1, \dots,$$

where $'$ denotes the derivative. Now choose two strictly positive continuous initial functions f_0 and g_0 and consider the corresponding sequences of distributions f_n and g_n. It is easy to show that the sequence of quotients

$$(0.5) \qquad q_n(x) = \frac{f_n(x)}{g_n(x)}$$

converges uniformly to a constant. In fact, from (0.4) it follows that

$$q_n(x) = \frac{f_0(\varphi_n(x))}{g_0(\varphi_n(x))}, \tag{0.6}$$

where $\varphi_n = G_0 \circ \cdots \circ G_{n-1}$. Further, according to (0.1), we have $|G_i'| \leq \lambda^{-1}$ and consequently, by the chain rule, $|\varphi_n'| \leq \lambda^{-n}$. Thus, the length of the intervals $\varphi_n([0,1])$ is at most λ^{-n}. Since the sequence of intervals $\varphi_n([0,1])$ is decreasing, i.e.,

$$\varphi_{n+1}([0,1]) = \varphi_n(G_n([0,1])) \subset \varphi_n([0,1]),$$

it has a unique common point, which we denote by x_*. This in turn implies that $\varphi_n(x)$ converges to x_* uniformly in x and, finally, due to (0.6),

$$\lim_{n\to\infty} q_n(x) = \frac{f_0(x_*)}{g_0(x_*)} \quad \text{uniformly in } x.$$

Observe that the sequences $\{f_n\}$ and $\{g_n\}$ converge uniformly to zero.

Example 2. Now let our state space be the unit circle S^1. (We obtain S^1 from [0, 1] by identifying the endpoints 0 and 1.) Let $F_i : S^1 \to S^1$ $(i = 0, 1, 2, \ldots)$ be a sequence of C^2 transformations satisfying the inequality $|F_i'(x)| \geq \lambda > 1$ $(x \in S^1, i = 0, 1, \ldots)$, which is analogous to (0.1). We define our dynamics by $x_{n+1} = F_n(x_n)$ $(n = 0, 1, \ldots)$ starting with an initial $x_0 \in S^1$. This time, however, the process never stops. As before, the corresponding sequence of densities f_n is defined by condition (0.3). To find a more explicit formula, observe that the inequality $F_i'(x) \neq 0$ implies that every point $x \in S^1$ has a finite number of counterimages which are functions of x. Thus

$$F_i^{-1}(x) = \{G_{i1}(x), \ldots, G_{ik_i}(x)\},$$

where k_i is the number of primages of each x for the function F_1. In a neighborhood of x, every inverse function $G_{i,j}$ of F_i is C^2 and satisfies the inequality $|G_{ij}'| \leq \lambda^{-1}$. Using these functions and condition (0.3), we immediately obtain

$$f_{n+1}(x) = \sum_{j=1}^{k_n} |G_{nj}'(x)| f_n(G_{nj}(x)). \tag{0.7}$$

Once again we may study the quotient (0.5). We will prove that it converges to the constant

$$c = \frac{\int_0^1 f_0(x)\,dx}{\int_0^1 g_0(x)\,dx}. \tag{0.8}$$

However, this cannot be done immediately. In the simplest case, where $F_n = F$ are the same for all n, the convergence $q_n(x) \to c$ is equivalent to the fact that, under the dynamical system (F, S^1), densities evolve asymptotically to an invariant density—the classical result of ergodic theory ([Re], [Ro], [Li]). We will come back to this example in §4 when our main result, Theorem 1, will be proved. Here we can explain only why the limit of the quotient $q_n(x)$, when it exists, must be equal to (0.8). This is due to the special property of the mapping $f_n \to f_{n+1}$ given by the formula (0.7).

Namely, it preserves the integral. Writing (0.5) in the form $f_{n+1}(x) = q_n(x)f_n(x)$ and integrating, we obtain, under assumption $q_n(x) \to c$, equality (0.8).

1. The asymptotic similarity. Let X be a nonempty set. Let $B(X)$ denote the space of all real-valued bounded functions $f : X \to \mathbb{R}$ with the supremum norm. Let $L = L(X)$ be an arbitrary fixed linear subspace of $B(X)$. For example, X might be a compact space and L could be the space of continuous functions on X. Let $L_+ = L_+(X)$ denote the subset of L consisting of the strictly (uniformly) positive functions, i.e.,

$$L_+ = \{f \in L : \inf f > 0\}.$$

We assume that L_+ is a nonempty set.

A linear operator $P : L \to L$ will be called *positive* if $P(L_+) \subset L_+$. Observe that a positive operator is monotonic, i.e., $Pf \le Pg$ whenever $f \le g$. By linearity, it is sufficient to verify that P positive implies $Pf \ge 0$ when $f \ge 0$. To prove the last property, fix an $f \ge 0$ and choose an arbitrary $h \in L_+$. Then $n^{-1}h + f \in L_+$ for $n = 1, 2, \ldots$, and consequently

$$n^{-1}Ph + Pf = P(n^{-1}h + f) \in L_+;$$

in particular, $n^{-1}Ph + Pf \ge 0$. Passing to the limit as $n \to \infty$, we have $Pf \ge 0$. Furthermore, a positive operator is always bounded. To see this, again choose an $h \in L_+$ and set $\alpha = \inf h$ and $\beta = \sup Ph$. We have

$$Pf = P\left(\frac{f}{h}\, h\right) \le P\left(\frac{\|f\|}{\alpha}\, h\right) \le \frac{\beta}{\alpha}\, \|f\|;$$

analogously

$$-Pf = P\left(-\, \frac{f}{h} h\right) \le \frac{\beta}{\alpha}\, \|f\|.$$

This implies

$$\|Pf\| \le \frac{\beta}{\alpha}\, \|f\|.$$

Our goal is to study a family $P(t,s)$ ($t \ge s$; $t, s \in \mathcal{D}(P)$) of linear positive operators. We will say that $P(t,s)$ is a *process* if $P(s,s) = I$ (identity) and

$$P(t,r)P(r,s) = P(t,s) \quad \text{for } t \ge r \ge s; \qquad t, r, s \in \mathcal{D}(P) \tag{1.1}$$

and if the domain $\mathcal{D}(P)$ is a subset of the real line which contains a sequence converging to $+\infty$. In applications, $\mathcal{D}(P)$ is either the set of nonnegative integers or the half-line $\mathbb{R}_+ = [0, \infty)$. We say the process is *positive* if $P(t,s)$ is a (linear) positive operator for all $t \ge s$ in $\mathcal{D}(P)$.

We now are ready to state our main result.

We will say $P(t,s)$ is an *eventually uniformly positive process* if $P(t,s)$ is a positive process and there is a subset $L_0 \subset L_+$ that is dense in L_+ and a constant $\alpha > 0$ such that, for every $f \in L_0$ and $s \in \mathcal{D}(P)$, the inequality

$$\frac{P(t,s)f}{\|P(t,s)f\|} \ge \alpha \tag{1.2}$$

holds for sufficiently large t, that is, for $t \geq t_0 = t_0(f, s)$. We will say that $P(t, s)$ has the property of *asymptotic similarity* if, for every $f \in L$, $g \in L_+$, and $s \in \mathcal{D}(P)$, there is a constant $c = c(f, g, s)$ such that

$$\lim_{t \to \infty} \left\| \frac{P(t,s)f}{P(t,s)g} - c \right\| = 0. \tag{1.3}$$

THEOREM 1. *Every eventually uniformly positive process has the property of asymptotic similarity.*

The proof will be given in the next section.

Now consider two special cases when $P(t, s)$ is generated by an operator or by a continuous semigroup.

When $P : L \to L$ is a positive operator, then

$$P(n, m) = P^{n-m}, \qquad n \geq m$$

is a process whose domain consists of nonnegative integers. Conditions (1.2) and (1.3) now simplify to

$$\frac{P^n f}{\|P^n f\|} \geq \alpha \quad \text{for } n \geq n_0(f) \tag{1.2$'$}$$

and

$$\lim_{n \to \infty} \left\| \frac{P^n f}{P^n g} - c(f, g) \right\| = 0, \tag{1.3$'$}$$

respectively. Assume now that P has a positive eigenfunction, i.e.,

$$P f_* = \lambda f_*$$

for some $\lambda > 0$ and $f_* \in L_+$. Then according to (1.3′),

$$\lambda^{-n} P^n f = \frac{f_* P^n f}{\lambda^n f_*} = \frac{f_* P^n f}{P^n f_*} \to c(f, f_*) f_*.$$

Since $\lambda^{-n} P^n f$ depends linearly on f and f_* is positive, the functional $\sigma(f) = c(f, f_*)$ is uniquely determined and linear. Thus we have

$$\lim_{n \to \infty} \|\lambda^{-n} P^n f - \sigma(f) f_*\| = 0 \quad \text{for } f \in L. \tag{1.4}$$

This shows that, for large n, the shape of the functions $P^n f$ is nearly the same and the amplitude changes in a geometrical progression.

Now let $P^t : L \to L, t \in \mathbb{R}_+$, be a continuous semigroup. Thus, we have

$$P^0 = I \text{ (identity)}; \qquad P^{t_1+t_2} = P^{t_2+t_2} = P^{t_1} P^{t_2} \quad \text{for } t_1, t_2 \geq 0$$

and

$$\lim_{t \to 0} \|P^t f - f\| = 0 \quad \text{for } f \in L.$$

We define a process $P(t, s)$, setting

$$P(t, s) = P^{t-s} \quad \text{for } t \geq s \geq 0.$$

Again, conditions (1.2) and (1.3) reduce to

$$\text{(1.2}''\text{)} \qquad \frac{P^n f}{\|P^n f\|} \geq \alpha \quad \text{for } t \geq t_0(f)$$

and

$$\text{(1.3}''\text{)} \qquad \lim_{t\to\infty} \left\| \frac{P^t f}{P^t g} - c(f,g) \right\| = 0.$$

Assume that for some $t_0 > 0$ the operator P^{t_0} has a positive eigenvalue λ corresponding to an eigenfunction $f_* \in L_+$. Following [L–R], we can find an explicit formula describing the behavior of $P^t f$. Define $\gamma = (1/t_0)\log\lambda$. According to (1.4), we have

$$\text{(1.5)} \qquad \lim_{n\to\infty} \|e^{-\gamma n t_0} P^{n t_0} f - \sigma(f) f_*\| = 0 \quad \text{for } f \in L.$$

Substituting $f = P^t f_*$ into (1.5), we obtain

$$\text{(1.6)} \qquad P^t f_* = \sigma(P^t f_*) f_* \quad \text{for } t \geq 0.$$

Due to the continuity of $P^t f_*$, with respect to t, the function $\beta(t) = \sigma(P^t f_*)$ is continuous. From (1.6), it follows that β satisfies the Cauchy equation

$$\beta(t_1 + t_2) = \beta(t_1)\beta(t_2) \quad \text{for } t_1 \geq 0; \qquad t_2 \geq 0.$$

Thus β is an exponential function. From (1.6), it follows also that $\beta(t_0) = e^{\gamma t_0}$. Thus $\beta(t) = e^{\gamma t}$ and (1.6) admits the form

$$\text{(1.7)} \qquad P^t f_* = e^{\gamma t} f_*.$$

Now we may repeat the argument used in the discrete time case and write

$$e^{-\gamma t} P^t f = \frac{f_* P^t f}{e^{\gamma t} f_*} = \frac{f_* P^t f}{P^t f_*},$$

and, according to (1.3″), we finally obtain

$$\text{(1.8)} \qquad \lim_{t\to\infty} \|e^{-\gamma t} P^t f - \sigma(f) f_*\| = 0,$$

with $\sigma(f) = c(f, f_*)$, which is the desired formula.

2. Proof of the convergence theorem. The proof will be given in several steps.

Step I. First observe that, for every $f \in L$ and $g \in L_+$, the inequality

$$a \leq \frac{f}{g} \leq b,$$

where a and b are constants, implies

$$a \leq \frac{P(t,s)f}{P(t,s)g} \leq b \quad \text{for } t \geq s; \qquad t, s \in \mathcal{D}(P).$$

To verify this, multiply the first inequality by g, apply operator $P(t,s)$, and then divide by $P(t,s)g$.

Step II. We now show the following extension of inequality (1.2). If the number α is replaced by an arbitrary $\alpha' < \alpha$, then (1.2) holds for every $f \in L_+$. To prove this, fix $\alpha' < \alpha$. For given $f \in L_+$, choose an $\varepsilon > 0$ such that

$$\frac{1+\varepsilon}{1-\varepsilon}\,\alpha' \le \alpha, \qquad \varepsilon < 1$$

and a $\delta > 0$ satisfying

$$\frac{\delta}{\inf f - \delta} \le \varepsilon, \qquad \delta < \inf f.$$

Since L_0 is dense in L_+, there is an $\bar{f} \in L_0$ such that $\|f - \bar{f}\| \le \delta$. We therefore have

$$\left|\frac{f}{\bar{f}} - 1\right| \le \frac{\delta}{\bar{f}} \le \frac{\delta}{\inf f - \delta} \le \varepsilon,$$

and, according to Step I,

$$\left|\frac{P(t,s)f}{P(t,s)\bar{f}} - 1\right| \le \varepsilon \quad \text{for } t \ge s; \qquad t,s \in \mathcal{D}(P).$$

Now let $t_0 = t_0(f,s)$ be chosen such that

$$\frac{P(t,s)\bar{f}}{\|P(t,s)\bar{f}\|} \ge \alpha \quad \text{for } t \ge t_0.$$

Then, finally,

$$\frac{P(t,s)f}{\|P(t,s)f\|} = \frac{P(t,s)f/P(t,s)\bar{f}}{\|P(t,s)f\|/\|P(t,s)\bar{f}\|} \cdot \frac{P(t,s)\bar{f}}{\|P(t,s)\bar{f}\|} \ge \frac{1-\varepsilon}{1+\varepsilon}\,\alpha \ge \alpha' \quad \text{for } t \ge t_0.$$

Step III. Let r be an arbitrary real number satisfying $r \ge 1/(\alpha')^2$. We claim that, for every $f, g \in L_+$ and $s \in \mathcal{D}(P)$, there is a $t_1 = t_1(f,g,s)$ such that

$$\sup \frac{P(t,s)f}{P(t,s)g} \le r \inf \frac{P(t,s)f}{P(t,s)g} \quad \text{for } t \ge t_1. \tag{2.1}$$

To prove this, choose $t_0(f,s)$ and $t_0(g,s)$ according to Step II and set $t_1 = \max(t_0(f,s), t_0(g,s))$. We have

$$\sup \frac{P(t,s)f}{P(t,s)g} \le \frac{\sup P(t,s)f}{\inf P(t,s)g} \le \frac{(\alpha')^{-1} \inf P(t,s)f}{\alpha' \sup P(t,s)g} \le r \inf \frac{P(t,s)f}{P(t,s)g} \quad \text{for } t \ge t_1.$$

Step IV. Choose $r > 1$ according to Step III. Assume that $f, g \in L_+$ and

$$0 < a \le \frac{f}{g} \le b,$$

where a, b are constants. Fix an $s \in \mathcal{D}(P)$. We claim that there is a $t_2 = t_2(f,g,a,b,s)$ and constants a_* and b_* such that

$$a_* \le \frac{P(t,s)f}{P(t,s)g} \le b_* \quad \text{for } t \ge t_2 \tag{2.2}$$

and

$$b_* - a_* \le \left(1 - \frac{1}{2r}\right)(b - a). \tag{2.3}$$

To prove the claim choose a number δ such that

$$0 < \delta < \min\left(a, \frac{b-a}{2(r-1)}\right) \tag{2.4}$$

and define $\rho = a - \delta$. Then $f - \rho g \in L_+$. Applying inequality (2.1) to $f - \rho g$ and g, we obtain

$$\sup \frac{P(t,s)(f-\rho g)}{P(t,s)g} \le r \inf \frac{P(t,s)(f-\rho g)}{P(t,s)g} \quad \text{for } t \ge t_2, \tag{2.5}$$

where $t_2 = t_2(f - \rho g, g, s)$. Now set

$$a_* = \inf \frac{P(t_2,s)f}{P(t_2,s)g}, \qquad b_* = \sup \frac{P(t_2,s)f}{P(t_2,s)g}.$$

Applying the result of Step I to functions $P(t_2,s)f$, $P(t_2,s)g$ and to the operator $P(t,t_2)$, we obtain immediately

$$a_* \le \frac{P(t,t_2)P(t_2,s)f}{P(t,t_2)P(t_2,s)g} \le b_* \quad \text{for } t \ge t_2.$$

From this and the chain rule (1.1), inequality (2.2) follows immediately. Moreover, (2.5) with $t = t_2$ implies

$$b_* - \rho \le r(a_* - \rho),$$

which, according to the definition of ρ, gives

$$b_* - a + \delta \le r(a_* - a + \delta).$$

Using this and the inequality $b - b_* \le r(b - b_*)$, we obtain

$$b - a \le r(b-a) + r(a_* - b_*) + (r-1)\delta.$$

Since, by the definition of δ, we have $(r-1)\delta \le \frac{1}{2}(b-a)$, the last inequality implies (2.3).

Step V. For any $f, g \in L_+$ and $s \in \mathcal{D}(P)$, define

$$a(t) = \inf \frac{P(t,s)f}{P(t,s)g}, \qquad b(t) = \sup \frac{P(t,s)f}{P(t,s)g}.$$

From Step I and the chain rule (1.1), it follows that $a(t)$ is an increasing function of t, $b(t)$ is decreasing, and $a(t) \le b(t)$. Moreover, using the result of Step IV, we can find a sequence $\tau_1 < \tau_2 < \cdots$ which converges to infinity and

$$b(\tau_n) - a(\tau_n) \le \left(1 - \frac{1}{2r}\right)^n (b(s) - a(s)).$$

This implies

$$\lim_{t\to\infty}(b(t)-a(t))=0$$

and proves condition (1.3) in the case when $f \in L_+$.

Step VI. Now let $f \in L$ be arbitrary, $g \in L_+$, and $s \in \mathcal{D}(P)$. In order to prove (1.3), in this case it is sufficient to show that there exist $f_1, f_2 \in L_+$ such that $f = f_1 - f_2$. To verify this, write

$$f_1 = f + kg, \qquad f_2 = kg,$$

where

$$k = 1 + \frac{\|f\|}{\inf g}.$$

The proof of Theorem 1 is completed. □

Analogous results for processes that are homogeneous in time (semigroups) were proved in [Ru] and [L–R]. These proofs were based on an additional assumption that X is a compact metric space.

Remark 1. From the chain rule (1.1), it follows that it is not necessary to verify condition (1.2) for all $s \in \mathcal{D}(P)$. It is sufficient to check it only for s sufficiently large, say $s \geq s_0$ and $f \in L_0$. In fact, for $s < s_0$ and $f \in L_+$, we have

$$\frac{P(t,s)f}{\|P(t,s)f\|} = \frac{P(t,s_0)f_0}{\|P(t,s_0)f_0\|} \quad \text{for } t \geq s,$$

where $f_0 = P(s_0,s)f \in L_+$. Then, applying Step II to the process $P(t,s)$ with $t \geq s \geq s_0$, we obtain

$$\frac{P(t,s_0)f_0}{\|P(t,s_0)f_0\|} \geq \alpha' \quad \text{for } t \geq t_0(f_0,s).$$

3. Expanding mappings. Using Theorem 1, we can easily finish the proof of the asymptotic similarity for the process described in Example 2. Instead, however, we will study a more general situation when the process is defined by a sequence of mappings on a compact manifold. The circle considered in Example 2 is a simple model of such manifold.

Thus, let X be a compact connected smooth (C^∞) manifold of dimension δ with a Riemannian metric $\|\cdot\|$. By ρ and μ we denote the distance and the measure corresponding to this metric. Let a C^2 mapping $F : X \to X$ be given. We define two numbers, the exponential factor λ_F and the second-order bound β_F.

The factor λ_F is defined by formula

$$\lambda_F = \inf \frac{\|dF(x)\xi\|}{\|\xi\|}, \tag{3.1}$$

where the infimum is taken for all $x \in X$ and $\xi \in T_x$, $\xi \neq 0$. Here, as usual, T_x denotes the tangent space to X at the point x and $dF(x)$ denotes the differential of F at x.

The geometrical meaning of λ_F is evident. If $\lambda_F > 0$, then the mapping F is locally invertible and every point $x \in X$ has a finite number of inverse images, a number independent of x. For every fixed x_0, there is a neighborhood U_{x_0} such that

$$F^{-1}(x) = \{G_1(x), \ldots, G_k(x)\} \quad \text{for } x \in U_{x_0}, \tag{3.2}$$

where G_i are smooth functions of the same order of regularity as F. (Thus, in our case, they are C^2.) When $\lambda_F > 1$, the mapping F is called expanding. In the special case when $X = S^1$ (unit circle), $\lambda_F = \inf |F'|$.

The parameter β_F is defined by formula

$$\beta_F = \sup_{x \in X} |\operatorname{grad} |\det dF(x)||. \tag{3.3}$$

In this definition, $|\det dF(x)|$ is the absolute value of the determinant of $dF(x)$. Observe that we do not assume that the manifold X is oriented, and thus only the absolute value of this determinant is well defined. The symbol $|\operatorname{grad} f|$ denotes the length of the gradient of f.

The geometrical interpretation of β_F is a little more complicated. Namely, $|\det dF(x)|$ is approximately equal to the ratio $\mu(F(U))/\mu(U)$, where U is a "small" neighborhood of x. Thus, β_F measures how quickly this ratio can change when x moves. If $X = S^1$, then $\beta_F = \sup |F''|$.

Now assume that $F_n : X \to X$ is a sequence of C^2 mappings ($n = 0, 1, \ldots$) with $\lambda_{F_n} > 0$. The Frobenius–Perron operators P_n corresponding to F_n are defined by a formula analogous to (0.3), namely,

$$\int_A P_n f \, d\mu = \int_{F_n^{-1}(A)} f \, d\mu \quad \text{for } A \subset X, \ A \text{ measurable.} \tag{3.4}$$

Using (3.2), it is easy to derive the explicit formula for P_n, namely,

$$P_n f(x) = \sum_i |\det dG_{ni}(x)| f(G_{ni}(x)), \tag{3.5}$$

where G_{ni} are the inverse functions to F_n defined in a neighborhood of x.

The process $P(n, m)$ is defined by the product

$$P(n, m) = P_{n-1} \cdots P_{m+1} P_m \quad \text{for } n > m > 0, \qquad P(n, n) = I \quad \text{for } n \geq 0$$

on the space $C(X)$ of all continuous functions on X.

THEOREM 2. *If all $\lambda_{F_n} > 0$ and in particular all F_n are locally invertible and if*

$$\liminf_{n \to \infty} \lambda_{F_n} > 1, \quad \sup_n \beta_{F_n} < \infty, \tag{3.6}$$

then $P(n, m)$ has the property of asymptotic similarity. Moroever, the constant c in condition (1.3) *is equal to $\int_X f \, d\mu / \int_X g \, d\mu$.*

Proof. Fix an integer n and consider an inverse transformation G_{ni} of F_n. Then

$$(dF_n)(G_{ni}(x)) dG_{ni}(x) = I$$

and

$$|(\det dF_n)(G_{ni}(x))| |\det dG_{ni}(x)| = 1 \tag{3.7}$$

for x belonging to the domain $\mathcal{D}(G_{ni})$ of G_{ni}. Since the manifold has dimension δ,

$$|\det dF_n(x)| \geq (\lambda_{F_n})^\delta \quad \text{and} \quad |\det dG_{ni}(y)| \leq \left(\frac{1}{\lambda_{F_n}}\right)^\delta \quad \text{for } y \in \mathcal{D}(G_{ni}). \tag{3.8}$$

Calculating the gradient of (3.7), we obtain

$$\begin{aligned}&|\det dG_{ni}(x)|(\mathrm{grad}|\det dF_n|)(G_{ni}(x))dG_{ni}(x)\\ &\quad = -|(\det dF_n)(G_{ni}(x))|\mathrm{grad}|\det dG_{ni}(x)|.\end{aligned}$$

From this and inequality $|vdG_{ni}(x)| \leq \lambda_{F_n}^{-1}|v|$, it follows that

$$\begin{aligned}&|(\det dF_n)(G_{ni}(x))||\mathrm{grad}|\det dG_{ni}(x)||\\ &\quad \leq \lambda_{F_n}^{-1}|\det dG_{ni}(x)||(\mathrm{grad}|\det dF_n|)(G_{ni}(x))|.\end{aligned}$$

Omitting the arguments to simplify notation, condition (3.3) defining β_{F_n} finally implies,

$$\frac{|\mathrm{grad}|\det dG_{ni}||}{|\det dG_{ni}|} \leq \lambda_n^{-1}\frac{|\mathrm{grad}|\det dF_n||}{|\det dF_n|} \leq \beta_n/\lambda_n^{\delta+1},$$

where we are using the abbreviations $\lambda_{F_n} = \lambda_n$ and $\beta_{F_n} = \beta_n$. Furthermore, writing

$$J_{ni} = |\det dG_{ni}|, \qquad |J'_{ni}| = |\mathrm{grad}|\det dG_{ni}||,$$

the second part of (3.8) and the last inequality become

$$J_{ni} \leq \left(\frac{1}{\lambda_n}\right)^{\delta}, \qquad \frac{|J'_{ni}|}{J_{ni}} \leq \frac{\beta_n}{\lambda_n^{\delta+1}}.$$

Now, using formula (3.5) for the Frobenius–Perron operator, we have

$$P_n f = \sum_i J_{ni}(f \circ G_{ni}).$$

Define the "regularity" of f for $f > 0$ to be (see [P–Y])

$$\mathrm{Reg}\ (f) = \sup_{x \in X}\frac{|\mathrm{grad}\ f(x)|}{f(x)}.$$

We may easily evaluate the regularity of $P_n f$. Namely,

$$\begin{aligned}\frac{|\mathrm{grad}\ P_n f|}{P_n f} &\leq \frac{\sum_i |J'_{ni}|(f \circ G_{ni})}{\sum_i J_{ni}(f \circ G_{ni})} + \frac{\sum_i J_{ni}|\mathrm{grad}\ f \circ G_{ni}||dG_{ni}|}{\sum_i J_{ni}(f \circ G_{ni})}\\ &\leq \frac{\beta_n}{\lambda_n^{\delta+1}} + \frac{1}{\lambda_n}\sup\frac{|\mathrm{grad}\ f|}{f} \quad \text{for } f \in C^1(x),\ f > 0.\end{aligned} \tag{3.9}$$

Hence, $\mathrm{Reg}\ (P_n f) \leq \frac{\beta_n}{\lambda_n^{\delta+1}} + \frac{1}{\lambda_n}\ \mathrm{Reg}\ (f)$. Let m_0 be such that $\lambda = \inf_{n \geq m_0}\lambda_n > 1$ and let $\beta = \sup \beta_n$. An induction argument yields the inequality

$$\mathrm{Reg}\ (P(n,m)f) \leq \frac{1}{\lambda^{n-m}}\ \mathrm{Reg}\ (f) + \frac{\beta}{\lambda^{\delta}(\lambda - 1)} \quad \text{for } n \geq m \geq m_0.$$

Set $k = 1 + \beta/(\lambda^{\delta}(\lambda - 1))$. Then for every strictly positive $f \in C^1(X)$ and every $m \geq m_0$, there is an $n_0 = n_0(f,m)$ such that

$$\mathrm{Reg}\ (P(n,m)f) \leq k \quad \text{for } n \geq n_0. \tag{3.10}$$

The last inequality implies that

$$P(n,m)f(y) \le e^{k\rho(x,y)}P(n,m)f(x) \tag{3.11}$$

for arbitrary $x, y \in X$. Thus, in particular,

$$\frac{P(n,m)f(x)}{\|P(n,m)f\|} \ge e^{-kd} \quad \text{for } n \ge n_0, \tag{3.12}$$

where $d = \sup_{x,y\in X} \rho(x,y)$. A straightforward application of Theorem 1 shows that $P(n,m)$ exhibits asymptotic similarity. Now let $g \in C^1(X)$ be strictly positive. From (3.11), it follows that

$$P(n,m)g(x) \le e^{kd} \min P(n,m)g \le \frac{e^{kd}}{\mu(X)} \int_X P(n,m)g\,d\mu.$$

Since $P(n,m)$ is the product of Frobenius–Perron operators which preserve the integral, we have

$$\int_X P(n,m)g\,d\mu = \int_X g\,d\mu,$$

and consequently

$$P(n,m)g(x) \le \frac{e^{kd}}{\mu(X)} \int_X g\,d\mu \quad \text{for } n \ge n_0.$$

This evaluation and (1.3) imply that

$$\lim_{n\to\infty} \|P(n,m)f - cP(n,m)g\| = 0$$

for $f \in C(X)$, $g \in C^1(X)$, $g > 0$. The last condition, in turn, implies

$$\lim_{n\to\infty} \left(\int_X P(n,m)f\,d\mu - c\int_X P(n,m)g\,d\mu \right) = 0. \tag{3.13}$$

Since the operators $P(n,m)$ preserve the integral in this example, they are uniformly continuous in the space of integrable functions on X. Thus, by an approximation argument, (3.13) holds for every integrable f and g. But due to the preservation of the integral, condition (3.13) reduces to

$$\int_X f\,d\mu - c\int_X g\,d\mu = 0,$$

which completes the proof. □

Remark 2. A close look at the proof of Theorem 2 shows that, in fact, we have proved a more general result. Namely, in order to derive (3.10) from (3.9), it is sufficient to assume that

$$\prod_{n=1}^{\infty} \lambda_n = \infty \quad \text{and} \quad \sup_n \sum_{i=1}^{n} \frac{\beta_{n-i}}{\lambda_{n-i}^{\delta+1}\lambda_{n-i+1}\cdots\lambda_n} < \infty, \tag{3.14}$$

and consequently, Theorem 2 is proved under assumptions (3.14), which are essentially weaker than (3.6). In particular, (3.14) can be satisfied even if infinitely many transformations F_n are not expanding but the remaining expand sufficiently strongly.

The asymptotic properties of expanding mappings were first studied by A. Rényi [Re] on interval [0, 1]. In fact, he studied them on S^1, since he assumed that the values at $x = 0$ and $x = 1$ are the same. The results of Rényi concerning the existence of an ergodic invariant measure were extended to manifolds by A. Avez [A], K. Krzyżewski and W. Szlenk [K–S], and K. Krzyżewski [K]. The last authors proved that every C^2 expanding mapping F has an invariant density f_* and that the sequence of iterates $P^n f$, where P is the Frobenius–Perron operator for F, converges uniformly to f_*.

Our Theorem 2 generalizes the convergence part of their results to the case when the densities are transformed by a sequence of transformations.

4. Piecewise-convex transformations. The main purpose for proving Theorem 1 was to establish a criterion implying asymptotic similarity, a criterion that is applicable to processes which fluctuate with time. Such an application was shown in the previous section. It is interesting, however, that even in the case when the process is described by one transformation and thus is homogeneous, we still get new results. In this section, we apply Theorem 1 to an old problem related to the existence of invariant measures for piecewise convex transformations. This problem also dates back to A. Rényi [Re], who prove the existence of an absolutely continuous ergodic invariant measure μ for the so-called r-adic transformations where $r > 1$ is not necessarily an integer,

$$F_r = rx(\text{mod } 1). \tag{4.1}$$

This system (F_r, μ) is exact (V. A. Rochlin [Ro]). From this and the characterization of exact systems of M. Lin [Li], it follows that, for every density f, the sequence of the iterates $\{P_r^n f\}_n$ of the Frobenius–Perron operator P_r of F_r converges strongly (i.e., in the L^1 norm) to the density $f_* = d\mu/dx$. Our goal is to show that, for a class of piecewise convex transformations, this convergence is uniform. This seems to be quite unexpected, since the limiting density, in general, is not continuous. In particular, it is not continuous for the transformation (4.1) if r is not an integer. In our treatment, *the functions are defined* at not just almost every but *at every point, so the uniform convergence will include the points of discontinuity of* f_*.

For simplicity, we will restrict ourselves to transformations F which are piecewise C^1. Recall that a real-valued C^1 function F defined on an interval Δ is *convex* if and only if its derivative F' is nondecreasing on Δ.

Let F be a given transformation on the half-open interval [0, 1) into itself. We will assume that F satisfies the following conditions:

(i) there is a partition $0 = a_0 < \cdots < a_r = 1$ such that for each integer i the restriction F_i of F to half-closed interval $[a_{i-1}, a_i)$ may be extended to the closed interval $[a_{i-1}, a_i]$ as a convex C^1 function;

(ii) $F(a_{i-1}) = 0, F'(a_{i-1}) > 0$ for $i = 1, \ldots, r$; and

(iii) $F([0, a_1)) = [0, 1), F'(0) > 1$.

We denote by P the Frobenius–Perron operator corresponding to F. As usual,

$$\int_A Pf(x)\,dx = \int_{F^{-1}(A)} f(x)\,dx \quad \text{for } A \subset [0, 1),\ A \text{ measurable.} \tag{4.2}$$

An elementary calculation shows that Pf may be written

$$Pf(x) = \sum_{i=1}^{r} G_i'(x) f(G_i(x)) \quad \text{for } 0 \leq x < 1, \tag{4.3}$$

where

$$G_i(x) = \begin{cases} F_i^{-1}(x) & \text{for } x \in F([a_{i-1}, a_i)), \\ a_i & \text{for } x \in [0,1) - F([a_{i-1}, a_i)), \end{cases}$$

and $G_i'(x)$ denotes the derivatives from the right so $G_i' = 0$ outside $F([a_{i-1}, a_i))$.

The functions G_i' are decreasing and are continuous from the right. Thus, if f is decreasing and continuous from the right, then Pf has the same property.

The following properties of the operator P given by formula (4.3) were proved in [L–Y]:

1°. The set

$$S = \bigcup_{n=0}^{\infty} F^{-n}(\{a_0, \ldots, a_r\})$$

is dense in [0, 1).

2°. If $f = 1_{[c,d)}$ is a characteristic function of an interval $[c, d)$ with the endpoints c and d belonging to S, then there exists an integer $n_0 = n_0(c, d)$ such that $P^n f$ for $n \geq n_0$ is a decreasing function on [0, 1).

Consider now the space L_* of all continuous from the right, piecewise-constant functions defined on [0, 1) and such that the points of discontinuity belong to S. In other words, L_* consists of the functions obtained by taking linear combinations of

$$1_{[c_i, d_i)}, \qquad c_i, d_i \in S.$$

Let L be the closure of L_* in the supremum norm topology. Due to property 1°, the space L in particular contains all continuous functions $f : [0,1) \to \mathbb{R}$. We define on L the process $P(n, m) = P^{n-m}$ for $n \geq m \geq 0$.

THEOREM 3. *The process $P(n, m)$ exhibits the property of asymptotic similarity.*

Proof. Let D_0 denote the subset of L_* which contains all normalized densities, i.e., all the functions satisfying

$$\int_0^1 f(x)\,dx = 1, \qquad f(x) \geq 0 \quad \text{for } 0 \leq x < 1. \tag{4.4}$$

Since P is a Frobenius–Perron operator, it preserves these conditions and $P(D_0) \subset D_0$. We shall use the following property of D_0, proved in [L–Y].

3°. There exists a number $M \geq 1$ (independent of f) such that

$$P^n f(x) \leq M \tag{4.5}$$

for every $f \in D_0$ and n sufficiently large, $n \geq n_0(f)$.

According to 2°, we may also assume that $P^n f$ are decreasing on [0, 1). Using this, we may evaluate $P^n f$ from below. Observe first that every decreasing function $g \in D$ with $g(0) \leq M$ satisfies

$$g(x) \geq \frac{1}{2M} \quad \text{for } 0 \leq x \leq \frac{1}{2M}. \tag{4.6}$$

In fact, suppose not and let $g(x_0) < 1/2M$ for an $x_0 \leq 1/2M$. Then

$$\begin{aligned} 1 = \int_0^1 g(x)\,dx &= \int_0^{x_0} g(x)\,dx + \int_{x_0}^1 g(x)\,dx \\ &\leq x_0 g(0) + (1 - x_0) g(x_0) < \frac{1}{2M}\,M + \frac{1}{2M} = \frac{M+1}{2M}, \end{aligned}$$

which contradicts the fact that $M \geq 1$. We may now extend our evaluation to the whole interval [0, 1) by the use of formula (4.3). Fix $f \in D_0$ and choose an $m \geq n_0(f)$. Define $g = P^m f$. From (4.3), it follows that

$$P^k g(x) \geq [\inf G_1']^k g(G_1^k(x)) \quad \text{for } 0 \leq x < 1,$$

where G_1^k is the kth iterate of G_1. Since $G_1(0) = 0$ and

$$G_1'(x) \leq G_1'(0) = 1/F_1'(0) < 1,$$

the integer k may be chosen such that $G_1^k(x) \leq \frac{1}{2M}$ for all $x \in [0, 1)$. Consequently, (4.6) implies

$$P^k g(x) \geq \delta, \qquad \text{where } \delta = \frac{[\inf(G_1')]^k}{2M}. \tag{4.7}$$

Observe that k does not depend on the choice of f. Define $n_1 = k + n_0(f)$. For $n \geq n_1$, we have $P^n f = P^k g$ with $g = P^{n-k} f$. Since $n - k \geq n_0$, we may apply inequality (4.7), which gives

$$P^n f(x) \geq \delta \quad \text{for } n \geq n_1(f), \quad 0 \leq x < 1. \tag{4.8}$$

From (4.5) and (4.8), it follows that, for every $f \in D_0$,

$$\frac{P^n f}{\|P^n f\|} \geq \frac{\delta}{M} \quad \text{for } n \geq n_1(f). \tag{4.9}$$

Now define, as in §1, $L_+ = \{f \in L : \inf f > 0\}$ and $L_0 = L_+ \cap L_*$. It is evident that L_0 is dense in L_+. Further, if $f \in L_0$, then $\lambda f \in D_0$ for some $\lambda > 0$. Thus (4.9) is also valid for every $f \in L_0$, completing the proof. □

We are now going to show that the operator P has a fixed point in the set L_+. To make this statement precise, we need to emphasize a subtle difference between formulas (4.2) and (4.3). For every given integrable f, formula (4.2) defines the function Pf almost everywhere. Formula (4.3) with given f defines $Pf(x)$ at each and every point $x \in [0, 1)$. We have proved Theorem 3 for the "precise" P defined by (4.3), and we want to find a fixed point f_* which satisfies

$$f_*(x) = \sum_{i=1}^{r} G_i'(x) f_*(G_i(x)) = Pf_*(x) \tag{4.10}$$

at every point $x \in [0, 1)$. Thus the classical technique of ergodic theory related with definition (4.2) is not helpful. In particular, we cannot use results proved in [L–Y] and [So], which assure the existence of an invariant density for a large class of piecewise-convex transformations, because the invariance there is understood as the equality $f = Pf$ almost everywhere.

Let $f \in D_0$ be an arbitrary function. We know that $g = P^{n_1} f$ satisfies

$$\delta \leq P^n g(x) \leq M \quad \text{for } n \geq 0, \quad 0 \leq x < 1$$

and that the functions $P^n g$ are continuous from the right and decreasing. Define

$$g_n = \frac{1}{n} \sum_{j=1}^{n} P^j g.$$

The functions g_n are again continuous from the right and are decreasing. They admit the values from the interval $[\delta, M]$. According to Helly's theorem (see [Lo] p. 12, Thm. 1.2.1), every bounded sequence of monotonic (decreasing or increasing) functions defined on an interval of the real line contains a pointwise-convergent subsequence. Thus from g_n we can select a subsequence g_{α_n} which converges to a decreasing function g_*. Further, since

$$|Pg_n(x) - g_n(x)| \leq \frac{2M}{n} \quad \text{for } 0 \leq x < 1,$$

the function g_* is a fixed point of P; that is,

$$g_*(x) = \sum_{i=1}^{r} G_i'(x) g_*(G_i(x)) \tag{4.11}$$

is satisfied in every point $x \in [0, 1)$. We don't know, however, if g_* is continuous from the right. Define

$$f_*(x) = g_*(x+0) = \lim_{\substack{y \to x \\ y > x}} g_*(y).$$

Passing to the right-hand limit in (4.11) and observing that G_i' are continuous from the right, we finally obtain (4.10).

It is evident that the function f_* has the following properties: it is decreasing, continuous from the right, and satisfies

$$\delta \leq f_*(x) \leq M, \qquad \int_0^1 g_*(x)\,dx = 1.$$

The last inequality is the consequence of the Lebesgue dominated convergence theorem. Thus, according to (1.4) with $\lambda = 1$, we have

$$\|P^n f - \sigma(f) f_*\| = 0 \quad \text{for } f \in L. \tag{4.12}$$

Since P preserves the integral and f_* is a normalized density, we have

$$\sigma(f) = \int_0^1 f(x)\,dx.$$

It should be emphasized that the convergence (4.12) holds, in particular, for every continuous bounded f and that the limiting function is in general discontinuous.

5. Integro-differential equations. In previous examples, time was discrete and the operators were generated by one-to-one transformations. We now present an application of Theorem 1 to a continuous-time system generated by an evolution equation of the form

$$\frac{du}{dt} = Au + Q(t)u \quad \text{for } t \geq s \tag{5.1}$$

with the initial condition

$$u(s) = f. \tag{5.2}$$

For simplicity, we will consider only the case when u assumes its values in the space of continuous functions.

Let X be a metric space and μ a finite Borel measure on X. Further, let $L = C(X)$ be the space of all real-valued bounded continuous functions on X with the supremum norm. We assume that A in (5.1) is the generator of a continuous (C_0) semigroup $P^t, t \geq 0$, of positive linear operators on $C(X)$ and that $Q(t)$, for $t \geq 0$, is a family of integral operators given by formula

$$Q(t)f(x) = \int_X K(t,x,y)f(y)\mu(dy), \tag{5.3}$$

where $K : \mathbb{R}_+ \times X \times X \to \mathbb{R}_+$ is a continuous kernel. More specifically, we make the following assumptions:

1°. The semigroup P^t satisfies condition (1.2″) and, in particular, it has the property of asymptotic similarity.

2°. For some value $t_0 > 0$, the operator P^{t_0} has an eigenvalue $\lambda > 0$ corresponding to an eigenfunction $f_* \in C_+(X)$.

3°. The functions

$$k_0(t) = \inf\{K(t,x,y) : x, y \in X\}$$

and

$$k_1(t) = \sup\{K(t,x,y) : x, y \in X\}$$

satisfy

$$\liminf_{t\to\infty} \frac{k_0(t)}{k_1(t)} > 0. \tag{5.4}$$

We define the process $U(t,s)$ by assuming

$$U(t,s)f = u(t)$$

where u is the solution of (5.1) and (5.2). Under our assumptions, $u(t)$ is uniquely defined for $t \geq s$ and strictly positive for strictly positive f. Thus U is a positive process.

The process $U(t,s)$ may be considered as a perturbation of $P(t,s) = P^{t-s}$ since the generator A of P^t is perturbed by the family of integral operators $Q(t)$. An important property of this perturbation is described by the following theorem.

THEOREM 4. *If A and $Q(t)$ satisfy conditions* 1° – 3°, *then $U(t,s)$ exhibits asymptotic similarity.*

Proof. Our main tool is the classical variation of constant formula for (5.1) using $Q(t)u = Q(t)U(t,s)f$. This formula gives

$$U(t,s)f = P^{t-s}f + \int_s^t P^{t-r}Q(r)U(r,s)f\,dr. \tag{5.5}$$

Let γ and f_* be as in formula (1.7). Define

$$\beta_0(t) = \frac{k_0(t)}{\sup f_*}, \qquad \beta_1(t) = \frac{k_1(t)}{\inf f_*}$$

and fix an $f \in C_+(X)$. Then

$$\beta_0(t)m(f)f_* \leq Q(t)f \leq \beta_1(t)m(f)f_*, \tag{5.6}$$

where

$$m(f) = \int_X f\,d\mu.$$

Using (5.5) and the second inequality (5.6), we obtain

$$U(t,s)f \le P^{t-s}f + \int_s^t P^{t-r}[\beta_1(r)m(U(r,s)f)f_*]\,dr.$$

Since $P^{t-r}f_* = e^{\gamma(t-r)}f_*$, this gives

$$U(t,s)f \le P^{t-s}f + \left\{\int_s^t e^{\gamma(t-r)}\beta_1(r)m(U(r,s)f)\,dr\right\} f_*.$$

Analogously, using the first inequality (5.6) we obtain

$$U(t,s)f \ge P^{t-s}f + \left\{\int_s^t e^{\gamma(t-r)}\beta_0(r)m(U(r,s)f)\,dr\right\} f_*,$$

and from this

$$(5.7)\qquad \frac{U(t,s)f(x)}{\|U(t,s)f\|} \ge \frac{e^{-\gamma(t-s)}P^{t-s}f(x) + \left\{\int_s^t e^{\gamma(s-r)}\beta_0(r)m(U(r,s)f)\,dr\right\} f_*(x)}{e^{-\gamma(t-s)}\|P^{t-s}f\| + \left\{\int_s^t e^{\gamma(s-r)}\beta_1(r)m(U(r,s)f)\,dr\right\} \|f_*\|}.$$

Since P^t satisfies condition (1.2″), we have

$$(5.8)\qquad \frac{P^{t-s}f(x)}{\|P^{t-s}f\|} \ge \alpha \quad \text{for } t \ge t_0(f,s).$$

Assume that s_0 is so large that

$$\frac{\beta_0(r)}{\beta_1(r)} = \frac{\inf f_*}{\sup f_*}\frac{k_0(r)}{k_1(r)} \ge \frac{1}{2}\frac{\inf f_*}{\sup f_*}\liminf_{t\to\infty}\frac{k_0(t)}{k_1(t)} =: \delta \quad \text{for } r \ge s_0.$$

From this definition of δ and (5.7) and (5.8), it follows that

$$\frac{U(t,s)f(x)}{\|U(t,s)f\|} \ge \min(\alpha,\delta)\,\frac{\min f_*}{\|f_*\|} \quad \text{for } s \ge s_0 \text{ and } t \ge t_0(f,s).$$

According to Remark 1, this completes the proof. □

Observe that if a semigroup $P^t, t \ge 0$, of positive operators satisfies condition (1.2″), then so does $e^{-\beta t}P^t$ for arbitrary real β. Further, if A generates P^t, then $A - \beta I$ (I = identity) generates $e^{-\beta t}P^t$. Thus, applying Theorem 4 to the equation

$$(5.9)\qquad \frac{du}{dt} + \beta u = Au + Q(t)u,$$

we obtain the following result.

COROLLARY 1. *If A generates a semigroup $P^t, t \ge 0$, satisfying conditions 1° and 2°, and if the family of integral operators* (5.3) *satisfies 3°, then the process generated by equation* (5.9) *and condition* (5.2) *has the property of asymptotic similarity.*

We close this section by showing a specific example of equation (5.9). Consider the integro-differential equation of the form

(5.10)
$$\frac{\partial u(t,x)}{\partial t} + a(x)\,\frac{\partial u(t,x)}{\partial x} + \beta u(t,x) = \int_0^1 K(t,x,y)u(t,y)\,dy, \quad 0 \le x \le 1, \quad t \ge 0.$$

We assume that the coefficient $a(x)$ is a C^1 function and that

$$a(0) = 0, \qquad a(x) > 0 \quad \text{for } x > 0. \tag{5.11}$$

We also assume that the kernel K in (5.10) satisfies condition 3° with $X = [0,1]$. To prove that equation (5.10) generates a process on $C([0,1])$ having the property of asymptotic similarity, it is sufficient to verify that the semigroup P^t generated by

$$Af = -a(x)\,\frac{df}{dx}, \qquad f \in C^1[0,1]$$

satisfies conditions 1° and 2°. This semigroup is given by the equation

$$\frac{\partial u(t,x)}{\partial t} + a(x)\,\frac{\partial u(t,x)}{\partial x} = 0 \quad \text{with } u(0,x) = f(x) \tag{5.12}$$

or more explicitly by formula

$$P^t f(x) = u(t,x) = f(\varphi(0;t,x)),$$

where $\varphi(t) = \varphi(t;t_0,x_0)$ is the solution of the ordinary differential equation $\varphi' = a(\varphi)$ with the initial condition $\varphi(t_0) = x_0$. Due to assumption (5.11), we have

$$\lim_{t\to\infty} \varphi(0;t,x) = 0$$

uniformly for $0 \le x \le 1$. Consequently,

$$\lim_{t\to\infty} \frac{P^t f(x)}{P^t f(y)} = \lim_{t\to\infty} \frac{f(\varphi(0;t,x)}{f(\varphi(0;t,y)} = \frac{f(0)}{f(0)} = 1$$

for every $f \in C_+(X)$, and the convergence is uniform. This implies (1.2″). Then 2° is satisfied with $f_* \equiv 1$ and $\lambda = 1$.

A result analogous to Theorem 4 was proved in [L–R] for homogeneous in time processes with X being a compact space. The proof presented here is much shorter. This is due to the fact that our main tool—Theorem 1—does not require any compactness assumption.

Theorem 4 can be easily generalized. We may replace, for example, the space $C(X)$ by a different subspace of $B(X)$ (bounded functions on X), and we could relax the assumption of the continuity of the kernel. We stated our result in a less general form for the sake of simplicity and to show its close relation to the results proved in [L–R].

REFERENCES

[A] A. Avez, *Propriétés ergodiques des endomorphismes dilatantes des variétés compactes*, C. R. Acad. Sci. Paris Sér A, 266 (1968), pp. 610–612.

[B] P. J. Bushell, *Hilbert's metric and positive contraction mappings in a Banach space*, Arch. Rational Mech. Anal., 52 (1973), pp. 330–338.

[C] J. E. Cohen, *Ergodic theorems in demography*, Bull. Amer. Math. Soc., 1 (1979), pp. 275–295.

[F–K] T. Fujimoto and U. Krause, *Asymptotic properties for inhomogeneous iterations of nonlinear operators*, SIAM J. Math. Anal., 19 (1988), pp. 841–853.

[G–K–R] M. Golubitsky, E. Keeler, and M. Rothschild, *Convergence of the age structure: Applications of the projective metric*, Theoret. Population Biol., 7 (1975), pp. 84–93.

[K] K. Krzyżewski, *Some results on expanding mappings*, Astérisque, 50 (1977), pp. 205–218.

[K–S] K. Krzyżewski and W. Szlenk, *On invariant measures for expanding differential mappings*, Studia Math., 33 (1969), pp. 83–92.

[L–R] A. Lasota and R. Rudnicki, *Asymptotic behaviour of semigroups of positive operators on $C(X)$*, Bull. Polish Acad. Sci. Math., 36 (1988), pp. 151–159.

[L–Y] A. Lasota and J. A. Yorke, *Exact dynamical systems and the Frobenius–Perron operator*, Trans. Amer. Math. Soc., 273 (1982), pp. 375–384.

[Li] M. Lin, *Mixing for Markov operators*, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 19 (1971), pp. 231–242.

[Lo] S. Lojasiewicz, *An Introduction to the Theory of Real Functions*, Wiley, New York, 1988.

[N] R. D. Nussbaum, *Some nonlinear weak ergodic theorems*, SIAM J. Math. Anal., 21 (1990), pp. 436–460.

[P–Y] G. Pianigiani and J. A. Yorke, *Expanding maps on sets which are almost invariants: Decay and chaos*, Trans. Amer. Math. Soc., 252 (1979), pp. 351–366.

[Re] A. Rényi, *Representation for real numbers and their ergodic properties*, Acta Math. Acad. Sci. Hungar., 8 (1957), pp. 477– 493.

[Rl] D. Ruelle, *Statistical mechanics of a one-dimensional lattice gas*, manuscript.

[Ro] V. A. Rochlin, *Exact endomorphisms of Lebesgue spaces*, Amer. Math. Soc. Transl. Ser. 2, 39 (1964), pp. 1–36.

[Ru] R. Rudnicki, *Asymptotic properties of the iterates of positive operators on $C(X)$*, Bull. Polish. Acad. Math. Sci., 34 (1986), pp. 181–187.

[Se] E. Seneta, *Nonnegative Matrices and Markov Chains*, Springer-Verlag, New York, 1981.

[So] J. Socala, *On the existence of invariant measures for Markov operators*, Ann. Polon. Math., 48 (1988), pp. 51–56.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 241–257, January 1996

013

A CENTER-UNSTABLE MANIFOLD THEOREM FOR PARAMETRICALLY EXCITED SURFACE WAVES*

LAWRENCE TURYN[†]

Abstract. When fluid in a rectangular tank sits upon a platform which is oscillating with sufficient amplitude, surface waves appear in the "Faraday resonance." Scientists and engineers have done bifurcation analyses which assume that there is a center manifold theory using a finite number of excited spatial modes. We establish such a center manifold theorem for Xiao-Biao Lin's model in which potential flow is assumed but an artificial dissipation term is included in the system of partial differential equations on the free surface. We use interpolation spaces developed by da Prato and Grisvard, establish maximal regularity for a family of evolution operators, and adapt the center manifold theory of Chow, Lin, and Lu.

Key words. parametric resonance, center manifold, surface waves, interpolation spaces, maximal regularity

AMS subject classifications. Primary, 76B15, 35G25, 34C45, 34G20; Secondary, 70J40, 35Q35, 47D06, 46B70

1. Introduction and summary.

Consider a rectangular tank filled with an incompressible homogeneous fluid to a depth h. If the base of the tank is made to oscillate, then what will be the behavior of the fluid? As long ago as 1831, Faraday observed fluid oscillations at one half the frequency of the base, so this phenomenon of parametric excitation is known as Faraday resonance. This and other historical references can be found in Benjamin and Ursell [3] and Miles and Henderson [18].

Lin [15] has established a mathematical formulation for this problem. As long as the amplitude of the excitation is sufficiently small, he has obtained (i) global existence and uniqueness and (ii) an approximation result which justifies a truncation to a finite number of modes. His model assumes potential flow and includes on the free boundary terms for surface tension and artificial viscosity. The latter is to some extent physically meaningful because Lin showed that at high wave numbers his artificial viscosity produces dissipation proportional to the square of the wave number and proportional to the total kinematic energy, i.e., the dissipation is consistent with that produced by kinematic viscosity.

On the other hand, several authors, e.g., Gu and Sethna [14], Holmes [13], and Silber and Knobloch [23], have analyzed bifurcation equations assumed to hold on a finite-dimensional center manifold corresponding to some neutrally stable mode(s). If there is friction in the system, this is only possible if the amplitude of the excitation is not required to be sufficiently small.

The purpose of this paper is to establish the existence of a local center manifold theorem for this problem. We introduce virtually no new mathematical techniques but instead apply a variety of results established by other authors.

Our mathematical formulation is due to Lin [15]. The velocity potential, as a function of the shape of the free surface and the potential on the free surface,

*Received by the editors May 25, 1993; accepted for publication (in revised form) April 15, 1994.

[†]Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435. This research was partially done while the author was visiting the School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332.

is substituted into Bernoulli's equation on the free surface. We use this functional relationship from Lin's paper, albeit in a time-independent version, in Theorem 6.1 below. Our techniques diverge from Lin's in that we then consider the problem as an ordinary differential equation in function spaces of space dependence rather than as an implicit equation in a Hilbert function space of both space and time dependence.

For our approach, we use a variation of constants formula in interpolation spaces, as in da Prato and Grisvard [8], Sinestrari [25], and da Prato and Lunardi [9], which gives "maximal regularity." Unfortunately, we could not use the methods of Henry [12] for our nonlinear problem, although we do mimic his methods for periodic, linear problems. The actual local center manifold theorem that we obtain is an application of Chow, Lin, and Lu [7], or we could have used the result of da Prato and Lunardi [9], which is also in the style of Henry [12, Chap. 6]. The method of Liapunov–Perron obtains an invariant manifold as a graph of a function defined by an integral operator.

The paper is organized in this way: In §2, the physical problem and a model for it are presented along with a linearization about the flat surface, i.e., undisturbed, solution. In §3, a semigroup of bounded linear operators is explicitly presented and shown to be analytic. In §4, a brief description is made of a well-known general method for defining interpolation spaces in which our semigroup has a maximal regularity property. In §5, we return to linear, time-periodic problems and show in Proposition 5.2 that there is a family of evolution operators which define integral operators with a generalization of the maximal regularity property for autonomous problems. We show that our linearized, time-periodic problem can be analyzed in this way. These generalizations are the only "new" results we have; all the other work in this paper consists of applying the work of other authors to our specific problem or, in the case of Lin's paper, taking a result in its entirety. In §6, the nonlinearities of our model are found to be nice enough viz. the integral operators of §5; here a result of Lin is essential. In §7, we adapt a center-unstable manifold theorem and an exponential attractivity theorem of other authors to our nonlinear, time-periodic problem. In §8, we mention a result of a sequel in preparation. We have calculated an approximate, local-center manifold for an example involving the (3, 2) and (2, 3) spatial modes.

In fact, experiments of Simonelli and Gollub [24] provide examples where there are a small number of unstable modes. As an aside, when we examined their experimental data for the onset of instability [24, p. 479, Fig. 4(a)] and used a theoretical result for the damped Mathieu equation from Turyn [27, §1], we came to the conclusion that the artificial viscosity term $-\mu\nabla_{\mathbf{x}}^2 u$ in equation (2.5) of our paper would have μ about two to three times the kinematic viscosity, 2.948 centipoises, of n-butyl alcohol at 20°C. This suggests that, although this artificial viscosity is consistent with dissipation of kinetic energy, as in Lin [15, §6], the mechanism may not be so simple.

Our original intention was to discuss regions more general than rectangular tanks, e.g., cylindrical tanks. We ran into difficulties in §6 when discussing boundary conditions to be satisfied by the nonlinearities. We hope to understand this better in the future.

2. The physical problem and its linearization. Let D be a bounded open domain in $\mathbb{R}^2$, with coordinates $\mathbf{x}$. We will define the smoothness we require of its boundary ∂D later. The moving coordinate z is fixed with respect to the oscillating container, with the positive axis pointing upwards. The free boundary is $S_F : z = v(\mathbf{x}, t)$ and we denote

$$\Omega_v = \Omega_v(t) := \{(\mathbf{x}, z) : -h < z < v(\mathbf{x}, t), \mathbf{x} \in D\},$$

with outward unit normal $\mathbf{n}$. We assume the velocity field $\mathbf{V}$ is the gradient of a potential $\phi = \phi(\mathbf{x}, z, t)$ which satisfies

$$\begin{cases} \nabla^2\phi = 0 & \text{in } \Omega_v, \quad (2.1)\\ \phi = u(\mathbf{x},t) & \text{on } S_F : z = v(\mathbf{x},t), \quad (2.2)\\ \frac{\partial\phi}{\partial\mathbf{n}} = 0 & \text{on } \partial\Omega_v \setminus S_F, \quad (2.3)\end{cases}$$

where $\nabla = (\nabla_{\mathbf{x}}, \frac{\partial}{\partial z})$. The latter boundary condition states that the normal velocity is zero on the sides and base of the container.

On the free boundary the potential ϕ and the shape v satisfy the "kinematic" condition

$$v_t = \phi_z - \nabla_{\mathbf{x}}\phi \cdot \nabla_{\mathbf{x}} v \quad \text{on } S_F, \tag{2.4}$$

where subscripts denote differentiation, and Bernoulli's equation

$$\phi_t = -\gamma\nabla\cdot\mathbf{n} - (g - \alpha(t))v - \frac{1}{2}\,|\nabla\phi|^2 + \mu\nabla_{\mathbf{x}}^2 u \quad \text{on } S_F, \tag{2.5}$$

where atmospheric pressure $p_0 \equiv 0$, density $\rho \equiv 1, \gamma$ is the coefficient of surface tension, g is the acceleration of gravity, the parametric excitation $\alpha(t)$ is the effect of the oscillating base of the container, and we have added an artificial dissipative term proportional to $\nabla_{\mathbf{x}}^2 u$, as in Lin [15]. Specifically, we have $\alpha(t) = a(d^2/dt^2)\cos(\omega t)$, where a is the amplitude of the oscillations of the base of the container.

On the free boundary, $u(\mathbf{x},t) = \phi(\mathbf{x}, v(\mathbf{x},t), t)$, so $u_t = \phi_t + \phi_z v_t$, etc. Denote $w(\mathbf{x},t) := \phi_z(\mathbf{x}, v(\mathbf{x},t), t)$, so that $w = N(v)u$ is a linear operator on u which depends nonlinearly on v. In Theorem 6.1, we will borrow from Lin [15] a result on the smooth dependence of w on u and v, in suitable function spaces of spatial dependence.

Denote by [ ] the operation of taking the mean over the domain D, i.e., $D[f] = \int_D f$. First, we modify the model so as to have the property $[u] = 0$ preserved in time; the property $[v] = 0$ is automatically preserved. Second, we replace the nonlinear surface tension term $-\gamma\nabla\cdot\mathbf{n}$, where $\mathbf{n} = (-\nabla_{\mathbf{x}} v, 1)/\sqrt{1 + |\nabla_{\mathbf{x}} v|^2}$ on the free surface, by $\gamma\nabla_{\mathbf{x}}^2 v$, its linearization about the undisturbed solution $v \equiv 0$. In Lin's paper one finds both modifications; however, in his paper the second modification is done purely for convenience. In fact, because his spaces $K(r,s)$ satisfy $K(r,0) \subseteq K(r,2)$ for $r > 2$, he can treat the nonlinear surface tension term as easily as its linearization. In our formulation, it seems we must make this second modification, otherwise the spaces E and F, described in §4, will not be suitable in §6 for $\mathbf{f} : F \to E$. It is not clear if this second modification is acceptable viz. physical experiments, for small oscillations.

In terms of v, u, w these modifications of (2.4) and (2.5) yield

$$\begin{cases} v_t = w + M_1(v,u), & (2.6)\\ u_t = \gamma\nabla_{\mathbf{x}}^2 v + \mu\nabla_{\mathbf{x}}^2 u - (g - \alpha(t))v + M_2(v,u) - [M_2(v,u)], & (2.7)\end{cases}$$

where

$$M_1(v,u) = w|\nabla_{\mathbf{x}} v|^2 - \nabla_{\mathbf{x}} u \cdot \nabla_{\mathbf{x}} v,$$

$$M_2(v,u) = -\frac{1}{2}\,|\nabla_{\mathbf{x}} u|^2 + \frac{1}{2}\,w^2(1 + |\nabla_{\mathbf{x}} v|^2).$$

We take as boundary conditions $\frac{\partial u}{\partial\mathbf{n}} = 0$ on ∂D, which follows from (2.3), and $\frac{\partial v}{\partial\mathbf{n}} = 0$ on ∂D. The latter is somewhat controversial; in fact, one could argue that instead of an artificial dissipation term $\mu\nabla_{\mathbf{x}}^2$ in equation (2.5), one should introduce damping in the boundary conditions. This has been brought to my attention by M. Silber. Recently, Simonelli and Gollub [24] have described experiments in which the free surface is at right-angle contact with the sidewalls of a rectangular container, and this is consistent with $\frac{\partial v}{\partial \mathbf{n}} = 0$ on ∂D. Douady [10] has described experiments in which the free surface is pinned at the boundary, i.e., $v = 0$ on ∂D, by the use of felt on the walls of the container. It appears that this boundary condition cannot be accommodated by our abstract framework.

To linearize equations (2.6) and (2.7), we need only replace w by $N(0)u$ in equation (2.6) and replace M_1 and M_2 by 0. To this end, let $A := -\Delta_N$ denote the Neumann Laplacian on the complex Hilbert space $\tilde{L}^2(D) = \{u \in L^2(D) : \int_D u = 0\}$, with domain $\mathcal{D}(A) = \{u \in \tilde{L}^2(D) : Au \in \tilde{L}^2(D), \frac{\partial u}{\partial \mathbf{n}} = 0 \text{ on } \partial D\}$. Throughout, assume that $D \subseteq \mathbb{R}^2$ is bounded and open and has the uniform C^m regularity property, as in Adams [1, p. 67], with m as large as needed later. It is well known that $A : \mathcal{D}(A) \subseteq \tilde{L}^2(D) \to \tilde{L}^2(D)$ is a strictly positive definite, self-adjoint linear operator with compact inverse. Let $\{\kappa_n^2\}$ be the eigenvalues of A and $\psi_n = \psi_n(\mathbf{x})$ be the corresponding orthonormalized eigenfunctions; without loss of generality, $\kappa_1^2 \leq \kappa_2^2 \leq \cdots$. As in Henry [12] or Pazy [21], one can define the fractional powers A^β, which in this situation are closed linear operators with domain $X^\beta := \{u \in \tilde{L}^2(D) : \sum_{n=1}^\infty \kappa_n^{4\beta}|(u, \psi_n)_{L^2}|^2 < \infty\}$. From now on we will write $\sum_n$ instead of $\sum_{n=1}^\infty$ and denote $u_n = (u, \psi_n)_{L^2(D)}$. X^β is a Hilbert space when given the inner product $(u, v) = \sum_n \kappa_n^{4\beta} u_n \bar{v}_n$.

From inspection of (2.1)–(2.3), Lin [15, §4] obtained the explicit result that

$$N(0)u = \sum_n \kappa_n \tan h(\kappa_n h) u_n \psi_n \quad \text{for } u = \sum_n u_n \psi_n.$$

It follows that for any $\beta, N(0) : X^\beta \to X^{\beta-(1/2)}$ is a bounded linear operator. In fact $N(0)$ is close to being $A^{1/2} = (-\Delta_N)^{1/2}$. To be precise, $N(0) = A^{1/2} + A_1$, where

$$A_1 u = \sum_n \kappa_n (\tan h(\kappa_n h) - 1) u_n \psi_n.$$

Since $0 < \kappa_n \to \infty$ as $n \to \infty, A_1 : X^\beta \to X^\beta$ is a bounded linear operator. Thus, the linearization of (2.6)–(2.7) about $u = v = 0$ is

$$\text{(2.8)} \qquad \frac{d}{dt}\begin{pmatrix} v \\ u \end{pmatrix} = B\begin{pmatrix} v \\ u \end{pmatrix} + B_1(t)\begin{pmatrix} v \\ u \end{pmatrix},$$

where

$$\text{(2.9)} \qquad B = \begin{pmatrix} 0 & A^{1/2} \\ -\gamma A & -\mu A \end{pmatrix}, \qquad B_1(t) = \begin{pmatrix} 0 & A_1 \\ -(g - \alpha(t))I & 0 \end{pmatrix}.$$

Our plan for the rest of the paper is, first, to analyze the semigroup of linear operators e^{tB} on spaces which will be suitable for subsequent analysis; second, to express the solution of (2.8) in terms of a family of evolution operators; third, to establish a local existence and uniqueness theorem for the nonlinear problem (2.6)–(2.7); and finally, to establish a local center manifold theorem for (2.6)–(2.7).

3. The semigroup e^{tB}. First, we analyze the semigroup e^{tB} on the Hilbert space $\mathbf{X} := X^\beta \times X^{\beta-(1/2)}$, for all $\beta \geq \frac{1}{2}$. We will be more specific about β when discussing the nonlinear problem in §6. Other choices of $\mathbf{X}$ are possible but less convenient for our subsequent analysis of $\mathcal{D}(B), \mathcal{D}(B^2)$, and the nonlinear problem. Choose $\mathcal{D}(B) = X^{\beta+(1/2)} \times X^{\beta+(1/2)}$, which is itself a Hilbert space. From (2.9) $B : \mathcal{D}(B) \subset \mathbf{X} \to \mathbf{X}$ is a closed, densely defined linear operator. In addition, B has a compact resolvent, as one can see by explicit calculation of $(\lambda I - B)^{-1}$ as the operator norm limit of natural finite-rank operators. The spectrum of B consists of those $\lambda \in \mathbb{C}$ which are eigenvalues and thus solve, for some $n \geq 1$,

$$\lambda^2 + \mu\kappa_n^2\lambda + \gamma\kappa_n^3 = 0.$$

For future reference, we note that $0 \notin \sigma(B)$. Let

$$\lambda_n^{\pm} := \frac{1}{2}\,\kappa_n^2(-\mu \pm \delta_n), \delta_n := [\mu^2 - 4\gamma\kappa_n^{-1}]^{1/2}.$$

The corresponding eigenfunctions of B are

$$\begin{pmatrix} \mu \pm \delta_n \\ -2\gamma \end{pmatrix} \psi_n.$$

For convenience, assume $\delta_n \neq 0$ for all n. If $\delta_n = 0$ for a single n, the forms of e^{tB} and Be^{tB} would be altered, but the results, (3.3) and following, would not be affected. By diagonalizing B in each of the subspaces spanned by

$$\left\{ \begin{pmatrix} \psi_n \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \psi_n \end{pmatrix} \right\},$$

after much calculation one finds that

$$e^{tB}\begin{pmatrix} v \\ u \end{pmatrix} = \sum_n \delta_n^{-1}(e^{t\lambda_n^+}C_n^+ + e^{t\lambda_n^-}C_n^-)\begin{pmatrix} v_n \\ u_n \end{pmatrix}\psi_n, \tag{3.1}$$

where $v_n := (v, \psi_n) = \int_D v\psi_n, u_n := (u, \psi_n)$, and

$$C_n^+ := \begin{pmatrix} \mu(\nu_n^+)^{-1} & \kappa_n^{-1} \\ -\gamma & -\frac{\gamma}{\mu}\nu_n^+\kappa_n^{-1} \end{pmatrix}, \qquad C_n^- := \begin{pmatrix} -\frac{\gamma}{\mu}(\nu_n^-)^{-1}\kappa_n^{-1} & -\kappa_n^{-1} \\ \gamma & \mu\nu_n^- \end{pmatrix}, \tag{3.2}$$

and $\nu_n^+ := -\mu\lambda_n^+/(\gamma\kappa_n), \nu_n^- := -\lambda_n^-/(\mu\kappa_n^2)$. One notes that because $\kappa_n \to \infty$ as $n \to \infty, \lim_{n\to\infty} \nu_n^{\pm} = 1$ and, in fact, $1 \leq \nu_n^+ \leq 2$ and $0 < \nu_n^- \leq 1$ for all n, so (3.1) and (3.2) will be convenient for subsequent analysis.

The norm on $\mathbf{X} = X^{\beta} \times X^{\beta-(1/2)}$ is given by

$$\left\| \begin{pmatrix} v \\ u \end{pmatrix} \right\|_{\mathcal{X}}^2 = \sum_n ((\kappa_n^{2\beta}|v_n|)^2 + (\kappa_n^{2\beta-1}|u_n|)^2),$$

so one sees immediately that (3.1) defines a C^0, i.e., strongly continuous, semigroup e^{tB} of bounded linear operators on $\mathbf{X}$. We note that, because $\kappa_n \to \infty$ as $n \to \infty, \sigma(B)$ consists of two infinite sequences of real numbers tending to $-\infty$ and possibly a finite number of complex conjugate pairs of eigenvalues $\{\lambda_n^{\pm}\}_{1 \leq n \leq n_0}$, all with Re $\lambda_n^{\pm} < 0$, so $\{e^{tB}\}_{t \geq 0}$ is uniformly bounded. We calculate that

$$Be^{tB}\begin{pmatrix} v \\ u \end{pmatrix} = \sum_n \delta_n^{-1}(\lambda_n^+ e^{t\lambda_n^+}C_n^+ + \lambda_n^- e^{t\lambda_n^-}C_n^-)\begin{pmatrix} v_n \\ u_n \end{pmatrix}\psi_n,$$

so there exists a constant M such that

$$\|tBe^{tB}\| \leq M \quad \text{for } t \geq 0. \tag{3.3}$$

In addition, e^{tB} is differentiable for $t > 0$ and $0 \notin \sigma(B)$. It follows from Pazy [21, Thm. 2.5.2] that (i) $\{e^{tB}\}_{t \geq 0}$ is an analytic semigroup of bounded linear operators on $\mathbf{X}$ and (ii) $-B$ is a sectorial operator, in the sense of Henry [12].

4. Interpolation spaces for autonomous problems. Throughout this section, θ is a fixed real number, $0 < \theta < 1$.

In the approach of Henry [12] to nonlinear problems and invariant manifolds, for a problem $\frac{dw}{dt} = Bw + f(w), w \in \mathbf{X}$, if $-B$ is a sectorial operator generating an analytic semigroup e^{tB}, then one defines $\mathbf{X}^\rho = \mathcal{D}((-B)^\rho)$ and hypothesizes that $f : \mathbf{X}^\rho \to \mathbf{X}$ for some $0 \leq \rho < 1$. Unfortunately, the nonlinear problem of ours has a specific nonlinearity which does not allow this approach. From Lin [15], we know that our nonlinearity takes $H^r(D) \times H^r(D)$ into $H^{r-1}(D) \times H^{r-1}(D)$ as long as $r > 2$; see Theorem 6.1 below. It seems that if $\mathbf{X} = X^{\beta_1} \times X^{\beta_2}$, whenever we choose $\beta_1, \beta_2 > \frac{1}{2}$, in order to have $H^{r-1}(D) \times H^{r-1}(D) \subseteq \mathbf{X}$, it follows that to have $\mathbf{X}^\rho \subseteq H^r(D) \times H^r(D)$, one must take $\rho = 1$. For example, the calculations of Chen and Triggiani [5] and Rodriguez-Bernal [22] confirm this for many choices of β_1, β_2; the details are omitted except to note that one can introduce $x = v$ and $y = A^{1/2}u$ and define operators

$$\mathcal{A} = A^{3/2} \quad \text{and} \quad B = \begin{pmatrix} 0 & I \\ -\gamma\mathcal{A} & -\mu\mathcal{A}^{2/3} \end{pmatrix}.$$

In addition, the form of e^{tB} leads one to suspect that, for all choices of β_1 and β_2, one must take $\rho = 1$.

We can use another approach using interpolation spaces, as in daPrato and Grisvard [8] and many others, including Butzer and Berens [4], Sinestrari [25], daPrato and Lunardi [9], and Angenent [2]. We will denote $E := D_B(\theta)$, to be defined below, and $F := D_B(\theta + 1) = \{w \in \mathcal{D}(B) : Bw \in D_B(\theta)\}$, and we will show in later sections, first, that the linear periodic problem (2.8) has some nice properties with regard to these spaces E and F and, second, that the nonlinearities in problem (2.6)–(2.7) take F into E nicely. The outstanding property of the spaces E and F is "maximal regularity," which has been put to good use by daPrato and Lunardi [9] to establish a center manifold theorem for autonomous problems.

We follow the exposition by Angenent [2] of the daPrato and Grisvard construction [8]. Suppose E_0 and E_1 are real Banach spaces with E_1 densely included in E_0 and $B : E_1 \to E_0$ is bounded and linear. Considered as an unbounded operator on E_0, we assume that B generates an analytic semigroup e^{tB}. Throughout this section, θ is a fixed real number, $0 < \theta < 1$. We define E_θ to be a space of traces, i.e., initial values, of a certain class of functions

$$Y_\theta := \{y \in C((0,1]; E_1) \cap C^1((0,1]; E_0 : \lim_{t \downarrow 0} t^{1-\theta}(\|y'(t)\|_{E_0} + \|y(t)\|_{E_1}) = 0\}$$

with norm

$$\|y\|_\theta := \sup_{0<t\leq 1} t^{1-\theta}(\|y'(t)\|_{E_0} + \|y(t)\|_{E_1}).$$

Specifically,

$$E_\theta := \{w : w = y(0) \text{ for some } y \in Y_\theta\}$$

with norm

$$\|w\|_\theta = \inf\{\|y\|_\theta : y \in Y_\theta, y(0) = w\}.$$

It is known that E_θ is a Banach space with this norm and that $E_\theta = D_B(\theta)$ as defined in daPrato and Lunardi [9]. Similarly, between the spaces E_1 and $E_2 := \mathcal{D}_{E_1}(B) = \{w \in E_1 : Bw \in E_1\}$, one can define $E_{1+\theta}$. If $E_0 = \mathbf{X}$ and $E_1 = \mathcal{D}(B)$, then $E_2 = \mathcal{D}(B^2), E_{1+\theta} = D_B(\theta+1)$, and $\mathcal{D}(B^2) \subseteq D_B(\theta+1) \subseteq \mathcal{D}(B) \subseteq D_B(\theta) \subseteq \mathbf{X}$, as in daPrato and Grisvard [8] and Sinestrari [25].

These interpolation spaces are defined by using continuous functions, so in the general theory of interpolation, $E_\theta \subseteq (E_1, E_0)_{\theta,\infty}$. The interpolation in Lions and Magenes [16] uses L^2 functions and can be notated $(\quad,\quad)_{1-\theta,2}$.

One should note that, in our problem, $E_0 = \mathbf{X}$ and $E_1 = \mathcal{D}(B)$ are spaces of functions of $\mathbf{x} \in D$, as are $D_B(\theta)$ and $D_B(\theta + 1)$; the time variable appears only in the definition of these spaces as initial values of functions of a time variable.

Because B generates an analytic semigroup e^{tB}, there exists ω, M such that $\|e^{tB}\|_{L(\mathbf{X})} \leq Me^{\omega t}$, where $L(\mathbf{X})$ denotes the space of bounded linear operators from $\mathbf{X}$ to $\mathbf{X}$. It is known that

$$\|e^{tB}\|_{L(Z)} \leq Me^{\omega t} \tag{4.1}$$

and that $-B$ is sectorial on Z, for any choice of Z being $\mathbf{X}, \mathcal{D}(B), D_B(\theta)$, or $D_B(\theta+1)$, although it might be necessary to increase M.

In our specific problem, the spectrum of B consists of $\{\lambda_n^\pm\}_{n\geq 1}$, where

$$\lambda_n^\pm = -\frac{1}{2}\,\kappa_n^2(\mu \mp \sqrt{\mu^2 - 4\gamma\kappa_n^{-1}}.$$

Because $0 < \kappa_1 \leq \kappa_2 \leq \cdots$, there exists $\omega < 0$ such that Re $\sigma(B) < \omega$ and there exists $M \geq 1$ such that (4.1) holds with that $\omega < 0$.

In our problem, we take $E_0 = \mathbf{X} = X^\beta \times X^{\beta-(1/2)}$ and $E_1 = \mathcal{D}(B) = X^{\beta+(1/2)} \times X^{\beta+(1/2)}$. One can calculate that $E_2 = \mathcal{D}(B^2) = \{w \in \mathcal{D}(B) : Bw \in \mathcal{D}(B)\} = X^{\beta+1} \times X^{\beta+1} \cap \{(v,u) : \gamma v + \mu u \in X^{\beta+(3/2)}\}$. The latter expresses an "interaction condition." One can observe that interpolation commutes with the operation of taking direct products, so that

$$\begin{aligned} E_\theta &= (X^{\beta+1/2} \times X^{\beta+1/2}, X^\beta \times X^{\beta-1/2})_{\theta,\infty} \\ &= (X^{\beta+1/2}, X^\beta)_{\theta,\infty} + (X^{\beta+1/2}, X^{\beta-1/2})_{\theta,\infty}. \end{aligned}$$

We recall that $X^\beta = \mathcal{D}(A^\beta)$ where $A = -\Delta_N$ on $\tilde{L}^2(D)$.

It is known from work of Grisvard [11] that, regarding $0 < \beta < 1$,

$$X^\beta = \left\{ \begin{array}{ll} \tilde{H}^{2\beta}(D), & 0 < \beta < \frac{3}{4} \\ \tilde{H}^{2\beta}(D) \cap \{u : \frac{\partial u}{\partial \mathbf{n}} = 0 \quad \text{on} \quad \partial D\}, & \frac{3}{4} < \beta < 1 \end{array} \right\},$$

where $H^s(D)$ is the Sobolev space $W^{s,2}(D)$ and, as before, a $\tilde{}$ indicates "mean value zero." Put loosely, in X^β the boundary condition $\frac{\partial u}{\partial \mathbf{n}} = 0$ applies if it makes sense, i.e., if $\nabla_{\mathbf{x}} u|_{\partial D}$ makes sense as a trace of a distribution on ∂D. For $1 < \beta < 2$,

$$\begin{aligned} X^\beta &= \{u \in X^{\beta-1} : \Delta_N u \in X^{\beta-1}\} \\ &= \left\{ \begin{array}{ll} \tilde{H}^{2\beta} \cap \{u : \frac{\partial u}{\partial \mathbf{n}} = 0 \quad \text{on} \quad \partial D\}, & 1 < \beta < \frac{7}{4} \\ \tilde{H}^{2\beta} \cap \{u : \frac{\partial u}{\partial \mathbf{n}} = \frac{\partial \Delta u}{\partial n} = 0 \quad \text{on} \quad \partial D\}, & \frac{7}{4} < \beta < 2 \end{array} \right\}. \end{aligned}$$

It is known from daPrato and Grisvard [8, §6.1] that

$$(X^{\beta+1/2}, X^{\beta})_{\theta,\infty} \subseteq \{u \in X^{\beta} : A^{\beta}u \in h_2^{\theta}\} \tag{4.2}$$

and

$$(X^{\beta+1/2}, X^{\beta-1/2})_{\theta,\infty} \subseteq \{u \in X^{\beta-1/2} : A^{\beta-1/2}u \in h_2^{2\theta}\}. \tag{4.3}$$

Here $h_2^s(D)$ stands for a Nikolsk'ii space, as in Nikolsk'ii [19], [20], Slobodeckii [26], and Adams [1]. These spaces are actually a subclass of the Besov spaces. The only property of these spaces we will use are the continuous imbeddings

$$H^s(D) \hookrightarrow h^s(D) \hookrightarrow H^{s-\varepsilon}(D) \tag{4.4}$$

for any $\varepsilon > 0$. We will use the property that interpolation preserves bounded linear operators in order to avoid the need for detailed examination of the Nikolsk'ii spaces.

For the moment, we leave our specific operator B and specific space $\mathbf{X}$ and return now to the general situation. We denote $E = E_\theta = D_B(\theta)$ and $F = E_{1+\theta} = D_B(\theta+1)$. The maximal regularity property mentioned above is as follows:

$$\text{for all } f \in C([0,\tau); E), \quad \int_0^{\cdot} e^{(\cdot-s)B} f(s)\, ds \in C([0,\tau]; F).$$

Proofs can be found in daPrato and Grisvard [8], Sinestrari [25], and Angenent [2], among others. We apply results of this sort to our specific operator B and space $\mathbf{X}$ in Propositions 4.1 and 4.2 below.

Many proofs of existence of invariant manifolds explicitly or implicitly involve weighted spaces; see, for example, Chow, Lin, and Lu [7] and van Gils and Vanderbauwhede [28]. For a Banach space $Z, \sigma \in \mathbb{R}, \eta \in \mathbb{R}$, define

$$C_\eta((-\infty,\sigma]; Z) = \{f : f \text{ continuous on } (-\infty,\sigma] \text{ and } \|f\|_{\eta,\sigma,-,Z} < \infty\},$$

where its norm is

$$\|f\|_{\eta,\sigma,-,Z} = \sup_{t\leq\sigma} e^{-\eta t}|f(t)|_Z.$$

Similarly define

$$C_\eta([\sigma,\infty); Z) = \{f : f \text{ continuous on } [\sigma,\infty) \text{ and } \|f\|_{\eta,\sigma,+,Z} < \infty\},$$

where

$$\|f\|_{\eta,\sigma,+,Z} = \sup_{t\geq\sigma} e^{\eta t}|f(t)|_Z.$$

Recall that in our problem $\|e^{tB}\|_{L(Z)} \leq Me^{\omega t}$, for some $\omega < 0$ and $M \geq 1$, from (4.1), where Z is any of $\mathbf{X}, \mathcal{D}(B), D_B(\theta), D_B(\theta+1)$. Here, θ is a fixed real number with $0 < \theta < 1$. We apply two results of daPrato and Lunardi [9, pp. 118–120] to our specific B and $\mathbf{X}$ to get the following propositions.

PROPOSITION 4.1. *With B as in (2.9), $\mathbf{X} = X^\beta \times X^{\beta-(1/2)}, \mathcal{D}(B) = X^{\beta+(1/2)} \times X^{\beta+(1/2)}$, and M, ω as in (4.1), choose any η such that $\omega + \eta < 0$. Then there exists $K_+ = K_+(\eta, \omega, \sigma, \theta)$ such that $f(\cdot) \mapsto \int_\sigma^{\cdot} e^{(\cdot-s)B} f(s)\, ds$ defines a bounded linear operator $\mathcal{L}_+ : C_\eta([\sigma,\infty); E) \to C_\eta([\sigma,\infty); F)$ with $\|\mathcal{L}_+\| \leq K_+$, where $E = D_B(\theta), F = D_B(\theta+1)$.*

PROPOSITION 4.2. *With B, M and ω, as in Proposition* 4.1, *choose any η such that $\omega - \eta < 0$. Then there exists $K_- = K_-(\eta, \omega, \sigma, \theta)$ such that*

$$f(\cdot) \mapsto \int_{-\infty}^{\cdot} e^{(\cdot - s)B} f(s)\, ds$$

defines a bounded linear operator $\mathcal{L}_- : C_\eta((-\infty, \sigma]; E) \to C_\eta((-\infty, \sigma]; F)$ with $\|\mathcal{L}_-\| \leq K_-$, where $E = D_B(\theta)$ and $F = D_B(\theta + 1)$.

These two propositions express a maximal regularity property useful for the proof of the existence of an invariant manifold for an autonomous problem. In the next section, these two propositions will imply similar properties for a periodic problem.

5. A family of evolution operators for equation (2.8). Given two normed linear spaces Z_1 and $Z_2, L(Z_1, Z_2)$ denotes the space of bounded linear operators with the operator norm.

In this section, we establish a new result for periodic problems considered in the framework of interpolation spaces.

DEFINITION. A family of evolution operators $\{\Phi(t,s)\}$ *consists of bounded linear operators on a space Z for $t \geq s$ satisfying $\Phi(s,s) = I, \Phi(t_1, s)\Phi(s, t_2) = \Phi(t_1, t_2)$ for all $t_1 \geq s \geq t_2$ in $\mathbb{R}$, and $\{\Phi(t,s) : t \geq s\}$ is strongly continuous in (t,s) with values in $L(Z)$.*

Assume $-B$ is sectorial on a Banach space .and $t \mapsto B_1(t) : \mathbb{R} \to L(E^\rho, E)$ is Hölder-continuous with exponent < 1. The work of Henry [12, §7.1] shows that the solutions of

$$\left\{ \begin{array}{l} \frac{dw}{dt} = Bw + B_1(t)w, \\ w(s) = w_0 \in E \end{array} \right\} \tag{5.1}$$

are given by $w(t) = \Phi(t,s)w_0$, where $\{\Phi(t,s)\}$ is a family of evolution operators which satisfies

$$\Phi(t,s) = e^{(t-s)B} + \int_s^t e^{(t-\tau)B} B_1(\tau)\Phi(\tau, s)\, d\tau \tag{5.2}$$

for $t \geq s$. Moreover, if $B_1(\cdot)$ is periodic with period T, then $\Phi(t+T, s+T) = \Phi(t,s)$ for $t \geq s$.

In our particular problem (2.8), we can take $\rho = 0$, i.e., $E^\rho = E$.

For the periodic problem (5.1), one can define the period map $U(t) = \Phi(t+T, t)$. In our problem, B has compact resolvent on $\mathbf{X}$. The same argument, approximation in the uniform operator norm by a sequence of finite rank operators, shows that B has compact resolvent on $\mathcal{D}(B)$. By interpolation, we see that B has compact resolvent on $E := D_B(\theta)$, where we fix any $\theta \in (0,1)$. From Henry [12, §7.2], the nonzero eigenvalues of $U(t)$ are independent of time t and constitute all of the spectrum of $U(t)$ with the exception of 0, which is in the continuous spectrum of $U(t)$ since it is a compact operator.

PROPOSITION 5.1 (Henry [12, Thm. 7.2.3 et seq.]). *If all of the characteristic multipliers of the problem* (2.8), *i.e ., problem* (5.1) *in our specific case, are of modulus less than* 1 *and bounded away from* 1, *then there exist $\tilde{M} \geq 1$ and $\tilde{\omega} < 0$ such that*

$$\|\Phi(t,s)\|_{L(E,E^\rho)} \leq \tilde{M}(t-s)^{-\rho} e^{\tilde{\omega}(t-s)}, \quad \text{for } t > s. \tag{5.3}$$

Proof. It suffices to show that $t \mapsto B_1(t) : [0,T] \to L(E^\rho, E)$ is Hölder continuous. Looking at (2.9) we see that $t \mapsto B_1(t) : [0,T] \to L(\mathcal{X})$ and $t \mapsto B_1(t) :$

$[0,T] \to L(\mathcal{D}(B))$ are Hölder continuous, so the desired result follows by interpolation of bounded linear operators. □

PROPOSITION 5.2. *Assume that $B_1(\cdot)$ is periodic with period $T, B_1(\cdot) : [0,T] \to L(E^\rho, E)$ is Hölder continuous, and (5.3) holds, apart from any assumption as to the sign of $\tilde{\omega}$. If η is chosen so that $\tilde{\omega}+\eta < 0$ and $\tilde{\omega}-\eta < 0$, then, for all $\sigma \in \mathbb{R}$, the map*

$$f(\cdot) \mapsto \int_{-\infty}^{\cdot} \Phi(\cdot, s) f(s)\, ds$$

defines a bounded linear operator

$$\mathcal{L}_- : C_\eta((-\infty,\sigma]; E) \to C_\eta((-\infty,\sigma]; F)$$

and the map

$$f(\cdot) \mapsto \int_{\sigma}^{\cdot} \Phi(\cdot, s) f(s)\, ds$$

defines a bounded linear operator

$$\mathcal{L}_+ : C_\eta([\sigma,\infty); E) \to C_\eta([\sigma,\infty); F).$$

Here, $E = D_B(\theta)$ and $F = D_B(\theta+1)$, as in §4.

Proof. We prove the result for $\mathcal{L}_-$; the result for $\mathcal{L}_+$ can be proved similarly. Fix any $f \in C_\eta((-\infty,\sigma]; E)$ and define $u(\cdot) = \int_{-\infty}^{\cdot} \Phi(\cdot, s) f(s)\, ds$. From (5.2),

$$u(t) = \int_{-\infty}^{t} e^{(t-s)B} f(s)\, ds + \int_{-\infty}^{t} \int_{s}^{t} e^{(t-\tau)B} B_1(\tau)\Phi(\tau,s)\, d\tau f(s)\, ds.$$

The first term is in $C_\eta((-\infty,\sigma]; F)$, by Proposition 4.2, with

$$\left\| \int_{-\infty}^{\cdot} e^{(\cdot - s)B} f(s)\, ds \right\|_{\eta,\sigma,-,F} \le K_- \|f\|_{\eta,\sigma,-,E}.$$

Rewrite the second term as

$$\int_{-\infty}^{t} e^{(t-s)B} \int_{-\infty}^{\tau} B_1(\tau)\Phi(\tau,s) f(s)\, ds\, d\tau := \int_{-\infty}^{t} e^{(t-\tau)B} g(\tau)\, d\tau.$$

Using Proposition 4.2 again, to complete the proof it will suffice to show that $g \in C((-\infty,\sigma]; E)$ and that there is a constant C, independent of f, such that $\|g\|_{\eta,\sigma,-,E} \le C\|f\|_{\eta,\sigma,-,E}$. We have for all $\tau \le \sigma$, denoting $m = \max_{0\le\tau\le T} \|B_1(\tau)\|_{L(E^\rho,E)}$,

$$\begin{aligned} e^{-\eta\tau} \left| \int_{-\infty}^{\tau} B_1(\tau)\Phi(\tau,s) f(s)\, ds \right|_E &\le m\tilde{M} \int_{-\infty}^{\tau} (\tau - s)^{-\rho} e^{\tilde{\omega}(\tau-s)} e^{-\eta(\tau-s)} e^{-\eta s} |f(s)|_E\, ds \\ &\le m\tilde{M} \|f\|_{\eta,\sigma,-,E} \cdot \int_{-\infty}^{\tau} (\tau-s)^{-\rho} e^{(\tilde{\omega}-\eta)(\tau-s)}\, ds. \end{aligned}$$

Since $\rho < 1$, $\int_{-\infty}^{\tau} (\tau-s)^{-\rho} e^{(\tilde{\omega}-\eta)(\tau-s)}\, ds < \infty$, and the proposition is proven. □

In our problem we will only use $\rho = 0$, i.e., $E^\rho = E$.

COROLLARY 5.3 (to Proposition 5.1). *For the linear homogeneous periodic problem (2.8), if $\|\alpha\|_\infty := \max_{0\le t\le T} |\alpha(t)|$ is sufficiently small, then the zero solution*

$v \equiv u \equiv 0$ is exponentially asymptotically stable in Z, where Z is any of the spaces $\mathbf{X}, \mathcal{D}(B), E$, and F.

Proof. Problem (2.8) defines $\Phi(t,s)$, which can be considered as a block diagonal matrix $(\Phi_n(t,s))_{n=1}^{\infty}$, where $\Phi_n(t,s)$ is a 2×2 principal fundamental matrix for the problem

$$\begin{pmatrix} \dot{v}_n \\ \dot{u}_n \end{pmatrix} = \begin{pmatrix} 0 & \kappa_n \tan h(\kappa_n h) \\ -\gamma\kappa_n^2 - (g - \alpha(t)) & -\mu\kappa_n^2 \end{pmatrix} \begin{pmatrix} v_n \\ u_n \end{pmatrix}. \tag{5.4}$$

It will suffice to show that the characteristic multipliers $\mu_{n,j}, j = 1,2, n \geq 1$ of (5.4) (i) go to zero as $n \to \infty$ for all $\|\alpha\|_\infty$ and (ii) have modulus less than 1 for all n if $\|\alpha\|_\infty$ is sufficiently small.

Fix an n and denote $c_n = \mu\kappa_n^2, a_n = (\gamma\kappa_n^2 + g)\kappa_n \tan h(\kappa_n h)$. Rescale the time variable by $\tau = c_n t$ and apply Grönwall's inequality to (5.4) with $\alpha \neq 0$ considered as a perturbation of (5.4) with $\alpha \equiv 0$ to conclude that

$$|\mu_n^{\pm}| < \left(-\frac{c_n}{2} + \chi_n + m \frac{\|\alpha\|_\infty}{c_n} \right), \tag{5.5}$$

where $\chi_n := 0$, if $c_n^2 \leq 4a_n$, and $\chi_n := \delta_n$, if $c_n^2 \geq 4a_n$, where $\delta_n := \frac{1}{2}\sqrt{c_n^2 - 4a_n}$, and m is a constant uniformly bounded in n. Results (i) and (ii) follow from easy but tedious asymptotics in (5.5), using $\kappa_n \to \infty$ as $n \to \infty$. □

6. The nonlinearities in equations (2.6)–(2.7). Fix a $\theta \in (0,1)$ and denote $E = D_B(\theta), F = D_B(\theta + 1)$, where B is in equation (2.9). We can write equations (2.6)–(2.7) abstractly as

$$\frac{d\mathbf{w}}{dt} = (B + B_1(t))\mathbf{w} + \mathbf{f}(\mathbf{w}), \tag{6.1}$$

where $\mathbf{f}(\mathbf{w}) = (f_1(\mathbf{w}), f_2(\mathbf{w})), \mathbf{w} = (v,u)$. The purpose of this section is to show that $\mathbf{f} : F \to E$ is C^∞, using a result of Lin [15].

Recall that

$$f_1(v,u) = w|\nabla_{\mathbf{x}} v|^2 - \nabla_{\mathbf{x}} u \cdot \nabla_{\mathbf{x}} v, \tag{6.2}$$

$$f_2(v,u) = M_2(v,u) - [M_2(v,u)], \qquad [\ \ ] = \text{ mean value over } D, \tag{6.3}$$

$$M_2(v,u) = -\frac{1}{2}\, |\nabla_{\mathbf{x}} u|^2 + \frac{1}{2}\, w^2(1 + |\nabla_{\mathbf{x}} v|^2),$$

and

$$w = N(v)u = \phi_z(\mathbf{x}, v(\mathbf{x},t), t)$$

is constructed from the solution ϕ of equations (2.1)–(2.3).

THEOREM 6.1 (Lin [15, Thm. 4.3, "fixed t" part of the proof]). *Fix any $r > 2$. For sufficiently small $|v|_{H^r(D)}$, the map $(v,u) \mapsto N(v)u : \tilde{H}^r(D) \times \tilde{H}^r(D) \to H^r(D)$ exists and is C^∞.*

The proof consists of mapping the region $\Omega_v := \{(\mathbf{x}, z) : -h < z < v(\mathbf{x}), \mathbf{x} \in D\}$ diffeomorphically to $\Omega_0 := \Omega_v|_{v\equiv 0}$ and solving a perturbation of Laplace's equation in Ω_0. The diffeomorphism exists because $|v|_{H^r(D)}$ is small and $r > 2$; the perturbation is small because $|v|_{H^r(D)}$ is, so the implicit function theorem can be applied. The proof uses a result of Zolesio [29] on multiplication of elements of Sobolev spaces of fractional order.

To show $\mathbf{f} : F \to E$ is C^∞, we use the following theorem.

THEOREM 6.2 (daPrato and Grisvard [8, Thm. 2.4]). *Suppose* $0 < \theta < 1, \mathbf{f} : X_1 \to X_2$ *is Fréchet differentiable and* $\mathbf{f}|_{Y_1} : Y_1 \to Y_2$ *is uniformly Lipschitz, where* $Y_i \hookrightarrow X_i$ *are continuously imbedded Banach spaces. Then* $\mathbf{f} : (Y_1, X_1)_{\theta;\infty} \to (Y_2, X_2)_{\theta;\infty}$ *is continuous.*

COROLLARY 6.3. *Suppose* $0 < \theta < 1, E = D_B(\theta), F = D_B(\theta+1)$, $\mathcal{U}_1$ *is an open neighborhood of* $0 \in \mathcal{D}(B)$ *and* $\mathcal{U}_2$ *is an open neighborhood of* $0 \in \mathcal{D}(B^2)$. *If* $\mathbf{f} : \mathcal{U}_1 \subset \mathcal{D}(B) \to \mathcal{X}$ *and* $\mathbf{f}|_{\mathcal{U}_2} : \mathcal{U}_2 \subset \mathcal{D}(B^2) \to \mathcal{D}(B)$ *is* C^∞, *then so is* $\mathbf{f} : \mathcal{V} \subset F \to E$, *for some* $\mathcal{V}$ *an open neighborhood of* $0 \in F$.

COROLLARY 6.4. *Suppose* $0 < \theta < 1, E = D_B(\theta), F = D_B(\theta+1), \mathbf{X} = X^\beta \times X^{\beta-(1/2)}, \mathcal{D}(B) = X^{\beta+(1/2)} \times X^{\beta+(1/2)}, \beta = \frac{r-1}{2}$, *and* $r > 2$. *If the region* D *is a rectangle, then there is an open neighborhood* $\mathcal{V}$ *of* $0 \in F$ *such that* $\mathbf{f}$ *defined by* (6.1)–(6.2) *satisfies* $\mathbf{f} : \mathcal{V} \to E$ *is* C^∞.

Proof. First of all, if $\beta = \frac{r-1}{2}$, then $\mathcal{D}(B) \subset \tilde{H}^r(D) \times \tilde{H}^r(D)$ and $\mathcal{D}(B^2) = (X^{\beta+1} \times X^{\beta+1}) \cap \{v, u) : \gamma v + \mu u \in H^{\beta+(3/2)}\} \subset \tilde{H}^{r+1}(D) \times \tilde{H}^{r+1}(D)$. By Theorem 6.1, for sufficiently small $\delta > 0, \mathbf{f}$ maps $\mathcal{U}_1 = \{(v,u) \in \mathcal{D}(B) : |v|_{\tilde{H}^r(D)} < \delta\}$ into $H^{r-1}(D) \times H^{r-1}(D)$, in fact, it maps into $\tilde{H}^{r-1}(D) \times \tilde{H}^{r-1}(D)$ because clearly $[f_2] = 0$ and $[f_1] = 0$ is noted in Lin [15, §6]. Again the result on multiplication of Zolesio [29] is used. Likewise, $\mathbf{f}$ maps $\mathcal{U}_2 = \{(v,u) \in \mathcal{D}(B^2) : |v|_{\tilde{H}^{r+1}(D)} < \delta\}$ into $\tilde{H}^r(D) \times \tilde{H}^r(D)$. The only thing remaining is to check that $\mathbf{f}(v,u)$ satisfies any boundary conditions required to have (i) $\mathbf{f}(v,u) \in X^\beta \times X^{\beta-(1/2)} = \mathbf{X}$ when $(v,u) \in \mathcal{D}(B) = X^{\beta+(1/2)} \times X^{\beta+(1/2)}$ and (ii) $\mathbf{f}(v,u) \in X^{\beta+(1/2)} \times X^{\beta+(1/2)}$, when $(v,u) \in \mathcal{D}(B^2)$.

The only boundary conditions which might need to be satisfied are of the form $\frac{\partial}{\partial \mathbf{n}} \Delta_N^i f_j = 0$, integer $i \geq 0, j = 1$ or 2. Recall that $\beta = \frac{r-1}{2}$. For example, if $\frac{5}{2} > r > 2$, then $X^{\beta-(1/2)} = \tilde{H}^{r-2}(D), X^\beta = \tilde{H}^{r-1}(D), X^{\beta+(1/2)} = \tilde{H}^r(D) \cap \{u : \frac{\partial u}{\partial \mathbf{n}} = 0 \text{ on } \partial D\}$. Another example is if $\frac{7}{2} > r > \frac{5}{2}$, then $X^{\beta-(1/2)} = \tilde{H}^{r-2}(D), X^\beta = \tilde{H}^{r-1}(D) \cap \{u : \frac{\partial u}{\partial \mathbf{n}} = 0 \text{ on } \partial D\}$, and $X^{\beta+(1/2)} = \tilde{H}^r(D) \cap \{u : \frac{\partial u}{\partial \mathbf{n}} = 0 \text{ on } \partial D\}$. Another example to note is that, if $\frac{9}{2} > r > \frac{7}{2}, X^{\beta+(1/2)} = \tilde{H}^r(D) \cap \{u : \frac{\partial u}{\partial \mathbf{n}} = 0, \frac{\partial \Delta u}{\partial \mathbf{n}} = 0 \text{ on } \partial D\}$.

In our problem, D is a rectangle, say $D = \{\mathbf{x} = (x_1, x_2) : 0 < x_j < \ell_j, j = 1, 2\}$ for some $\ell_1, \ell_2 > 0$. The eigenfunctions are $\psi_n(\mathbf{x}) = \cos(\omega_{1,n} x_1) \cos(\omega_{2,n} x_2)$, where $\omega_{1,n}^2 + \omega_{2,n}^2 = \kappa_n^2$. Because of this explicit information that the ψ_n's are even functions of both x_1 and x_2 with respect to 0, we see that for $r > 2, u, v, w \in \tilde{H}^r(D)$ are even functions in x_1 and x_2 and that

$$f_1 = w|\nabla_{\mathbf{x}} v|^2 - \nabla_{\mathbf{x}} u \cdot \nabla_{\mathbf{x}} v \in \tilde{H}^{r-1}(D),$$

$$M_2 = -\frac{1}{2}\, |\nabla_{\mathbf{x}} u|^2 + \frac{1}{2}\, w^2(1 + |\nabla_{\mathbf{x}} v|^2) \in \tilde{H}^{r-1}(D),$$

$$f_2 = M_2 - [M_2]$$

are even functions of x_1 and x_2. Again, the result on multiplication of Zolesio [29] is used. This implies that $\frac{\partial}{\partial \mathbf{n}} \Delta^i f_j = 0$ on ∂D whenever it makes sense in terms of a trace of a distribution. It follows that (i) $\mathbf{f}$ satisfies the boundary conditions to be in $\mathbf{X} = X^\beta \times X^{\beta-(1/2)}$ whenever $(v,u) \in \mathcal{D}(B) = X^\beta \times X^{\beta-(1/2)}$ and (ii) $\mathbf{f}$ satisfies boundary conditions to be in $\mathcal{D}(B) = X^{\beta+(1/2)} \times X^{\beta+(1/2)}$ whenever $(v,u) \in \mathcal{D}(B^2) = (X^{\beta+1} \times X^{\beta+1}) \cap \{(v,u) : \gamma v + \mu u \in H^{\beta+(3/2)}\}$.

This completes the proof of Corollary 6.4. □

THEOREM 6.5. *For the nonlinear periodic problem* (2.6)–(2.7), *if* $\beta = \frac{r-1}{2} > \frac{1}{2}$, *there is a neighborhood* $\mathcal{U}$ *of* $\mathbf{0} \in F = D_B(\theta+1) \subset \mathcal{D}(B) = X^{\beta+(1/2)} \times X^{\beta+(1/2)}$ *in which there is existence and uniqueness of solutions, locally forward in time.*

We note that if $\beta > 1$, i.e., $r > 3$, the solutions are classical, i.e., $C^2(D) \times C^2(D)$, by the Sobolev imbedding theorem. Theorem 6.5 can be proven by a standard use of the contraction mapping theorem, along with Proposition 4.1 and Corollary 6.4.

We also have the following theorem.

THEOREM 6.6 (linearized stability). *For the nonlinear periodic problem* (2.6)–(2.7), *if* $\beta > \frac{1}{2}$, *then* $(v, u) = 0 \in F$ *is exponentially asymptotically stable in the sense of Liapunov, i.e., locally in* F, *if* $\|\alpha\|_\infty := \max_{0 \le t \le T} |\alpha(t)|$ *is sufficiently small.*

Theorem 6.6 can be proven in the usual way, using Proposition 4.1, Corollary 5.3, and Corollary 6.4; see, e.g., the proof of daPrato and Lunardi [9, Thm. 2.2]. The restriction "locally in F" is due to the same restriction in Corollary 6.4.

7. A local center-unstable manifold theorem for problem (2.6)–(2.7).

As usual, we fix a $\theta \in (0,1)$.

DEFINITION. *A* C^k *local center-unstable manifold in* Z *for a* T*-periodic problem, i.e., periodic with period* T,

$$\frac{d\mathbf{w}}{dt} = (B + B_1(t))\mathbf{w} + \mathbf{f}(\mathbf{w}), \mathbf{w} \in Z \tag{7.1}$$

is a set of the form

$$\mathcal{M} = \{(t, \xi_1 + \mathbf{h}(t, \xi_1)) : \xi_1 \in \mathcal{U}_1(t), t \in \mathbb{R}\} \subseteq \mathbb{R} \times Z,$$

where for all $t, Z = Z_1(t) \oplus Z_2(t), \mathcal{U}_1(t)$ *is a neighborhood of* $0 \in Z_1(t)\, \mathbf{h}(t, \cdot) : \mathcal{U}_1 \to Z_2(t)$ *is* $C^k, \mathbf{h}(t, 0) \equiv \mathbf{0}, d_{\mathbf{z}}\mathbf{h}(t, 0) \equiv 0$, *where* d *denotes a Fréchet derivative, and* $\mathcal{M}$ *is invariant forward in time* t *for* (7.1).

We note from the proof of Corollary 5.3 that, for the linear problem (2.8), there are at most finitely many unstable modes, corresponding to characteristic multipliers of modulus greater than 1, no matter how large $\|\alpha\|_\infty := \max_{0 \le t \le T} |\alpha(t)|$. We assume

(H) α is such that (2.8) has a positive finite number ν of characteristic multipliers greater than or equal to 1.

Correspondingly, there is for each $t \in \mathbb{R}$ a finite-rank projection $P_1(t)$ on the Hilbert space $\mathcal{X}$; we denote $P_2(t) = I - P_1(t)$. In an abuse of notation we use $P_1(t)$ to denote the restriction of $P_1(t)$ to $E = D_B(\theta)$ or $F = D_B(\theta + 1)$; the latter are not Hilbert spaces, but the linear operators $P_1(t)$ are still uniformly bounded in t, by interpolation.

In fact, one can choose $P_1(t)$ so that the evolution family $\{\Phi(t,s)P_1(s)\}$ satisfies

$$\Phi(t,s)P_1(s) = P_1(t)\Phi(t,s), \quad t \ge s.$$

Moreover, because $\dim P_1(s)E = \nu < \infty$ for all s, the evolution family $\{\Phi(t,s)P_1(s)\}$ can be extended to satisfy

$$\Phi(t,s)P_1(s) = P_1(t)\Phi(t,s),$$

$$\Phi(t,s)P_1(s)\Phi(s,\tau)P_1(\tau) = \Phi(t,\tau)P_1(\tau)$$

for all t, s, τ; in an abuse of notation, we have not bothered to renotate this extended family. Recall that $\nu < \infty$. By hypothesis (H), there exists $M_1 \geq 1, \omega_1 > 0$ such that

$$\|\Phi(t,s)P_1(s)\|_{L(E,F)} \leq M_1 e^{-\omega_1(t-s)} \quad \text{for } t \leq s. \tag{7.2}$$

$$\|\Phi(t,s)P_1(s)\|_{L(E)} \leq M_1 e^{-\omega_1(t-s)} \quad \text{for } t \leq s,$$

i.e., backwards in time. In fact, since all of the characteristic multipliers of $\Phi(t+T, T)P_1(t)$ are of modulus greater than or equal to 1, one may choose positive ω_1 as close to 0 as desired, although perhaps at the expense of increasing M_1. Because of estimate (7.2) and the fact that $\dim P_1(s)E = \nu < \infty$ for all s, one has a backwards maximal regularity result, as in daPrato and Lunardi [9, Thm. 2.4].

LEMMA 7.1(corollary to Proposition 5.2). *Assume that $B_1(\cdot)$ is periodic with period $T, B_1(\cdot) : [0,T] \to L(E^\rho, E)$ is Hölder continuous, hypothesis* (H) *holds, and η is chosen so that $\omega_1 - \eta < 0$, where ω_1 is as in* (7.2). *Then for all $\sigma \in \mathbb{R}$, the map*

$$f(\cdot) \mapsto \int_\sigma^\cdot \Phi(\cdot, s)P_1(s)f(s)\,ds$$

defines a bounded linear operator

$$\tilde{\mathcal{L}}_+ : C_\eta((-\infty,\sigma]; E) \to C_\eta((-\infty,\sigma]; F).$$

By hypothesis (H) and the definitions of $P_1(t)$ and $P_2(t)$, we have that there exists $M_2 \geq 1, \tilde{\omega} < 0$ such that

$$\|\Phi(t,s)P_2(s)\|_{L(E,F)} \leq M_2 e^{\tilde{\omega}(t-s)}, \quad \text{for } t \geq s \tag{7.3}$$

by Proposition 5.1.

LEMMA 7.2 (corollary to Proposition 5.2). *Assume that $B_1(\cdot)$ is periodic with period $T, B_1(\cdot) : [0,T] \to L(E^\rho, E)$ is Hölder continuous, hypothesis* (H) *holds, and η is chosen so that $\tilde{\omega} - \eta < 0$, where $\tilde{\omega}$ is as in* (7.3). *Then for all $\sigma \in \mathbb{R}$, the map*

$$f(\cdot) \mapsto \int_{-\infty}^\cdot \Phi(\cdot, s)P_2(s)f(s)\,ds$$

defines a bounded linear operator

$$\tilde{\mathcal{L}}_- : C_\eta((-\infty,\sigma]; E) \to C_\eta((-\infty,\sigma]; F).$$

To help in what follows, here are some comparisons of notation in related papers. Chow and Lu's [6] $-\alpha - \eta, -\beta + \eta$, correspond to our $\omega, \tilde{\omega}$, and their $\eta, C_\eta(\mathbb{R}^-, X)$ correspond to our $\eta, C_\eta((-\infty,\sigma]; F)$. Chow, Lin, and Lu's [7] α and $-\beta$ correspond to our ω_1 and $\tilde{\omega}$, and their γ and $E_\sigma^-(-\gamma, X)$ correspond to our $-\eta$ and $C_\eta((-\infty,\sigma]; F)$. DaPrato and Lunardi's [9, proof of Thm. 3.1] ω_1 and $-\mu$ correspond to our $\tilde{\omega}$ and η. Chow and Lu's [6] condition $\beta + (k-1)\eta > 0$ corresponds to our condition $\tilde{\omega} - k\eta < 0$. Chow, Lin, and Lu's [7] spectral gap conditions $\alpha < \gamma \leq k\gamma < \beta$ correspond to our conditions $\omega_1 + \eta < 0, \eta < 0, \tilde{\omega} - k\eta < 0, \tilde{\omega} < 0$, and hypothesis (H). DaPrato and Lunardi's spectral gap condition $\lambda_1 < 0 \leq \lambda_2$ and $\omega_1 + \mu < 0$ correspond to our conditions $\tilde{\omega} < 0$, hypothesis (H), and $\tilde{\omega} - \eta < 0$.

THEOREM 7.3. *Assume* (H). *For any integer $k \geq 1$, there exists a local center-unstable manifold $\mathcal{M}$ of dimension ν for the nonlinear periodic problem* (2.6)–(2.7) *and hence for a model of the Faraday resonance. If we fix* $0 < \theta < 1, r > 2$, *then* $\mathcal{M} \subset \mathbb{R} \times F$, *where* $F = D_B(\theta + 1) \subset \mathcal{D}(B) = X^{r/2} \times X^{r/2} \subset \tilde{H}^r(D) \times \tilde{H}^r(D)$. *If* $r + \theta > 3$ *then the solutions are classical.*

Proof. Fix any $r > 2$ and $0 < \theta < 1$ and define $E = D_B(\theta), F = D_B(\theta + 1)$. Fix any $\sigma \in \mathbb{R}$. Use Lemma 7.2 to define a bounded linear operator $\tilde{\mathcal{L}}_-$, and use Lemma 7.1 to define a bounded linear operator $\tilde{\mathcal{L}}_+$. The η that one chooses is determined by the ω_1 in Lemma 7.1 and $\tilde{\omega}$ in Lemma 7.2. One chooses $\eta < 0$ such that $\tilde{\omega} - k\eta < 0$ and $\omega_1 - \eta < 0$. By Corollary 6.4, there is a neighborhood of $\mathbf{0} \in \mathcal{U} \subset F$ such that $\mathbf{f} : \mathcal{U} \to E$ is C^∞. The proofs of Chow and Lu [6, §§3 and 4] for autonomous problems and Chow, Lin, and Lu [7, Lem. 3.1 and 3.2 and Thm. 3.3] for nonautonomous problems work just as well. First, one defines a bounded linear operator $\mathcal{T} : C_\eta((-\infty, \sigma]; E) \to C_\eta((-\infty, \sigma]; F)$ by $\mathbf{f}(\cdot) \mapsto \int_\sigma^\cdot \Phi(\cdot, s)P_1(s)\mathbf{f}(s)\,ds + \int_{-\infty}^\cdot \Phi(\cdot, s)P_2(s)\mathbf{f}(s)\,ds$. After making the usual cutoff function alteration of $\mathbf{f}$ to get a function $\tilde{\mathbf{f}}$ with sufficiently small Lipschitz constant, the solution of the integral equation

$$\varphi(t) = \Phi(t, \sigma)\xi_1 + \int_\sigma^t \Phi(t, s)P_1(s)\tilde{\mathbf{f}}(\varphi(s))\,ds + \int_{-\infty}^t \Phi(t, s)P_2(s)\tilde{\mathbf{f}}(\varphi(s))\,ds$$

is denoted by $\varphi(t; \sigma, \xi_1)$ for $t \le \sigma, \xi_1 \in \mathcal{U}_1(\sigma)$. The function $\mathbf{h}$ which gives the integral manifold $\mathcal{M}$ is then defined by

$$\mathbf{h}(\sigma, \xi_1) = \int_{-\infty}^\sigma \Phi(\sigma, s)P_2(s)\tilde{\mathbf{f}}(\varphi(s; \sigma, \xi_1))\,ds = \varphi(\sigma; \sigma, \xi_1) - \xi_1.$$

See also daPrato and Lunardi [9] for autonomous problems. The conclusion about when the solutions are classical follows from (4.2)–(4.4) and the fact that $F = D_B(\theta + 1) = \{\mathbf{w} \in D(B) : B\mathbf{w} \in D_B(\theta)\}$. □

We note that, as usual, $\mathcal{M}$ may depend on k and the choice of the cutoff function used in the proof. If, in addition, $P_1(t+T) \equiv P_1(t)$, then $\mathbf{h}(t+T, \xi_1) \equiv \mathbf{h}(t, \xi_1)$ follows from $\Phi(t + T, s + T) \equiv \Phi(t, s)$. So, if the linearization has only decaying modes and simple periodic solutions, then we get a periodic center manifold.

THEOREM 7.4 (exponential attractivity with asymptotic phase). *For any integer* $k \ge 1$ *the local center-unstable manifold* $\mathcal{M}$ *for problems* (2.6)–(2.7) *is locally exponentially attractive, i.e., there exists a neighborhood* $\mathcal{U}$ *of* $\mathcal{M}$ *such that, if* $\mathbf{w}(t) = (v(t), u(t))$ *is a solution of* (2.6)–(2.7) *which exists for* $t \in [\sigma, \infty)$ *and such that* $(t, \mathbf{w}(t)) \in \mathcal{U}$ *for* $t \in [\sigma, \infty)$, *then there exists a solution* $\mathbf{w}^*(t)$ *of* (2.6)–(2.7), $(\sigma, \mathbf{w}^*(\sigma)) \in \mathcal{M}$, *such that* $|\mathbf{w}(t) - \mathbf{w}^*(t)|_F \to 0$ *as* $t \to \infty$, *exponentially.*

Proof. If necessary, by taking $\mathcal{U}$ smaller than in Theorem 7.3, one can make the Lipschitz constant of $\tilde{\mathbf{f}}$ as small as one needs. The proof is then similar to that of Chow and Lu [6, Thm. 5.1]; again, one uses the integral operators $\tilde{\mathcal{L}}_\pm$. □

The hypothesis that $(t, \mathbf{w}(t)) \in \mathcal{U}$ for $\sigma \le t < \infty$ ensures that the solution of (2.6)–(2.7) also satisfies the problem with a cutoff of $\mathbf{f}$.

8. An example. We mention in passing a result of a sequel in preparation. For a square tank, Silber and Knobloch [23], following experimental work of Simonelli and Gollub [24], studied excitation of the (3, 2) and (2, 3) spatial modes. We were able to calculate an approximate, local center manifold which reduces the dynamics to that of a periodic system of ordinary differential equations

$$\left\{ \begin{array}{l} \dot{x} = xq(t, x, y) \\ \dot{y} = yq(t, y, x) \end{array} \right\},$$

where $q(t, x, y) = a(t)x^2 + b(t)y^2$. Here, x, y correspond to the center directions for the (3, 2) and (2, 3) spatial modes.

Acknowledgments. The hospitality of the School of Mathematics of the Georgia Institute of Technology is gratefully acknowledged. It is my pleasure to thank Shui-Nee Chow, Anibal Rodriguez-Bernal, Wenxian Shen, and Mary Silber for their help; in particular, the impetus for this work came out of discussions with Chow and Silber. I also thank an anonymous referee for some helpful suggestions. The referee also made me aware of a paper by Lunardi [17]; it appears that, for my particular applied problem, my work uses simpler spaces than in her paper on fully nonlinear parabolic equations.

REFERENCES

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

[2] S. B. ANGENENT, *Nonlinear analytic semiflows*, Proc. Roy. Soc. Edinburgh Sect. A, 115 (1990), pp. 91–107.

[3] T. B. BENJAMIN AND F. URSELL, *The stability of the plane free surface of a liquid in vertical periodic motion*, Proc. Roy. Soc. London Ser. A, 225 (1954), pp. 505–515.

[4] P. L. BUTZER AND H. BERENS, *Semigroups of Operators and Approximations*, Springer-Verlag, New York, 1967.

[5] S. CHEN AND R. TRIGGIANI, *Characterization of domains of fractional powers of certain operators arising in elastic systems, and applications*, J. Differential Equations, 88 (1990), pp. 279–293.

[6] S.-N. CHOW AND K. LU, *Invariant manifolds for flows in Banach spaces*, J. Differential Equations, 74 (1988), pp. 285–317.

[7] S.-N. CHOW, X.-B. LIN, AND K. LU, *Smooth invariant foliations in infinite-dimensional spaces*, J. Differential Equations, 94 (1991), pp. 266–291.

[8] G. DAPRATO AND P. GRISVARD, *Equations d'evolution abstraites nonlineaires de type parabolique*, Ann. Mat. Pura Appl. (4), 120 (1979), pp. 329–396.

[9] G. DAPRATO AND A. LUNARDI, *Stability, instability, and center manifold theorem for fully nonlinear autonomous parabolic equations in Banach space*, Arch. Rational Mech. Anal., 101 (1988), pp. 115–141.

[10] S. DOUADY, *Experimental study of the Faraday instability*, J. Fluid Mech., 221 (1990), pp. 383–409.

[11] P. GRISVARD, *Caracterisation de quelques espaces d'interpolation*, Arch. Rational Mech. Anal., 25 (1967), pp. 40–63.

[12] D. HENRY, *Geometric theory of semilinear parabolic equations*, in Lecture Notes in Mathematics, vol. 840, Springer-Verlag, Berlin, 1981.

[13] P. HOLMES, *Chaotic motions in a weakly nonlinear model for surface waves*, J. Fluid Mech., 162 (1986), pp. 365–388.

[14] X. M. GU AND P. R. SETHNA, *Resonant surface waves and chaotic phenomena*, J. Fluid Mech., 183 (1987), pp. 543–565.

[15] X.-B. LIN, *Existence theory for damped gravity waves in a closed rectangular basin*, Arch. Rational Mech. Anal., 114 (1991), pp. 267–295.

[16] J. L. LIONS AND E. MAGENES, *Nonhomogeneous Boundary Value Problems and Applications*, vol. I, Springer-Verlag, New York, 1972.

[17] A. LUNARDI, *Stability of the periodic solutions to fully nonlinear parabolic equations in Banach spaces*, Differential Integral Equations, 1 (1988), pp. 253–279.

[18] J. MILES AND D. HENDERSON, *Parametrically forced surface waves*, Ann. Rev. Fluid Mech., 22 (1990), pp. 143–165.

[19] S. M. NIKOL'SKII, *On imbedding, continuation, and approximation theorems for differentiable functions of several variables*, Uspekhi Mat. Nauk, 16 (1961), pp. 63–114 (in Russian); English translation, Russian Math. Surveys, 16 (1961), pp. 55–104.

[20] ———, *Approximation of Functions of Several Variables and Imbedding Theorems*, Springer-Verlag, Berlin, 1975.

[21] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[22] A. RODRIGUEZ-BERNAL, *On the generation of analytic semigroups by a class of damped wave equations*, to appear.

[23] M. SILBER AND E. KNOBLOCH, *Parametrically excited surface waves in square geometry*, Phys. Letters A, 137 (1989), pp. 349–354.

[24] F. SIMONELLI AND J. P. GOLLUB, *Surface wave mode interactions: Effects of symmetry and degeneracy*, J. Fluid Mech., 199 (1989), pp. 471–494.

[25] E. SINESTRARI, *On the abstract Cauchy problem of parabolic type in spaces of continuous functions*, J. Math. Anal. Appl., 107 (1985), pp. 16–66.

[26] L. N. SLOBODECKII, *Generalized Sobolev spaces and their application to boundary value problems for partial differential equations*, Leningradskiĭ Gosudarstvennyĭ Pedagogičeskiĭ Institut im A. I. Gercena Učenye Zapiski, 197 (1958), pp. 54–112 (in Russian); English translation, Amer. Math. Soc. Transl. Ser. 2, 57 (1966), pp. 207–275.

[27] L. TURYN, *The damped Mathieu equation*, Quart. Appl. Math., 51 (1993), pp. 389–398.

[28] S. A. VAN GILS AND A. VANDERBAUWHEDE, *Center manifolds and contractions on a scale of Banach spaces*, J. Funct. Anal., 72 (1987), pp. 209–224.

[29] J. L. ZOLESIO, *Multiplication dans les espaces de Besov*, Proc. Roy. Soc. Edinburgh Sect. A, 78 (1977), pp. 113–117.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 258–273, January 1996

014

THE REGULARIZATION OF LINEAR DIFFERENTIAL-ALGEBRAIC EQUATIONS*

LEONID V. KALACHEV† AND ROBERT E. O'MALLEY, JR.‡

Abstract. Differential-algebraic equations naturally arise in numerous important applied contexts, e.g., electrical circuits, constrained dynamical motion and fluid dynamics. The authors show how the structure of the solution space for some typical problems can be illuminated through the introduction of various regularizations which often can be given natural physical interpretations.

Key words. differential-algebraic equations, constrained differential equations, regularization

AMS subject classifications. 34A09, 34A12, 34D15

1. An index-one problem.

Consider the constrained differential equation

$$\left\{\begin{aligned} \dot{U}_0 &= A(t)U_0 + B(t)V_0 + f(t), \\ 0 &= C(t)U_0 + D(t)V_0 + g(t) \end{aligned}\right. \tag{1.1}$$

consisting of a linear differential system for the m-vector U_0 and a linear constraint relating U_0 and the n-vector V_0. If we make the strong assumption that

$$\text{the matrix } D(t) \text{ remains nonsingular,} \tag{1.2}$$

we can solve for

$$V_0 = -D^{-1}(CU_0 + g), \tag{1.3}$$

and there simply remains the linear differential system

$$\dot{U}_0 = (A - BD^{-1}C)U_0 + (f - BD^{-1}g) \tag{1.4}$$

for U_0. The standard existence theory for linear ordinary differential equations (cf., e.g., Coddington and Levinson [5]) implies that if the coefficients A, B, f, C, D, and g are continuous functions on some bounded interval of, say, $t \geq 0$, then the solution space for this so-called index-one differential-algebraic equation (or DAE) (1.1) will be m-dimensional, being parameterized, for example, by the m-components of the initial vector $U_0(0)$. Problem (1.1)–(1.2) is usually called index-one because one differentiation of the algebraic constraint in (1.1) provides $\dot{U}_0$ and $\dot{V}_0$ as explicit functions of U_0, V_0, and t.

Such DAEs and their nonlinear generalizations arise in many practical, important contexts, ranging from the Navier–Stokes equations to Euler–Lagrange equations and to Kirchhoff's laws. Constant coefficient problems were classically solved through the use of matrix pencils (cf. Gantmacher [8]), but more general DAEs were seldom

*Received by the editors March 25, 1994; accepted for publication (in revised form) April 25, 1994. This research was supported in part by National Science Foundation grant DMS-9107196.

†Department of Applied Mathematics, University of Washington, FS-20, Seattle, WA 98195. Current address: Department of Mathematical Sciences, University of Montana, Missoula, MT 59812.

‡Department of Applied Mathematics, University of Washington, FS-20, Seattle, WA 98195.

studied very thoroughly. DAEs have, however, been of substantial recent interest to the numerical analysis community (cf., e.g., Brenan et al. [2], Hairer et al. [14], Hairer and Wanner [15], and Griepentrog et al. [13]). Our study indicates how regularization can effectively reveal the solution structure for DAEs. Following the general approach of Tikhonov and Arsenin [33], we will define a continuum of problems P_ϵ for small positive ϵ values and determine the DAE's solutions as limiting solutions of P_ϵ for $\epsilon \to 0$ (for, say, $t > 0$). This relates to earlier analytical and numerical studies involving such regularizations by Campbell [3], Lötstedt [22], Knorrenschild [20], Hanke [16], and Eich and Hanke [7], but our approach is more closely based on singular perturbation theory (cf., e.g., O'Malley [26]). Some further nonlinear generalizations are contained in Kalachev and O'Malley [19].

One way to regularize problem (1.1) is to consider it as the limit of the singularly perturbed $(m+n)$-dimensional linear system

$$\begin{cases} \dot{u} = Au + Bv + f, \\ -\epsilon D\dot{v} = Cu + Dv + g \end{cases} \tag{1.5}$$

as $\epsilon \to 0$. Rewriting this as

$$\begin{cases} \dot{u} = Au + Bv + f, \\ \epsilon\dot{v} = -D^{-1}Cu - v - D^{-1}g \end{cases} \tag{1.6}$$

emphasizes that the small *positive* parameter ϵ was introduced in (1.5) in a manner to achieve the appropriate asymptotic stability for (1.6), which implies fast initial dynamics where v generally decays rapidly like a constant vector multiple of the scalar function $e^{-t/\epsilon}$. The well-known Tikhonov–Levinson theory for singularly perturbed initial value problems (cf. Smith [31], O'Malley [26], and Vasil'eva et al. [35]) directly establishes that (1.5) has an $(m+n)$-dimensional solution space, parametrized by any prescribed bounded initial vector

$$\begin{pmatrix} u(0) \\ v(0) \end{pmatrix},$$

with solutions having the asymptotic form

$$\begin{cases} u(t,\epsilon) = U_0(t) + O(\epsilon), \\ v(t,\epsilon) = V_0(t) + \beta_0(t/\epsilon) + O(\epsilon) \end{cases} \tag{1.7}$$

on fixed bounded intervals $0 \le t \le T$, where

$$\begin{pmatrix} U_0 \\ V_0 \end{pmatrix}$$

satisfies the original DAE (1.1) (and therefore (1.3) and (1.4)) and where

$$\beta_0(\tau) = e^{-\tau}[v(0) + D^{-1}(0)C(0)u(0) + D^{-1}(0)g(0)] \tag{1.8}$$

is the decaying solution of the vector initial layer problem $\frac{d\beta_0}{d\tau} = -\beta_0$, $\beta_0(0) = v(0) - V_0(0)$ on the stretched interval $\tau = \frac{t}{\epsilon} \ge 0$. We note that this limiting initial layer correction $\beta_0(\tau)$ will be trivial when the initial value for $v(0)$ is consistent with the constraint $C(0)u(0) + D(0)v(0) + g(0) = 0$, so $\beta_0(0) = 0$. Otherwise, the solution of

(1.6) will feature nonuniform convergence in the v variable at $t = 0$, i.e., an initial Heaviside function discontinuity in the $\epsilon \to 0$ limit. One way to solve the original DAE is to numerically integrate the regularized problem (1.5) using a stiff integrator for a sequence of small $\epsilon > 0$ values. (Using the perturbation term $-\epsilon D(0)\dot{v}$ may be computationally preferable to that used in (1.5), since one won't then need to calculate $D^{-1}(t)$ for all $t > 0$ of interest. Then, however, the t interval must be restricted so that $-D^{-1}(0)D(t)$ remains stable.) More simply, one can just integrate the nonstiff mth-order initial value problem for U_0 (cf. (1.4)) with $U_0(0) = u(0)$ and use (1.3) to obtain V_0.

We note that the artificially introduced parameter ϵ may often by provided a physical interpretation (see below). Indeed, when a DAE (1.1) results from the neglect of rapid transients in a dynamic model, $\frac{1}{\epsilon}$ may represent the size of the smallest neglected decay rate.

2. A pure index-two problem in Hessenberg form. More challenging DAEs than (1.1)–(1.2) arise when the matrix $D(t)$ in (1.1) is identically zero, when it has *fixed* positive rank r less than n, or when it involves isolated turning points where $D(t)$ becomes singular or changes rank. We will first consider the "pure" index-two situation when $D(t) \equiv 0$ and then the case when $0 < r < n$. Specifically, let us first consider the DAE

$$\begin{cases} \dot{U}_0 = A(t)U_0 + B(t)V_0 + f(t), \\ 0 = C(t)U_0 + g(t) \end{cases} \tag{2.1}$$

in the situation that the number m of differential equations is not exceeded by the number n of constraints and that the $n \times n$ matrix

$$C(t)B(t) \text{ remains nonsingular} \tag{2.2}$$

for all $t \geq 0$. We will also assume somewhat more smoothness than before. The classical approach to solving (2.1)–(2.2) involves differentiating the constraint with respect to t to yield $C(AU_0 + BV_0 + f) + \dot{C}U_0 + \dot{g} = 0$ and solving this equation for V_0. Since the remaining problem for U_0 is like the index-one problem (1.1)–(1.2), it is natural to call (2.1)–(2.2) an index-two problem. See Gear and Petzold [10] and Griepentrog et al. [13] for comparisons of various index concepts. Very crudely, the index is the number of differentiations necessary to convert a DAE to an ordinary differential equation in standard form.

Here, let us consider the regularized problem

$$\begin{cases} \dot{u} = Au + Bv + f, \\ -\epsilon CBv = Cu + g, \end{cases} \tag{2.3}$$

where the scaled vector $-\epsilon CBv$ is used as a slack variable, so the given constraint need only be satisfied in the limit as $\epsilon v \to 0$. Eliminating

$$v = -\frac{1}{\epsilon}(CB)^{-1}[Cu + g] \tag{2.4}$$

leaves us with the singularly perturbed differential system

$$\epsilon\dot{u} = [-B(CB)^{-1}C + \epsilon A]u - B(CB)^{-1}g + \epsilon f \tag{2.5}$$

for u. We note that an alternative procedure (cf. Kalachev and O'Malley [18]) would be to differentiate the constraint in (2.3) to obtain a singularly perturbed differential system for u and v. Both approaches suggest more impulsive initial behavior for v than for u.

The Tikhonov–Levinson theory does not apply for the singular singular-perturbation problem (2.5) because the limiting Jacobian matrix

$$-Q(t) \equiv -B(CB)^{-1}C \tag{2.6}$$

generally has a constant rank $n < m$. Note that Q is a projection onto the range of C since $Q^2 = Q$, $CQ = C$, and rank $Q \leq$ rank C. We can, nonetheless, seek an asymptotic solution to the initial value problem for (2.5) in the form

$$u(t, \epsilon) = U(t, \epsilon) + \alpha(\tau, \epsilon), \tag{2.7}$$

where the outer expansion

$$U(t, \epsilon) = U_0(t) + \epsilon U_1(t) + \epsilon^2 U_2(t) + \cdots \tag{2.8}$$

must naturally satisfy (2.5) for all $t > 0$ as a formal power series in ϵ, while the initial layer correction $\alpha(\tau, \epsilon)$ must be a decaying solution of the homogeneous initial layer system

$$\frac{d\alpha}{d\tau} = [-Q(\epsilon\tau) + \epsilon A(\epsilon\tau)]\, \alpha \tag{2.9}$$

on $\tau \equiv \frac{t}{\epsilon} \geq 0$, thereby providing any necessary nonuniform convergence near $t = 0$ if $\alpha(0, \epsilon) = u(0) - U(0, \epsilon)$ is nontrivial (cf. Vasil'eva and Butuzov [34], O'Malley [26], and Vasil'eva et al. [35] for treatments of such singular problems).

Equating coefficients termwise in (2.5) implies that the terms of the outer expansion must successively satisfy

$$\begin{cases} 0 = -QU_0 - B(CB)^{-1}g, \\ \dot{U}_0 - AU_0 = -QU_1 + f, \\ \dot{U}_1 - AU_1 = -QU_2, \\ \text{etc.} \end{cases} \tag{2.10}$$

Introducing the complementary rank-$(m - n)$ projection

$$P(t) \equiv I - B(CB)^{-1}C \tag{2.11}$$

to Q, note that $CP = 0$ and $PQ = 0$. We can naturally seek any m-vector z by first finding Qz, then Pz, and finally z as their sum. We observe the widespread use of more general projection matrices in Griepentrog et al. [13] and in other earlier work cited. Indeed, the direct use of a matrix pseudoinverse of C is naturally suggested by the formulation (2.1).

Note that (2.10)(a) implies that

$$QU_0 = -B(CB)^{-1}g, \tag{2.12}$$

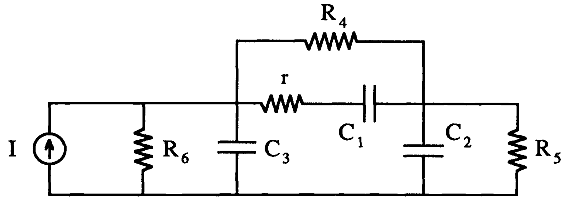


FIG. 1

while multiplication of (2.10)(b) by P yields $P\dot{U}_0 = PAU_0 + Pf$. Substituting $P\dot{U}_0 = (PU_0)^{\bullet} - \dot{P}U_0$ and replacing U_0 by $PU_0 + QU_0$ finally provides the linear differential equation

$$(PU_0)^{\bullet} = (PA + \dot{P})(PU_0) + Pf - (PA + \dot{P})B(CB)^{-1}g \tag{2.13}$$

for the projection PU_0. This implies an $(m - n)$-dimensional solution space for the DAE (2.1)–(2.2), since it remains only to integrate the differential system (2.13) for PU_0 using an *arbitrary* initial vector $P(0)U_0(0)$. Then, we will have

$$U_0 = PU_0 - B(CB)^{-1}g. \tag{2.14}$$

Since $CU_0 = -g$, (2.4) implies that we must still determine the original vector $V_0 = -(CB)^{-1}CU_1$. Note, however, that (2.10)(b) now specifies $QU_1 = f + AU_0 - \dot{U}_0$. But $CP = 0$ implies that $CU_1 = CQU_1$, so

$$\begin{aligned} V_0 &= -(CB)^{-1}C(f + AU_0 - \dot{U}_0) \\ &= -(CB)^{-1}C[(A - PA - \dot{P})(PU_0 - B(CB)^{-1}g) + (B(CB)^{-1}g)^{\bullet} + f - Pf]. \end{aligned} \tag{2.15}$$

To get the limiting solution then requires us to use both derivatives $\dot{P}$ and $(Qg)^{\bullet}$. Integrating the nonstiff m-dimensional system (2.13) for PU_0 would seem far easier numerically than trying to directly integrate the singularly perturbed initial value problem (2.5) to obtain U_0 as its limiting solution for $t > 0$ (since the limiting state matrix $-Q(t)$ has $m - n$ zero eigenvalues as well as n stable ones).

The limiting initial layer correction is necessarily given by

$$\alpha_0(\tau) = e^{-\tau}Q(0)[u(0) + B(0)(C(0)B(0))^{-1}g(0)], \tag{2.16}$$

since $\alpha_0(\tau)$ must be a decaying solution of $\frac{d\alpha_0}{d\tau} = -Q(0)\alpha_0$ with $\alpha_0(0) = u(0) - U_0(0)$ already known. Thus, $\alpha_0(\tau) \equiv Q(0)\alpha_0(\tau)$ and $P(0)\alpha_0(\tau) \equiv 0$, i.e., α_0 remains completely in the n-dimensional stable eigenspace of $Q(0)$. This again shows that we can take

$$P(0)U_0(0) = P(0)u(0) \tag{2.17}$$

to be arbitrary. Moreover, $\alpha_0(\tau)$ will provide the initial Heaviside jump for u, unless $\alpha_0(\tau) \equiv 0$ because the initial value for $u(0)$ is consistent with the constraint, i.e., $Q(0)u(0) = -B(0)(C(0)B(0))^{-1}g(0)$. It will, in turn, imply an initial delta-function impulse $-\frac{1}{\epsilon}(C(0)\ B(0))^{-1}C(0)\alpha_0(\frac{t}{\epsilon})$ for v. Allowing such inconsistency is particularly important for systems with switching (cf. Opal and Vlach [28]) and for stabilizing numerical drift off constraints. It is important to emphasize that the rapid motion in u is orthogonal to the constraint, i.e., in the range of $Q(0)$.

Example 1. Consider the simple linear electrical circuit pictured in Fig. 1 above.

For simplicity, let the capacitances C_j, the resistances R_j, and the input current I all have unit values and, for now, let the resistance r be zero. The voltage-current relations and the Kirchhoff voltage and current laws provide us the DAE:

$$(2.18)\qquad \begin{cases} \dot{v}_j = i_j, j = 1, 2, 3, \\ v_k = i_k, k = 4, 5, 6, \\ v_j = v_{j+3}, j = 1, 2, 3, \\ v_1 + v_2 = v_3, \\ i_1 + i_3 + i_4 + i_6 = 1, \\ \text{and} \\ i_1 - i_2 + i_4 - i_5 = 0. \end{cases}$$

Thus, we have three differential and nine algebraic equations for the twelve unknowns i_ℓ and v_ℓ, $\ell = 1, 2, \ldots, 6$. We naturally eliminate v_k and i_k, $k = 4, 5$, and 6, as well as i_2 and i_3. This leaves us the DAE

$$(2.19)\qquad \begin{cases} \dot{v} = Av + Bi_1 + f, \\ 0 = B^T v, \end{cases}$$

where

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 1 & -1 & 0 \\ -1 & 0 & -1 \end{pmatrix}, \quad v = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \quad \text{and} \quad f = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Since $B^T B = 3 > 0$, this problem is of the form (2.1)–(2.2) with $C = B^T$, and we therefore proceed to consider the slack variable regularization

$$(2.20)\qquad \begin{cases} \dot{v} = Av + Bi_1 + f, \\ -3\epsilon i_1 = B^T v \end{cases}$$

for an artificial small parameter $\epsilon > 0$. Since we can rewrite the present constraint as $v_1 + v_2 + 3\epsilon i_1 = v_3$, we can physically interpret our regularization as corresponding to the introduction of a small resistance $r = 3\epsilon$ in the circuit in series with the first capacitor or, for example, a resistance ϵ in series with each capacitor. This regularization coincides with the common technique (from Chua [4] and elsewhere) of introducing small parasitics and determining the network's solution through a limiting process.

Our regularization procedure reduces (2.18) to solving the singular singularly perturbed system

$$(2.21)\qquad \epsilon\dot{v} = -\left(\frac{1}{3}BB^T - \epsilon A\right)v + \epsilon f$$

with the limiting Jacobian matrix $-\frac{1}{3}BB^T$ having rank one and one stable and two zero eigenvalues. We naturally seek an asymptotic solution of (2.21) in the form

$$(2.22)\qquad v(t, \epsilon) = V(t, \epsilon) + \alpha(\tau, \epsilon),$$

where the outer expansion satisfies $V \sim \sum_{j\geq 0}^{\infty} V_j \epsilon^j$ and the initial layer correction $\alpha \to 0$ as $\tau \equiv \frac{t}{\epsilon} \to \infty$. Since the outer expansion will solve (2.21) termwise, we will successively need $\frac{1}{3}BB^T V_0 = 0$, $\frac{1}{3}BB^T V_1 = AV_0 + f - \overset{\bullet}{V}_0$, etc. The first condition requires that

$$V_{10} + V_{20} = V_{30}, \tag{2.23}$$

so it simply restricts the limiting outer solution $V_0 = (V_{10} \;\; V_{20} \;\; V_{30})^T$ to the constraint, while the second condition requires the componentwise equations

$$\begin{cases} V_{11} + V_{21} - V_{31} = -\overset{\bullet}{V}_{10}, \\ V_{11} + V_{21} - V_{31} = V_{10} - V_{20} - \overset{\bullet}{V}_{20}, \\ -V_{11} - V_{21} + V_{31} = -V_{10} - V_{30} + 1 - \overset{\bullet}{V}_{30}. \end{cases} \tag{2.24}$$

Eliminating V_{30} (through (2.23)) and then $V_{11} + V_{21} - V_{31}$ provides us the differential equations $\overset{\bullet}{V}_{j0} = -V_{j0} + \frac{1}{3}$, $j = 1$ and 2, for V_{10} and V_{20}. Imposing the initial conditions for $v_1(0)$ and $v_2(0)$, we then have

$$\begin{cases} V_{10}(t) = \dfrac{1}{3} + \left(v_1(0) - \dfrac{1}{3}\right) e^{-t}, \\ V_{20}(t) = \dfrac{1}{3} + \left(v_2(0) - \dfrac{1}{3}\right) e^{-t}, \\ \text{and} \\ V_{30}(t) = \dfrac{2}{3} + \left(v_1(0) + v_2(0) - \dfrac{2}{3}\right) e^{-t}. \end{cases} \tag{2.25}$$

We cannot generally expect the initial value $V_{30}(0)$ to agree with the prescribed $v_3(0)$, so there is a need for an initial layer correction term $\alpha_0(\tau)$. Later terms $V_k(t)$ in the outer expansion can be readily determined successively, up to specifying the initial values $V_{1k}(0)$ and $V_{2k}(0)$, but they will not be physically relevant.

If the initial conditions are consistent, i.e., $v_3(0) = v_1(0) + v_2(0)$, there is no need for $\alpha_0(\tau) = \alpha(\tau, 0)$. Otherwise, we will have a nontrivial limiting initial layer correction $\alpha_0(\tau) = (0 \;\; 0 \;\; \alpha_{30}(\tau))^T$ determined through the resulting scalar problem $\frac{d\alpha_{30}}{d\tau} = -\frac{1}{3}\alpha_{30}, \alpha_{30}(0) = v_3(0) - V_{30}(0)$. Thus,

$$\alpha_{30}(\tau) = e^{-\tau/3}(v_3(0) - v_1(0) - v_2(0)) \tag{2.26}$$

provides an initial Heaviside jump in v_3 in the limit $\epsilon \to 0$.

The slack variable constraint $i_1 = -\frac{B^T}{3\epsilon}(V + \alpha) \equiv I(t, \epsilon) + \frac{1}{\epsilon}\beta(\tau, \epsilon)$ implies the limiting current

$$I_0 \equiv -\frac{1}{3}B^T V_1 = -\frac{1}{3}(V_{11} + V_{21} - V_{31}) = \frac{1}{3}\left(V_{10} - \frac{1}{3}\right) = \frac{1}{3}\left(v_1(0) - \frac{1}{3}\right) e^{-t} \tag{2.27}$$

for $t > 0$ and the initial layer correction

$$\frac{1}{\epsilon}\beta_0\left(\frac{t}{\epsilon}\right) \equiv \frac{1}{3\epsilon}\alpha_{30}\left(\frac{t}{\epsilon}\right), \tag{2.28}$$

which features an initial delta function impulse whenever the initial voltages are inconsistent with the constraint. This reflects the rapid initial charging of the first capacitor when the series resistance r is small.

Example 2. If we linearize the Navier–Stokes equations about an incompressible flow with mean velocity $(U \;\; 0 \;\; 0)^T$ and introduce the moving coordinate $\xi = x - Ut$, we obtain a system of the form

$$(2.29)\qquad \begin{cases} \dfrac{\partial u}{\partial t} = \mu\left(\dfrac{\partial^2 u}{\partial \xi^2} + \dfrac{\partial^2 u}{\partial y^2} + \dfrac{\partial^2 u}{\partial z^2}\right) - \dfrac{\partial p}{\partial \xi} + F_1(\xi, y, z, t), \\ \dfrac{\partial v}{\partial t} = \mu\left(\dfrac{\partial^2 v}{\partial \xi^2} + \dfrac{\partial^2 v}{\partial y^2} + \dfrac{\partial^2 v}{\partial z^2}\right) - \dfrac{\partial p}{\partial \xi} + F_2(\xi, y, z, t), \\ \dfrac{\partial w}{\partial t} = \mu\left(\dfrac{\partial^2 w}{\partial \xi^2} + \dfrac{\partial^2 w}{\partial y^2} + \dfrac{\partial^2 w}{\partial z^2}\right) - \dfrac{\partial p}{\partial \xi} + F_3(\xi, y, z, t), \\ \dfrac{\partial u}{\partial \xi} + \dfrac{\partial v}{\partial y} + \dfrac{\partial w}{\partial z} = 0 \end{cases}$$

(cf. Criminale and Drazin [6]). Note that the latter equation enforces irrotationality. If we next Fourier transform our variables in ξ, y, and z by, for example, setting

$$\widehat{u}(\alpha, \beta, \gamma, t) \equiv \iiint_{-\infty}^{\infty} u(\xi, y, z, t) e^{i(\alpha\xi + \beta y + \gamma z)} d\xi \, dy \, dz,$$

we obtain a DAE of the form

$$(2.30)\qquad \begin{cases} \dfrac{dm}{dt} = \mu(\alpha^2 + \beta^2 + \gamma^2)m - iC^T\widehat{p} + F(\alpha, \beta, \gamma, t), \\ 0 = iCm, \end{cases}$$

where $m = (\widehat{u} \;\; \widehat{v} \;\; \widehat{w})^T$, $C = (\alpha \;\; \beta \;\; \gamma)$, and $F = (\widehat{F_1} \;\; \widehat{F_2} \;\; \widehat{F_3})^T$. We regularize this problem by introducing the scalar slack variable

$$(2.31)\qquad -\epsilon\widehat{p} = iCm,$$

which finally implies the singular singularly perturbed system

$$(2.32)\qquad \epsilon\frac{dm}{dt} = (-C^TC + \epsilon\mu(\alpha^2 + \beta^2 + \gamma^2)I)m + \epsilon F(\alpha, \beta, \gamma, t).$$

Again, we would expect to find a solution in the composite form

$$(2.33)\qquad m(t, \epsilon) = M(t, \epsilon) + \eta(\tau, \epsilon),$$

where $\eta \to 0$ as $\tau = \frac{t}{\epsilon} \to \infty$. The primary physical interest is to obtain the limiting outer solution M_0. The solution of (2.29) could ultimately be obtained by taking the inverse Fourier transform.

A physical interpretation of the regularization parameter can be obtained by realizing that introducing ϵ in (2.31) allowed us to approximate the incompressible solution by a slightly compressible solution. This, then, corresponds to the penalty function method, which has been extremely valuable numerically and analytically (cf.

Hughes et al. [17] and Temam [32]), where $\frac{1}{\epsilon}$ is a Lamé parameter. We note that such a DAE would also be obtained from the Navier–Stokes equations through appropriate discretizations (cf. Gresho et al. [12]).

3. A more complicated index-two problem. A higher index DAE can also occur such that the matrix $D(t)$ in (1.1) is transformable to a block diagonal form of fixed positive rank $r < n$. Specifically, let us suppose that a smooth nonsingular matrix $R(t)$ exists so that D can be decomposed as

$$D(t) = R(t)\begin{pmatrix} \widetilde{E}(t) & 0 \\ 0 & 0 \end{pmatrix} R^{-1}(t), \tag{3.1}$$

where $\widetilde{E}(t)$ remains a nonsingular $r \times r$ matrix. (Allowing nontrivial Jordan blocks in the nullspace of $D(t)$ might naturally be considered later.) If we use the corresponding splitting

$$V_0(t) = R(t)\begin{pmatrix} X_0(t) \\ Y_0(t) \end{pmatrix}, \tag{3.2}$$

the index-two DAE (1.1) will be transformed to the form

$$\begin{cases} \dot{U}_0 = A(t)U_0 + B(t)X_0 + C(t)Y_0 + f(t), \\ 0 = D(t)U_0 + E(t)X_0 + g(t), \\ 0 = F(t)U_0 + h(t) \end{cases} \tag{3.3}$$

(with different coefficient matrices than before). Note that transformations such as (3.2) are common in the singular perturbations literature (cf. Wasow [36], O'Malley [25], and Kreiss et al. [21]). Their numerical implementation is, however, nontrivial, often being carried out through orthogonal transformations.

We will assume that any such preliminary transformations have already been made and shall assume that (3.3) holds with

$$\begin{cases} m \geq n - r > 0, \\ \text{where the } r \times r \text{ matrix } E(t) \text{ and the } (n-r) \times (n-r) \text{ matrix} \\ F(t)C(t) \text{ both remain nonsingular.} \end{cases} \tag{3.4}$$

Our earlier analysis indicates that we should anticipate having an $(m - n + r)$-dimensional solution space for (3.3)–(3.4) which could be obtained by using the regularization

$$\begin{cases} \dot{u} = Au + Bx + Cy + f, \\ -\epsilon E\dot{x} = Du + Ex + g, \\ -\epsilon FCy = Fu + h \end{cases} \tag{3.5}$$

or, equivalently, the singular singularly perturbed system

$$\begin{cases} \epsilon\dot{u} = [-C(FC)^{-1}F + \epsilon A]u + \epsilon Bx - C(FC)^{-1}h + \epsilon f, \\ \epsilon\dot{x} = -E^{-1}Du - x - E^{-1}g, \end{cases} \tag{3.6}$$

which can be solved using the projection matrix $C(FC)^{-1}F$ and its complement to determine the limiting u, x, and y vectors.

The classical way to solve (3.3)–(3.4) is to solve (3.3)(b) for

$$X_0 = -E^{-1}DU_0 - E^{-1}g \tag{3.7}$$

and to differentiate (3.3)(c) to get

$$0 = F(AU_0 + BX_0 + CY_0 + f) + \dot{F}U_0 + \dot{h} = 0.$$

Solving for

$$Y_0 = -(FC)^{-1}[(FA + \dot{F})U_0 + FBX_0 + Ff + \dot{h}], \tag{3.8}$$

there simply remains the DAE

$$\begin{cases} \dot{U}_0 = \mathcal{A}(t)U_0 + b(t), \\ 0 = F(t)U_0 + h(t) \end{cases} \tag{3.9}$$

for U_0, where

$$\mathcal{A}(t) \equiv A - BE^{-1}D - C(FC)^{-1}[FA + \dot{F} - FBE^{-1}D]$$

and

$$b(t) \equiv f - BE^{-1}g - C(FC)^{-1}[-FBE^{-1}g + Ff + \dot{h}].$$

Because $F\mathcal{A} + \dot{F} = 0$ and $Fb + \dot{h} = 0$, it follows that the constraint (3.9)(b) defines an invariant manifold for the differential equation (3.9)(a).

We will use a new method to attack (3.9) which is also successful for nonlinear generalizations (cf. Kalachev and O'Malley [19]). Specifically, we will introduce a scaled Lagrange multiplier $\lambda(t)$ through the singularly perturbed DAE

$$\begin{cases} \dot{u}(t) = \mathcal{A}(t)u + b(t) + F^T(t)\lambda, \\ -\epsilon\lambda = F(t)u + h(t). \end{cases} \tag{3.10}$$

Note Gear's effective introduction of such a multiplier when $\epsilon = 0$ [9] and Lubich's progress for analogous nonlinear problems [23]. Equivalently, we will examine the singular singular-perturbation problem

$$\epsilon\dot{u} = [-F^T(t)F(t) + \epsilon\mathcal{A}(t)]u + [-F^T(t)h(t) + \epsilon b(t)], \tag{3.11}$$

for which the limiting Jacobian $-F^TF$ has rank $n - r$. If we seek an outer power series expansion

$$U(t, \epsilon) \sim \sum_{j=0}^{\infty} U_j(t)\epsilon^j \tag{3.12}$$

for (3.11), we successively need

$$\begin{cases} F^TFU_0 + F^Th = 0, \\ F^TFU_1 = \mathcal{A}U_0 - \dot{U}_0 + b, \\ \text{etc.} \end{cases} \tag{3.13}$$

It is now convenient to directly introduce the rank-$(n-r)$ projection matrix

$$Q(t) \equiv F^T(FF^T)^{-1}F \tag{3.14}$$

(noting that $F^TFQ = F^TF$, $QF^TF = F^TF$, and rank $Q \leq$ rank F) and to denote its complement by

$$P(t) \equiv I - F^T(FF^T)^{-1}F. \tag{3.15}$$

Note that (3.13)(a) implies that

$$QU_0 = -F^T(FF^T)^{-1}h, \tag{3.16}$$

so

$$U_0 = PU_0 - F^T(FF^T)^{-1}h. \tag{3.17}$$

Multiplying (3.13)(b) by P implies that $P(\mathcal{A}U_0 - \dot{U}_0 + b) = 0$, so U_0 will be determined through the resulting differential equation

$$(\dot{PU_0}) = (P\mathcal{A} + \dot{P})(PU_0) + Pb - (P\mathcal{A} + \dot{P})F^T(FF^T)^{-1}h \tag{3.18}$$

for PU_0 using arbitrary initial conditions $P(0)U_0(0) = P(0)u(0)$. Since P has rank $m-n+r$, the solution space for (3.3)–(3.4) is clearly of dimension $m-n+r$. An initial impulse in the range of $Q(0)$ will be necessary if $u(0)$ is inconsistent with the constraint $Fu + h = 0$. A straightforward numerical integration of the nonstiff initial value problem for (3.18) provides a practical method to obtain solutions (3.17) of the DAE (3.3)–(3.4). We note that Ascher and Lin [1] and Petzold et al. [29] discuss the relationship of related regularization techniques to Baumgarte stabilization and trust region methods.

4. Index-three problems.

4.1. An example. For the DAEs considered thus far, the constraints we have been concerned with coincided with the algebraic equations. For higher-index DAEs, "hidden" constraints occur in addition to the obvious explicit ones. A well-known index-three example in Hessenberg form (cf. Brenan et al. [2]) is

$$\begin{cases} \dot{U}_0 = W_0 + f(t), \\ \dot{V}_0 = U_0 + g(t), \\ V_0 = h(t). \end{cases} \tag{4.1}$$

Providing the nonhomogeneous terms are sufficiently smooth, the unique solution

$$\begin{cases} V_0 = h(t), \\ U_0 = \dot{h}(t) - g(t), \\ W_0 = \ddot{h}(t) - \dot{g}(t) - f(t) \end{cases} \tag{4.2}$$

is obtained by backwards substitution. Note that no initial conditions are needed.

If we differentiate (4.1)(c) once to get $\overset{\bullet}{V}_0$, we find the resulting hidden constraint

$$U_0 + g(t) = \overset{\bullet}{h}(t). \tag{4.3}$$

We will now attempt to regularize the DAE (4.1)–(4.3) by using a Lagrange multiplier λ to account for the original constraint and by using a slack variable w in the hidden constraint. Thus, we consider

$$\begin{cases} \overset{\bullet}{u} = w + f, \\ \overset{\bullet}{v} = u + g + \lambda, \\ \epsilon\lambda = -v + h, \\ \epsilon w = -u - g + \overset{\bullet}{h}. \end{cases} \tag{4.4}$$

Eliminating λ and w yields the two-dimensional singularly perturbed system

$$\begin{cases} \epsilon\overset{\bullet}{u} = -u - g + \overset{\bullet}{h} + \epsilon f, \\ \epsilon\overset{\bullet}{v} = \epsilon u + \epsilon g - v + h. \end{cases} \tag{4.5}$$

Its limiting solution will have the form

$$\begin{cases} u(t,\epsilon) = U_0(t) + \alpha_0(t/\epsilon) + O(\epsilon), \\ v(t,\epsilon) = V_0(t) + \beta_0(t/\epsilon) + O(\epsilon), \end{cases} \tag{4.6}$$

where the outer limit satisfies $U_0(t) = -g + \overset{\bullet}{h}$ and $V_0(t) = h$, the decaying initial layer term $\alpha_0(\tau) = e^{-\tau}(u(0) - U_0(0))$ is nontrivial unless we have the consistent initial value $u(0) = -g(0) + \overset{\bullet}{h}(0)$, and $\beta_0(\tau) = e^{-\tau}[v(0) - V_0(0)]$ is decaying, but nontrivial, unless we have the consistent $v(0) = h(0)$. Since $w = -\frac{1}{\epsilon}(u + g - \overset{\bullet}{h})$, (4.6) implies that

$$w(t,\epsilon) = -\frac{1}{\epsilon}\alpha_0\left(\frac{t}{\epsilon}\right) + O(1), \tag{4.7}$$

i.e., w generally involves a delta function impulse at $t = 0$. Thus, the constraints provoke initial jumps in the constrained quantities.

4.2. A more general problem. In previously considering the DAE

$$\begin{cases} \overset{\bullet}{U}_0 = A(t)U_0 + B(t)X_0 + C(t)Y_0 + f(t), \\ 0 = D(t)U_0 + E(t)X_0 + g(t), \\ 0 = F(t)U_0 + h(t), \end{cases} \tag{4.8}$$

we assumed that the matrices E and FC both remained nonsingular. Let us now make the assumption that

$$\begin{cases} m \geq 2(n-r) - k, \text{ that the } r \times r \text{ matrix } E(t) \\ \text{is nonsingular, and that there is a nonsingular matrix} \\ S(t) \text{ such that } FC = S(t)\begin{pmatrix} H(t) & 0 \\ 0 & 0 \end{pmatrix} S^{-1}(t), \\ \text{where } H(t) \text{ is a } k \times k \text{ matrix of full rank.} \end{cases} \tag{4.9}$$

If we differentiate (4.8)(c) and introduce

$$Y_0 = S(t)\begin{pmatrix} Z_0 \\ W_0 \end{pmatrix},$$

we obtain a hidden constraint

$$(FA + \dot{F})U_0 + FBX_0 + FCS\begin{pmatrix} Z_0 \\ W_0 \end{pmatrix} + Ff + \dot{h} = 0.$$

Multiplying by S, it takes the form

$$(4.10) \qquad \begin{cases} 0 = GU_0 + LX_0 + HZ_0 + \ell, \\ 0 = KU_0 + MX_0 + k. \end{cases}$$

Solving (4.8)(b) for X_0 and (4.10)(a) for Z_0 finally gives us an index-two DAE of the form

$$(4.11) \qquad \begin{cases} \dot{U}_0 = \widetilde{A}(t)U_0 + \widetilde{C}(t)W_0 + \widetilde{f}(t), \\ 0 = \widetilde{K}(t)U_0 + \widetilde{g}(t), \\ 0 = F(t)U_0 + h(t), \end{cases}$$

provided we also assume that

(4.12) the matrix $F(t)\widetilde{C}(t)$ remains nonsingular.

Here $\widetilde{K} \equiv K - ME^{-1}D$ and $\widetilde{C}$ consists of the last $n - r - k$ columns of CS. Using a Lagrange multiplier and a slack variable for (4.11), as we did for (4.4), we can readily determine the $(m-2n+2r+k)$-dimensional subspace of solutions for (4.8)–(4.9)–(4.12).

Detailed study of the canonical forms for higher-index DAEs, as well as their solution structure, could be carried out further. It would be quite analogous to the study of singular arc problems in optimal control (cf. Moylan and Moore [24], O'Malley and Jameson [27], Saberi and Sannuti [30], and Geerts [11]). Our use of a regularizing parameter is analogous to the frequent introduction of a cheap control cost for both analytical and numerical study of singular arcs.

4.3. A final example. Consider the motion of a particle which slides under gravity on a moving, but never vertical, plane

$$(4.13) \qquad a(t)x + b(t)y + z = d(t),$$

with a normal constraining force proportional to a scalar M_0. Newton's laws then imply the system

$$(4.14) \qquad \begin{cases} \ddot{x} + a(t)M_0 = 0, \\ \ddot{y} + b(t)M_0 = 0, \\ \ddot{z} + M_0 = g. \end{cases}$$

Introducing the position $U_0 = (x \;\; y \;\; z)^T$ and the corresponding velocity $V_0 = \dot{U}_0$, we can conveniently rewrite (4.14) as the DAE

$$(4.15) \qquad \begin{cases} \dot{P}_0 = \begin{pmatrix} 0 & I \\ 0 & 0 \end{pmatrix} P_0 - \begin{pmatrix} 0 \\ E^T \end{pmatrix} M_0 + \begin{pmatrix} 0 \\ h \end{pmatrix}, \\ 0 = -(E \;\; 0)P_0 + d, \end{cases}$$

where

$$P_0 = \begin{pmatrix} U_0 \\ V_0 \end{pmatrix},$$

$E = (a\ b\ 1)$, and $h = (0\ \ 0\ \ g)^T$. Differentiating the constraint provides us the additional "hidden" constraint

$$(\dot{E}\ E)P_0 = \dot{d}. \tag{4.16}$$

We will regularize the DAE (4.15)–(4.16) by considering

$$\begin{cases} \dot{p} = \begin{pmatrix} 0 & I \\ 0 & 0 \end{pmatrix} p - \begin{pmatrix} 0 \\ E^T \end{pmatrix} m - \begin{pmatrix} E^T \\ 0 \end{pmatrix} \lambda + \begin{pmatrix} 0 \\ h \end{pmatrix}, \\ -\epsilon\lambda = -(E\ 0)p + d, \\ \epsilon(1 + a^2 + b^2)m = (\dot{E}\ E)p - \dot{d}, \end{cases} \tag{4.17}$$

since

$$(\dot{E}\ E)\begin{pmatrix} 0 \\ E^T \end{pmatrix} = EE^T = 1 + a^2 + b^2 > 0.$$

Rewriting (4.17) as $\dot{u} = v - E^T\lambda$, $\dot{v} = -E^Tm + h$, $\epsilon\lambda = Eu - d$, and $\epsilon(1 + a^2 + b^2)m = \dot{E}u + Ev - \dot{d}$. Thus, we finally obtain the singularly perturbed differential system

$$\begin{cases} \epsilon\dot{u} = -E^TEu + \epsilon v + E^Td, \\ \epsilon\dot{v} = -\dfrac{E^T}{1 + a^2 + b^2}(\dot{E}u + Ev - \dot{d}) + \epsilon h. \end{cases} \tag{4.18}$$

It is a singular system because E^TE has rank 1. We nonetheless seek a solution

$$\begin{cases} u(t,\epsilon) = U(t,\epsilon) + \alpha(\tau,\epsilon), \\ v(t,\epsilon) = V(t,\epsilon) + \beta(\tau,\epsilon) \end{cases} \tag{4.19}$$

with exponentially decaying initial layer corrections α and β.

To proceed further, we note that

$$Q \equiv \frac{E^TE}{1 + a^2 + b^2} \tag{4.20}$$

is a projection such that $QE^T = E^T$ and $EQ = E$. Equating coefficients in (4.18), we find that the first two terms of the outer expansion must satisfy

$$\begin{cases} 0 = -QU_0 + \dfrac{E^Td}{1 + a^2 + b^2}, \\ \dfrac{\dot{U}_0}{1 + a^2 + b^2} = -QU_1 + \dfrac{V_0}{1 + a^2 + b^2}, \\ 0 = -\dfrac{E^T\dot{E}}{1 + a^2 + b^2}U_0 - QV_0 + \dfrac{E^T\dot{d}}{1 + a^2 + b^2}, \\ \dot{V}_0 = -\dfrac{E^T\dot{E}}{1 + a^2 + b^2}U_1 - QV_1 + h. \end{cases} \tag{4.21}$$

(4.21)(a) and (c) imply that

$$
(4.22)\qquad \begin{cases} U_0 = PU_0 + \dfrac{E^T d}{1+a^2+b^2} \\ \text{and} \\ V_0 = PV_0 - \dfrac{E^T \dot{E}}{1+a^2+b^2}(PU_0 + \dfrac{E^T d}{1+a^2+b^2}) + \dfrac{E^T \dot{d}}{1+a^2+b^2}, \end{cases}
$$

where P is the complementary projection to Q. Multiplying (4.21)(b) and (d) by P then yields the linear differential system

$$
(4.23)\qquad \begin{cases} (\dot{PU_0}) = \dot{P}(PU_0) + PV_0 + \dfrac{\dot{P}E^T d}{1+a^2+b^2}, \\ (\dot{PV_0}) = -\dfrac{\dot{P}E^T \dot{E}}{1+a^2+b^2}(PU_0) + \dot{P}(PV_0) + \dfrac{E^T}{1+a^2+b^2}\left(\dot{d} - \dfrac{\dot{E}E^T d}{1+a^2+b^2}\right) + Ph. \end{cases}
$$

Since the initial values for PU_0 and PV_0 are arbitrary, the solution space for (4.13)–(4.14) will be four-dimensional. Moreover, u will feature an initial jump if $u(0)$ is inconsistent with the given constraint, i.e., $E(0)u(0) \neq d(0)$, while v will have an initial jump and M_0 will have initial delta-function behavior unless $v(0)$ satisfies the hidden constraint $\dot{E}(0)u(0) + E(0)v(0) = \dot{d}(0)$.

REFERENCES

[1] U. Ascher and P. Lin, *Sequential regularization methods for higher index DAEs with constraint singularities* I: *Linear index-two case*, SIAM J. Numer. Anal., 33 (1996), to appear.

[2] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *Numerical Solution of Initial Value Problems in Differential-Algebraic Equations*, North–Holland, Amsterdam, 1989.

[3] S. L. Campbell, *Regularization of linear time varying singular systems*, Automatica J. IFAC, 20 (1984), pp. 365–370.

[4] L. O. Chua, *Introduction to Nonlinear Network Theory*, McGraw–Hill, New York, 1969.

[5] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, 1955.

[6] W. O. Criminale and P. G. Drazin, *The evolution of linearized perturbations of parallel flows*, Stud. Appl. Math., 83 (1990), pp. 123–157.

[7] E. Eich and M. Hanke, *Regularization methods for constrained mechanical multibody systems*, preprint.

[8] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959.

[9] C. W. Gear, *Differential-algebraic equation index transformations*, SIAM J. Sci. Stat. Comp., 9 (1988), pp. 39–47.

[10] C. W. Gear and L. R. Petzold, *Differential-algebraic systems and matrix pencils*, in Lecture Notes in Math. 973, Springer-Verlag, Berlin, 1983, pp. 75–89.

[11] T. Geerts, *Structure of linear-quadratic control*, Ph.D. thesis, Technische Universiteit Eindhoven, Eindhoven, The Netherlands, 1989.

[12] P. M. Gresho, S. T. Chan, R. L. Lee, and C. D. Upson, *A modified finite element method for solving the time-dependent incompressible Navier-Stokes equations, part* I: *Theory*, Internat. J. Numer. Methods Fluids, 4 (1984), pp. 557–598.

[13] E. Griepentrog, M. Hanke, and R. März, eds., *Proc. Berlin Seminar on Differential-Algebraic Equations*, Humboldt-Universität, Berlin, 1992.

[14] E. Hairer, C. Lubich, and M. Roche, *The numerical solution of differential-algebraic systems by Runge-Kutta methods*, in Lecture Notes in Math. 1409, Springer-Verlag, Berlin, 1989.

[15] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations* II: *Stiff and Differential-Algebraic Problems*, Springer-Verlag, Berlin, 1991.

[16] M. HANKE, *Regularization methods for higher index differential-algebraic equations*, preprint.

[17] T. J. R. HUGHES, W. K. LIU, AND A. BROOKS, *Finite-element analysis of incompressible viscous flows by the penalty function method*, J. Comput. Physics, 30 (1979), pp. 1–60.

[18] L. V. KALACHEV AND R. E. O'MALLEY, JR., *Regularization of linear differential-algebraic equations*, Tech. report, University of Washington, Seattle, WA, 1991.

[19] ———, *Regularization of nonlinear differential-algebraic equations*, SIAM J. Math. Anal., 25 (1994), pp. 614–629.

[20] M. KNORRENSCHILD, *Regularisierungen von Differentiell-Algebraischen Systemen-Theoretische und Numerische Aspekte*, Ph.D. Thesis, Rhein-Westfälische Technische Hochschule, Aachen, Germany, 1988.

[21] H.-O. KREISS, N. K. NICHOLS, AND D. L. BROWN, *Numerical methods for stiff two-point boundary value problems*, SIAM J. Numer. Anal., 23 (1986), pp. 325–368.

[22] P. LÖTSTEDT, *On the relation between singular perturbation problems and differential-algebraic equations*, Tech. report, Uppsala University, Uppsala, Sweden, 1985.

[23] CH. LUBICH, *Integration of stiff-mechanical systems by Runge–Kutta methods*, preprint.

[24] P. J. MOYLAN AND J. B. MOORE, *Generalizations of singular optimal control theory*, Automatica J. IFAC, 7 (1971), pp. 591–598.

[25] R. E. O'MALLEY, JR., *Boundary value problems for linear systems of ordinary differential equations involving many small parameters*, J. Math. Mech., 18 (1969), pp. 835–856.

[26] ———, *Singular Perturbation Methods for Ordinary Differential Equations*, Springer-Verlag, New York, 1991.

[27] R. E. O'MALLEY, JR. AND A. JAMESON, *Singular perturbations and singular arcs, part* I, IEEE Trans. Automatic Control, 20 (1975), pp. 218–226.

[28] A. OPAL AND J. VLACH, *Consistent initial conditions of linear switched networks*, IEEE Trans. Circuits Systems, 37 (1990), pp. 364–372.

[29] L. R. PETZOLD, Y. REN, AND T. MALY, *Numerical solution of differential-algebraic equations with ill-conditioned constraints*, Tech. report, University of Minnesota, Minneapolis, MN, 1993.

[30] A. SABERI AND P. SANNUTI, *Cheap and singular controls for linear quadratic regulators*, IEEE Trans. Automat. Control, 32 (1987), pp. 208–219.

[31] D. R. SMITH, *Singular Perturbation Theory*, Cambridge University Press, Cambridge, 1985.

[32] R. TEMAM, *Navier-Stokes Equations*, North–Holland, Amsterdam, 1977.

[33] A. N. TIKHONOV AND V. Y. ARSENIN, *Methods of Solving Ill-Posed Problems*, Wiley, New York, 1977.

[34] A. B. VASIL'EVA AND V. F. BUTUZOV, *Singularly Perturbed Equations in the Critical Case*, Moscow State University Press, Moscow, 1978.

[35] A. B. VASIL'EVA, V. F. BUTUZOV, AND L. V. KALACHEV, *The Boundary Function Method for Singular Perturbation Problems*, Society for Industrial and Applied Mathematics, Philadelphia, 1994.

[36] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, Wiley, New York, 1965.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 274–285, January 1996

STABILITY AND CONVERGENCE OF EXTENSION SCHEMES TO CONTINUOUS FUNCTIONS IN GENERAL METRIC SPACES*

E. LE GRUYER[†] AND J. C. ARCHER[†]

Abstract. For any E, E' general metric spaces, we formulate the concept of stability of an extension scheme $\mathcal{E}$ (φ continuous mapping from some closed subset of E into E', $\mathcal{E}(\varphi)$ continuous and extending φ). We show that, when $E' = \mathrm{I\!R}$, stable extension schemes always exist and that the classical extension schemes in the literature are instable. We also show that, when E' is complete, any stable extrapolation scheme $\mathcal{E}$ (φ mapping from some discrete and closed subset of E into E', $\mathcal{E}(\varphi)$ continuous and extending φ) has a unique extension to a stable extension scheme: this result establishes a link between the problem of extrapolation, which usually refers to numerical analysis, and the problem of extension, which also concerns pure mathematics.

Key words. stability, convergence, extension of functionals

AMS subject classifications. 41A05, 46A22, 47B50, 65D05

1. Introduction.

This paper concerns the extension of maps, from some subset of a metric space E into some metric space E', to continuous maps from E to E'. We consider schemes, such as Tietze's scheme, which accomplish such extensions. In particular, we are interested in the *stability* of extension schemes: heuristically, an extension scheme $\mathcal{E}$ to continuous functions from E to E' is said to be *stable* if a small alteration of a datum (φ, x) (φ continuous map from some subset of E to E', $x \in E$) has a small impact on the value $\mathcal{E}(\varphi)(x)$ of the extension $\mathcal{E}(\varphi)$ of φ.

In this article, we give a mathematical formulation of this heuristic concept of stability by introducing the following three notions: DV-stability (DV = data value), which controls errors in φ-values; Ω-stability, which controls errors in x; and DS-stability (DS = data site), which controls errors in the domain of φ. Aside from formulation, DV-stability and Ω-stability have been known for a long time. As far as we know, DS-stability appears for the first time in [1].

Our first main result (Theorem 3.3) states that, when $E' = \mathrm{I\!R}$, stable and convergent extension schemes exist for any metric space E. Here the starting point is a paper by Mc Shane [5] . We highlight this result by showing the instability of some extension schemes in the literature [2]–[5], [7], [8]. In fact, we do not know of any other stable extension scheme with the exception of the above result and the case $E = \mathrm{I\!R}$.

In numerical analysis, one is interested in the extension of maps whose domain (the set of data sites) is discrete and closed, in particular finite. In this case data sites are called *poles*, such an extension is called an *extrapolation*, and the corresponding extension scheme is called an *extrapolation scheme.* With this precision of terminology, our other main result (Theorem 5.1) states that any stable extrapolation scheme extends to a unique stable extension scheme. This result shows that the problem of finding stable extrapolation schemes (only such schemes have an interest in numerical analysis) cannot have an elementary solution.

2. Notations, moduli of continuity.

(E,d), (E',d') denote any two metric spaces.

* Received by the editors September 14, 1993; accepted for publication (in revised form) April 13, 1994.

[†] Institut National de Sciences Appliqueés, 20 Avenue des Buttes de Coësmes, 35043 Rennes cedex, France.

In this paper a *total* (resp., a *partial*) mapping from E to E' is a mapping from E to E' whose domain is E (resp., a subset of E).

$\Pi(E)$ denotes the set of nonempty subsets of E.

$\Delta(E)$ denotes the set of discrete and closed nonempty subsets of E.

$\Omega(E,E')$ denotes the set of Ω-continuous partial mappings (see Definition 2.1) from E to E' whose domain is an element of $\Pi(E)$.

$\mathcal{D}(E,E')$ denotes the set of Ω-continuous partial mappings from E to E' whose domain is an element of $\Delta(E)$ (see Remark 2.1).

We define $\mathcal{C}^o(E,E')$ as the set of total continuous mappings from E to E'.

For $x \in E$, $A \in \Pi(E)$, $d(x,A)$ denotes the usual distance from x to A:

$$d(x,A) := \inf\{d(x,a) : a \in A\}.$$

Let δ denote the (generalised pseudo) Hausdorff distance on $\Pi(E)$:
for $A, B \in \Pi(E)$,

$$\delta(A,B) := \sup\left(\sup_{a\in A} d(a,B), \sup_{b\in B} d(b,A)\right)$$

(in restriction to closed bounded nonempty subsets of E, δ is a distance).

The symbol | denotes restriction to.

For φ, ψ partial mappings from E to E', A a nonempty subset of $\mathrm{dom}(\varphi)\cap\mathrm{dom}(\psi)$, $d'_A(\varphi,\psi)$ denotes the usual (generalised) distance from $\varphi|_A$ to $\psi|_A$:

$$d'_A(\varphi,\psi) := \sup\{d'(\varphi(a),\psi(a)) : a \in A\}.$$

In this paper, a modulus of continuity (resp., a weakly concave modulus of continuity) is a mapping ω from $\mathbb{R}_+$ to $\mathbb{R}_+$ which satisfies

- $\omega(0) = 0$ and ω is continuous at 0,
- ω is increasing : $h_1 \le h_2 \Longrightarrow \omega(h_1) \le \omega(h_2)$,
- ω is subadditive :

$$\forall h_1, h_2 \in \mathbb{R}_+,\ \ \omega(h_1+h_2) \le \omega(h_1) + \omega(h_2) \quad \left(\text{resp.}, h \longmapsto \frac{\omega(h)}{h} \text{ is decreasing}\right).$$

DEFINITION 2.1. *We say that a partial mapping φ from (E,d) to (E',d') is Ω-continuous if a modulus of continuity ω exists which satisfies*

$$\forall x,y \in \mathrm{dom}(\varphi),\ \ d'(\varphi(x),\varphi(y)) \le \omega(d(x,y)). \tag{1}$$

PROPOSITION 2.1. *Let φ be a uniformly continuous partial mapping from E to E'. Then φ is Ω-continuous if and only if $\alpha, \beta \in \mathbb{R}_+$ exist such that, for any $h \in \mathbb{R}_+$, any $x, y \in \mathrm{dom}(\varphi)$, $d(x,y) \le h \Longrightarrow d'(\varphi(x),\varphi(y)) \le \alpha h + \beta$. In particular, if φ is bounded and uniformly continuous then φ is Ω-continuous.*

Proof. Necessity follows from a result by Stechkin which asserts that for any modulus of continuity ω, a weakly concave modulus of continuity $\overline{\omega}$ exists such that $\omega \le \overline{\omega} \le 2\omega$. Sufficiency follows from a lemma by Mc Shane [5, p. 389]. □

Remark 2.1. Any partial mapping from E into E' whose domain is an element of $\Delta(E)$ is continuous but not always Ω-continuous : function $f(n) := n^2$, n an integer, is a simple example of a (uniformly) continuous function that is not Ω-continuous.

Let $\varphi \in \Omega(E,E')$. When $\mathrm{dom}(\varphi)$ is disconnected (it is the general case in extension theory), the usual definition of the modulus of continuity of φ does not apply. We need a slightly stronger definition.

Definition 2.2. *For $\varphi \in \Omega(E, E')$, we define the weakly concave modulus $\overline{\omega}_\varphi$ of continuity of φ as the infimum of the weakly concave moduli of continuity ω which satisfy* (1).

The existence of $\overline{\omega}_\varphi$ follows from Stechkin's result. This outer definition is immediately seen to be equivalent to the following inner one:

$$\overline{\omega}_\varphi := \sup\{\overline{\omega}_{\varphi,a,b} : a, b \in \operatorname{dom}(\varphi)\}$$

where

$$(2) \qquad \begin{array}{rcll} \overline{\omega}_{\varphi,a,b}(h) & := & \inf(h, d(a,b)) \dfrac{d'(\varphi(a), \varphi(b))}{d(a,b)} & \text{for } a \neq b, \\ \overline{\omega}_{\varphi,a,b}(h) & := & 0 & \text{for } a = b. \end{array}$$

We close this section with two properties of weakly concave moduli of continuity which will be essential to the proof of Theorem 3.3.

Proposition 2.2. 1. *Let $\varphi, \psi \in \Omega(E, E')$ with the same domain A. Then*

$$\| \overline{\omega}_\varphi - \overline{\omega}_\psi \|_{\mathbb{R}_+} \leq 2d'_A(\varphi, \psi).$$

2. *Let $\varphi \in \Omega(E, E')$, $A, B \subset \operatorname{dom}(\varphi)$, $A, B \in \Pi(E)$. Then*

$$\| \overline{\omega}_{\varphi|A} - \overline{\omega}_{\varphi|B} \|_{\mathbb{R}_+} \leq 4\overline{\omega}_\varphi(\delta(A, B)).$$

Proof. 1. Let $h \in \mathbb{R}_+$. From Definition 2.2, we have

$$\begin{array}{rcl} |\overline{\omega}_\varphi(h) - \overline{\omega}_\psi(h)| & = & \left| \sup\limits_{a,b \in A} \overline{\omega}_{\varphi,a,b}(h) - \sup\limits_{a,b \in A} \overline{\omega}_{\psi,a,b}(h) \right| \\ & \leq & \sup\limits_{a,b \in A} |\overline{\omega}_{\varphi,a,b}(h) - \overline{\omega}_{\psi,a,b}(h)| . \end{array}$$

But, from (2), we have

$$\begin{array}{rcl} |\overline{\omega}_{\varphi,a,b}(h) - \overline{\omega}_{\psi,a,b}(h)| & \leq & |d'(\varphi(a), \varphi(b)) - d'(\psi(a), \psi(b))| \\ & \leq & d'(\varphi(a), \psi(a)) + d'(\varphi(b), \psi(b)) \\ & \leq & 2d'_A(\varphi, \psi). \end{array}$$

2. First we note that, for any $h \geq 0$, $\overline{\omega}_{\varphi|A}(h) \leq \overline{\omega}_\varphi(h) \leq \overline{\omega}_\varphi(\infty)$. Thus, if $\delta(A, B) = \infty$, we have $\| \overline{\omega}_{\varphi|A} - \overline{\omega}_{\varphi|B} \|_{\mathbb{R}_+} \leq 2\overline{\omega}_\varphi(\delta(A, B))$. Otherwise, set $\delta := \delta(A, B) + \varepsilon$ ($\varepsilon > 0$) and let $h \in \mathbb{R}_+$. Without loss of generality, let us assume that $\overline{\omega}_{\varphi|B}(h) \leq \overline{\omega}_{\varphi|A}(h)$. For each $a \in A$, we choose $b(a) \in B$ such that $d(a, b(a)) \leq \delta$. We have

$$\begin{aligned} \overline{\omega}_{\varphi|A}(h) - \overline{\omega}_{\varphi|B}(h) &\leq \sup_{a_1, a_2 \in A} \overline{\omega}_{\varphi,a_1,a_2}(h) - \sup_{a_1, a_2 \in A} \overline{\omega}_{\varphi,b(a_1),b(a_2)}(h) \\ &\leq \sup_{a_1, a_2 \in A} \left| \overline{\omega}_{\varphi,a_1,a_2}(h) - \overline{\omega}_{\varphi,b(a_1),b(a_2)}(h) \right| . \end{aligned}$$

For the sake of simplicity, from now on, we set $b_1 := b(a_1)$, $b_2 := b(a_2)$, and

$$\delta_{\varphi,a_1,a_2}(h) := |\overline{\omega}_{\varphi,a_1,a_2}(h) - \overline{\omega}_{\varphi,b_1,b_2}(h)|.$$

We must bound $\delta_{\varphi,a_1,a_2}(h)$.

First case. $h \geq d(a_1, a_2)$, $h \geq d(b_1, b_2)$. In this case we have from (2)

$$\delta_{\varphi,a_1,a_2}(h) = |d'(\varphi(a_1), \varphi(a_2)) - d'(\varphi(b_1), \varphi(b_2))|.$$

Thus

$$\delta_{\varphi,a_1,a_2}(h) \leq d'(\varphi(a_1), \varphi(b_1)) + d'(\varphi(a_2), \varphi(b_2)) \leq 2\overline{\omega}_\varphi(\delta).$$

Second case. $h \leq d(a_1, a_2)$, $h \leq d(b_1, b_2)$. In this case, (2) gives

$$\delta_{\varphi,a_1,a_2}(h) = h \left| \frac{d'(\varphi(b_1), \varphi(b_2))}{d(b_1, b_2)} - \frac{d'(\varphi(a_1), \varphi(a_2))}{d(a_1, a_2)} \right|.$$

We write $\delta_{\varphi,a_1,a_2}(h) = |\Delta_1 + \Delta_2|$ where

$$\Delta_1 := h(d(a_1, a_2) - d(b_1, b_2)) \frac{d'(\varphi(b_1), \varphi(b_2)) + d'(\varphi(a_1), \varphi(a_2))}{2d(a_1, a_2)d(b_1, b_2)},$$

$$\Delta_2 := h(d(a_1, a_2) + d(b_1, b_2)) \frac{d'(\varphi(b_1), \varphi(b_2)) - d'(\varphi(a_1), \varphi(a_2))}{2d(a_1, a_2)d(b_1, b_2)}.$$

Since, in our case, $h \leq \inf(d(a_1, a_2), d(b_1, b_2))$ and since, in any case,

$$|d(a_1, a_2) - d(b_1, b_2)| \leq \sup(d(a_1, a_2), d(b_1, b_2)),$$

we have

$$h \frac{|d(a_1, a_2) - d(b_1, b_2)|}{2d(a_1, a_2)d(b_1, b_2)} \leq 1.$$

Using the weak concavity of $\overline{\omega}_\varphi$, we obtain

$$|\Delta_1| \leq \overline{\omega}_\varphi \left(h \frac{|d(a_1, a_2) - d(b_1, b_2)|}{2d(a_1, a_2)} \right) + \overline{\omega}_\varphi \left(h \frac{|d(a_1, a_2) - d(b_1, b_2)|}{2d(b_1, b_2)} \right).$$

Since $h \leq \inf(d(a_1, a_2), d(b_1, b_2))$ and $|d(a_1, a_2) - d(b_1, b_2)| \leq 2\delta$, we obtain

$$|\Delta_1| \leq 2\overline{\omega}_\varphi(\delta).$$

To bound $|\Delta_2|$, we have

$$h \left(\frac{d(a_1, a_2) + d(b_1, b_2)}{2d(a_1, a_2)d(b_1, b_2)} \right) \leq 1.$$

Thus $|\Delta_2| \leq |d'(\varphi(b_1), \varphi(b_2)) - d'(\varphi(a_1), \varphi(a_2))| \leq 2\overline{\omega}_\varphi(\delta)$. Definitively, $\delta_{\varphi,a_1,a_2}(h) \leq 4\overline{\omega}_\varphi(\delta)$ in this second case.

Third case. $d(b_1, b_2) \leq h \leq d(a_1, a_2)$ or $d(a_1, a_2) \leq h \leq d(b_1, b_2)$. Without loss of generality, let us assume that $d(a_1, a_2) \leq h \leq d(b_1, b_2)$. Statement (2) gives in this case

$$\delta_{\varphi,a_1,a_2}(h) = \left| d'(\varphi(a_1), \varphi(a_2)) - \left(\frac{h}{d(b_1, b_2)} \right) d'(\varphi(b_1), \varphi(b_2)) \right|.$$

We write $\delta_{\varphi,a_1,a_2}(h) = |\Delta_3 + \Delta_4|$ where

$$\Delta_3 = \frac{d(b_1,b_2) - h}{d(b_1,b_2)} d'(\varphi(a_1), \varphi(a_2)),$$

$$\Delta_4 = \frac{h}{d(b_1,b_2)} (d'(\varphi(a_1), \varphi(a_2)) - d'(\varphi(b_1), \varphi(b_2)).$$

We have immediately $|\Delta_4| \leq 2\overline{\omega}_\varphi(\delta)$.

Since

$$|\Delta_3| \leq \frac{d(b_1,b_2) - h}{d(b_1,b_2)} \overline{\omega}_\varphi(d(a_1,a_2)),$$

we obtain, using the weak concavity of $\overline{\omega}_\varphi$ again,

$$|\Delta_3| \leq \overline{\omega}_\varphi \left(\frac{d(a_1,a_2)(d(b_1,b_2) - h)}{d(b_1,b_2)} \right).$$

Now, in our case, we have $d(b_1,b_2) - h \leq d(b_1,b_2) - d(a_1,a_2) \leq 2\delta$. Thus $|\Delta_3| \leq 2\overline{\omega}_\varphi(\delta)$. Finally $\delta_{\varphi,a_1,a_2}(h) \leq 4\overline{\omega}_\varphi(\delta)$ in this third case.

As $\varepsilon > 0$ is arbitrary and $\overline{\omega}_\varphi$ is continuous, the stated conclusion follows. □

3. Extension schemes.

DEFINITION 3.1. *We call C^o-extension (resp., C^o-extrapolation) scheme from E to E' any mapping*

$$\mathcal{E} : \Omega(E,E')(resp., \mathcal{D}(E,E')) \longrightarrow C^o(E,E') : \varphi \longmapsto \mathcal{E}(\varphi)$$

which satisfies

$$\forall x \in \mathrm{dom}(\varphi), \quad \mathcal{E}(\varphi)(x) = \varphi(x). \tag{3}$$

We give the following definitions for C^o-extension schemes. Similar definitions hold for C^o-extrapolation schemes.

DEFINITION 3.2. *Let $\mathcal{E}$ be a C^o-extension scheme from E to E'.*

(i) *We say that $\mathcal{E}$ reproduces the constants iff*

$$\forall \varphi \in \mathrm{dom}(\mathcal{E}), \quad \varphi = constant \Longrightarrow \mathcal{E}(\varphi) = (the\ same)\ constant.$$

(ii) *$\mathcal{E}$ is said to be DV-stable iff a positive constant C exists such that for any φ, $\psi \in \mathrm{dom}(\mathcal{E})$ with the same domain A, we have*

$$d'_E(\mathcal{E}(\varphi), \mathcal{E}(\psi)) \leq C d'_A(\varphi, \psi). \tag{4}$$

(iii) *$\mathcal{E}$ is said to be Ω-stable iff a positive constant C exists such that, for any $\varphi \in \mathrm{dom}(\mathcal{E})$,*

$$\overline{\omega}_{\mathcal{E}(\varphi)} \leq C\overline{\omega}_\varphi. \tag{5}$$

(iv) *$\mathcal{E}$ is said to be DS-stable iff a positive constant C exists such that, for any $\varphi \in \mathrm{dom}(\mathcal{E})$, any $A, B \subset \mathrm{dom}(\varphi)$, $A, B \in \Pi(E)$,*

$$d'_E(\mathcal{E}(\varphi|_A), \mathcal{E}(\varphi|_B)) \leq C\overline{\omega}_\varphi(\delta(A,B)). \tag{6}$$

(v) *$\mathcal{E}$ is said to be strongly convergent iff a constant positive C exists such that, for any $\varphi \in \Omega(E,E')$, for any $A \subset \mathrm{dom}(\varphi)$, $A \in \Pi(E)$,*

$$d'_{\mathrm{dom}(\varphi)}(\mathcal{E}(\varphi|_A),\varphi) \leq C\overline{\omega}_\varphi(\delta(A,\mathrm{dom}(\varphi))). \tag{7}$$

The smallest such C (which are attained), in (4)–(7), are called the constants *of DV-stability, Ω-stability, DS-stability, strong convergence.*

Let us remember here the usual definition of a convergent scheme : $\mathcal{E}$ is said to be convergent if, for any total Ω-continuous function f from E into E', for any $F \in \Pi(E)$,

$$\lim_{A\in\Delta(E),\delta(A,F)\to 0} d'_F(\mathcal{E}(f|_A),f) = 0. \tag{8}$$

Remark 3.1. It is easy to check that the univariate piecewise linear extrapolation scheme is a DV-stable, Ω-stable, DS-stable, strongly convergent $\mathcal{C}^o$-extrapolation scheme. The constants of DV-stability, strong convergence, DS-stability are resp., 1, 2, 2. The constant of Ω-stability is smaller than 2 (this constant is 1 if we use concave moduli of continuity in place of weakly concave ones).

The following proposition contains elementary formal consequences of above definitions. Here, $\mathcal{E}$ denotes any extrapolation or extension scheme from E to E'.

PROPOSITION 3.1. 1. *If $\mathcal{E}$ is strongly convergent, then $\mathcal{E}$ is convergent and $\mathcal{E}$ reproduces the constants.*

2. *If $\mathcal{E}$ is DS-stable or Ω-stable, then $\mathcal{E}$ is strongly convergent.*

Proof. 1. The proof of part (1) is obvious.

2. The result is immediate if $\mathcal{E}$ is DS-stable. Let us assume now that $\mathcal{E}$ is Ω-stable and let $\varphi \in \Omega(E,E')$, $A \in \Pi(E)$, $A \subset \mathrm{dom}(\varphi)$, $x \in \mathrm{dom}(\varphi)$. For any $\eta > 0$, we choose $y \in A$ such that $d(x,y) \leq d(x,A) + \eta$. Since $y \in A$, $A \subset B$ and since $\mathcal{E}$ is an extrapolation (or extension) scheme, we have

$$\varphi(y) = \mathcal{E}(\varphi|_A)(y) = \mathcal{E}(\varphi|_B)(y).$$

Thus

$$d'(\mathcal{E}(\varphi|_A)(x),\varphi(x)) \leq d'(\mathcal{E}(\varphi|_A)(x),\mathcal{E}(\varphi|_A)(y)) + d'(\varphi(y),\varphi(x)).$$

Applying the Ω-stability of $\mathcal{E}$, we obtain

$$\begin{array}{rcl} d'(\mathcal{E}(\varphi|_A)(x),\varphi(x)) & \leq & (1+C)\overline{\omega}_\varphi(d(x,A)+\eta) \\ & \leq & (1+C)\overline{\omega}_\varphi(\delta(A,\mathrm{dom}(\varphi))+\eta). \end{array}$$

The result follows since $\overline{\omega}_\varphi$ is continuous and η is arbitrary and since the above inequality holds for any $x \in \mathrm{dom}(\varphi)$. □

Now, we turn to the first main result of this paper. Our starting point is the following result by Mc Shane [5] (see also [8, p. 63, footnote]).

THEOREM 3.2. *Let (E,d) be any metric space and ω be a modulus of continuity. Then, for any partial mapping φ from E to $\mathbb{R}$ which satisfies*

$$\forall y,z \in \mathrm{dom}(\varphi),\ \ |\varphi(y)-\varphi(z)| \leq \omega(d(y,z)), \tag{9}$$

the mapping $\mathcal{M}(\varphi)$ defined, for any $x \in E$, by

$$\mathcal{M}(\varphi)(x) := \sup\{\varphi(a)-\omega(d(x,a)) : a \in \mathrm{dom}(\varphi)\}, \tag{10}$$

extends φ and satisfies

$$\forall y, z \in E, \quad |\mathcal{M}(\varphi)(y) - \mathcal{M}(\varphi)(z)| \leq \omega(d(y.z)). \tag{11}$$

This scheme $\mathcal{M}$, restricted to those φ which satisfy (9), is Ω-instable and DS-instable; see §4.6. We can however adapt (10) to obtain a stable $\mathcal{C}^o$-extension scheme.

THEOREM 3.3. *Let (E, d) be any metric space. Let us define a scheme $\mathcal{E}$ as follows. For $\varphi \in \Omega(E, \mathbb{R})$, $x \in E$,*

$$\mathcal{E}(\varphi)(x) := \sup\{\varphi(a) - \overline{\omega}_\varphi(d(x,a)) : a \in \mathrm{dom}(\varphi)\}. \tag{12}$$

Then $\mathcal{E}$ is a $\mathcal{C}^o$-extension scheme which is Ω-stable with 1 as constant of Ω-stability, DV-stable with a constant of DV-stability smaller than 3, and DS-stable with a constant of DS-stability smaller than 6. Moreover, for bounded φ, we have, for any $x \in E$,

$$\inf\{\varphi(a) : a \in \mathrm{dom}(\varphi)\} \leq \mathcal{E}(\varphi)(x) \leq \sup\{\varphi(a) : a \in \mathrm{dom}(\varphi)\}. \tag{13}$$

Proof. Let $\varphi \in \Omega(E, \mathbb{R})$. Since, by the definition of $\overline{\omega}_\varphi$, we have

$$|\varphi(y) - \varphi(z)| \leq \overline{\omega}_\varphi(d(y,z))$$

for any $y, z \in \mathrm{dom}(\varphi)$, Mc Shane's theorem applies. Therefore $\mathcal{E}$ is Ω-stable with 1 as a constant of Ω-stability.

Now let $\varphi, \psi \in \Omega(E, \mathbb{R})$ with the same domain A. From (12) we have, for any $x \in E$,

$$\begin{aligned}|\mathcal{E}(\varphi)(x) - \mathcal{E}(\psi)(x)| &\leq \sup\{|\varphi(a) - \overline{\omega}_\varphi(d(x,a)) - \psi(a) + \overline{\omega}_\psi(d(x,a))| : a \in A\} \\ &\leq \sup\{|\varphi(a) - \psi(a)| : a \in A\} \\ &\quad + \sup\{|\overline{\omega}_\varphi(d(x,a)) - \overline{\omega}_\psi(d(x,a))| : a \in A\}.\end{aligned}$$

Now, using Proposition 2.2(1), we have

$$\sup\{\overline{\omega}_\varphi(d(x,a)) - \overline{\omega}_\psi(d(x,a))| : a \in A\} \leq 2d'_A(\varphi, \psi).$$

Therefore $\mathcal{E}$ is DV-stable with a constant of DV-stability smaller than 3.

Now we prove the DS-stability of $\mathcal{E}$. Let $\varphi \in \Omega(E, \mathbb{R})$, A, B be nonempty subsets of $\mathrm{dom}(\varphi)$, and $x \in E$. Without loss of generality, let us assume that

$$\mathcal{E}(\varphi|_B)(x) \leq \mathcal{E}(\varphi|_A)(x).$$

If $\delta(A, B) = \infty$, let $b \in B$. From (12), we have

$$|\mathcal{E}(\varphi|_A)(x) - \mathcal{E}(\varphi|_B)(x)| \leq \sup\{|\varphi(a) - \overline{\omega}_{\varphi|_A}(d(x,a)) - \varphi(b) + \overline{\omega}_{\varphi|_B}(d(x,b))| : a \in A\}.$$

Now, for any $a \in A$, we have

$$\sup\{|\varphi(a) - \varphi(b)|, \overline{\omega}_{\varphi|_A}(d(x,a)), \overline{\omega}_{\varphi|_B}(d(x,b))\} \leq \overline{\omega}_\varphi(\infty).$$

Therefore, in this case, we obtain $\mathcal{E}(\varphi|_A)(x) - \mathcal{E}(\varphi|_B)(x)| \leq 3\overline{\omega}_\varphi(\delta(A, B))$.

If $\delta(A, B) < \infty$, let us set $\delta := \delta(A, B) + \eta$ ($\eta > 0$). For each $a \in A$, we choose $b(a) \in B$ such that $d(a, b(a)) \leq \delta$. From (12) we have

$$\begin{aligned}&|\mathcal{E}(\varphi|_A)(x) - \mathcal{E}(\varphi|_B)(x)| \\ &\quad\leq \sup_{a \in A}\{|\varphi(a) - \overline{\omega}_{\varphi|_A}(d(x,a)) - \varphi(b(a)) + \overline{\omega}_{\varphi|_B}(d(x, b(a)))|\} \\ &\quad\leq \sup\{|\varphi(a) - \varphi(b(a))| : a \in A\} \\ &\qquad + \sup\{|\overline{\omega}_{\varphi|_B}(d(x, b(a))) - \overline{\omega}_{\varphi|_A}(d(x,a))| : a \in A\}.\end{aligned}$$

By definition of $\overline{\omega}_\varphi$, we have $|\varphi(a)-\varphi(b(a))| \le \overline{\omega}_\varphi(\delta)$.

Now, setting $h := d(x,a)$, $k := d(x,b(a))$, we have

$$|\overline{\omega}_{\varphi|_B}(k) - \overline{\omega}_{\varphi|_A}(h)| \le |\overline{\omega}_{\varphi|_B}(k) - \overline{\omega}_{\varphi|_B}(h)| + |\overline{\omega}_{\varphi|_B}(h) - \overline{\omega}_{\varphi|_A}(h)|.$$

Since $\overline{\omega}_{\varphi|_B}$ is subadditive, we have

$$|\overline{\omega}_{\varphi|_B}(k) - \overline{\omega}_{\varphi|_B}(h)| \le \overline{\omega}_{\varphi|_B}(|h-k|) \le \overline{\omega}_{\varphi|_B}(\delta) \le \overline{\omega}_\varphi(\delta),$$

and, using Proposition 2.2(2),

$$|\overline{\omega}_{\varphi|_B}(h) - \overline{\omega}_{\varphi|_A}(h)| \le 4\overline{\omega}_\varphi(\delta).$$

Since $\eta > 0$ is arbitrary and since $\overline{\omega}_\varphi$ is continuous, we infer that $\mathcal{E}$ is DS-stable with a constant of DS-stability smaller than 6.

Finally, let us assume that $\varphi \in \Omega(E,\mathrm{I\!R})$ is bounded. By (12), we have $\mathcal{E}(\varphi)(x) \le \sup\{\varphi(a) : a \in \mathrm{dom}(\varphi)\}$. Now let $\eta > 0$ and choose $b \in A$ such that

$$\varphi(b) \ge \sup\{\varphi(a) : a \in \mathrm{dom}(\varphi)\} - \eta.$$

We have, from (12) again,

$$\mathcal{E}(\varphi)(x) \ge \varphi(b) - \overline{\omega}_\varphi(d(x,b)).$$

Now, by definition of $\overline{\omega}_\varphi$, we have

$$\begin{aligned} \overline{\omega}_\varphi(d(x,b)) &\le \sup\{\varphi(a) : a \in \mathrm{dom}(\varphi)\} - \inf\{\varphi(a) : a \in \mathrm{dom}(\varphi)\} \\ &\le \varphi(b) - \inf\{\varphi(a) : a \in \mathrm{dom}(\varphi)\} + \eta. \end{aligned}$$

Therefore $\mathcal{E}(\varphi)(x) \ge \inf\{\varphi(a) : a \in \mathrm{dom}(\varphi)\} - \eta$. The result follows since η is arbitrary. □

Remark 3.2. Henceforth, we denote by $\overline{\mathcal{E}}^-$ the $\mathcal{E}$ described in (12). We also define, for $\varphi \in \Omega(E,\mathrm{I\!R})$, $x \in E$,

$$\overline{\mathcal{E}}^+(\varphi)(x) := \inf\{\varphi(y) + \overline{\omega}_\varphi(d(x,y)) : y \in \mathrm{dom}(\varphi)\}. \tag{14}$$

It can be seen that $\overline{\mathcal{E}}^+$ satisfies Theorem 3.3, that $\overline{\mathcal{E}}^-(\varphi) \le \overline{\mathcal{E}}^+(\varphi)$ for any $\varphi \in \Omega(E,\mathrm{I\!R})$, and that $\overline{\mathcal{E}}^-(\varphi) < \overline{\mathcal{E}}^+(\varphi)$ for most of $\varphi \in \Omega(E,\mathrm{I\!R})$.

4. Study of the stability of some $\mathcal{C}^o$-extension schemes in the literature.

4.1. We begin with the scheme $\mathcal{T}$ yielded by the proof of Urysohn–Tietze's extension theorem from Dieudonné's handbook [2, pp. 89, 90]. We need a description of $\mathcal{T}(\varphi)$ only for those $\varphi \in \Omega(E,\mathrm{I\!R})$ such that $\inf\{\varphi(x) : x \in \mathrm{dom}(\varphi)\} = 1$ and $\sup\{\varphi(x) : x \in \mathrm{dom}(\varphi)\} = 2$:

$$\begin{aligned} \mathcal{T}(\varphi)(a) &:= \varphi(a) && \text{for } a \in \mathrm{dom}(\varphi), \\ \mathcal{T}(\varphi)(x) &:= \inf_{a \in \mathrm{dom}(\varphi)} \left\{ \frac{\varphi(a)d(x,a)}{d(x,\mathrm{dom}(\varphi))} \right\} && \text{for } x \notin \mathrm{dom}(\varphi). \end{aligned}$$

Let us show that $\mathcal{T}$ is Ω-instable in every metric space into which a compact interval $[\alpha,\beta]$ of $\mathrm{I\!R}$ can be metrically embedded. It is sufficient to prove the result when $E = [1,2]$. For $A := \{1, a, b, 2\} (1 \leq a \leq b \leq 2)$ and $f :=$ identity, we obtain from the definition of $\mathcal{T}$

$$\mathcal{T}(f|_A)(x) = \frac{a(x-a)}{b-x} \quad \text{for} \quad \frac{a+b}{2} \leq x \leq \frac{a^2+b^2}{a+b}.$$

Now, for any strictly positive integer n, let us set

$$a_n := \frac{3}{2} - \frac{1}{2n}, \quad b_n := \frac{3}{2} + \frac{1}{2n}, \quad A_n := \{1, a_n, b_n, 2\}, \quad x_n := \frac{3}{2}, \quad y_n := \frac{a_n^2 + b_n^2}{a_n + b_n}.$$

A calculation shows that

$$y_n - x_n = \frac{1}{6n^2} \quad \text{and} \quad \mathcal{T}(f|_{A_n})(y_n) - \mathcal{T}(f|_{A_n})(x_n) = \frac{1}{n}.$$

Thus $\mathcal{T}$ cannot be Ω-stable.

4.2. Now we consider the original proof of the Urysohn–Tietze extension theorem. In the metric case, this proof induces a $\mathcal{C}^o$-extension scheme $\mathcal{U}$ which is DS-instable in every metric space in which a compact interval $[\alpha, \beta]$ $(\alpha < \beta)$ of $\mathbb{R}$ can be metrically embedded. It is sufficient to prove the result when $E = [-1, 1]$. Let us recall the definition of $\mathcal{U}(\varphi)$ for $\varphi \in \mathcal{D}(E, E)$. Setting

$$A := \left\{x \in \operatorname{dom}(\varphi) : -1 \leq \varphi(x) \leq -\frac{1}{3}\right\}, \quad B := \left\{x \in \operatorname{dom}(\varphi) : \frac{1}{3} \leq \varphi(x) \leq 1\right\},$$

we first define

$$g_0 : E \longrightarrow E : x \longmapsto \frac{1}{3} \frac{d(x,A) - d(x,B)}{d(x,A) + d(x,B)}.$$

Noticing that $\frac{3}{2}(\varphi - g_0) \in \mathcal{D}(E, E)$, we can reiterate this definition with $\frac{3}{2}(\varphi - g_0)$ in place of φ, producing g_1 in place of g_0, etc. Then $\mathcal{U}(\varphi)$ is defined as $\sum_{n=0}^{\infty} g_n$.

Now, let $A_\varepsilon := \{-1, -\frac{1}{3} - \frac{\varepsilon}{2}, \frac{1}{3} + \frac{\varepsilon}{2}, 1\}$, $B_\varepsilon := \{-1, -\frac{1}{3} + \frac{\varepsilon}{2}, \frac{1}{3} - \frac{\varepsilon}{2}, 1\}$, and $f :=$ identity. Computer-aided computations show that

$$\lim_{\varepsilon \to 0} \left| \mathcal{U}(f|_{A_\varepsilon}) \left(\frac{2}{3}\right) - \mathcal{U}(f|_{B_\varepsilon}) \left(\frac{2}{3}\right) \right| \simeq \frac{1}{6}.$$

Since $\lim_{\varepsilon \to 0} \delta(A_\varepsilon, B_\varepsilon) = 0$, we infer that $\mathcal{U}$ cannot be DS-stable.

4.3. Valentine [7, p. 107] proved, in a way different from Mc Shane, that any partial Ω-continuous function φ from a metric space (E, d) to $\mathbb{R}$ can be extended to the whole space by a function $\overline{\varphi}$ which has the same weakly concave modulus of continuity. His proof is an adaptation of the proof of the Hahn–Banach extension theorem; the axiom of choice is used via Zorn's lemma. Now, using the axiom of choice again, we can select such a $\overline{\varphi}$ for each φ and therefore obtain an Ω-stable $\mathcal{C}^o$-extension scheme with 1 as the constant of Ω-stability. It is hopeless to try to prove the DV-stability and the DS-stability of such a scheme.

4.4. Dugundji [3] proved that any continuous partial function φ of closed domain from a metric space to a locally convex linear space can be extended to a continuous function on the whole space (it is, as far as we know, the strongest result about the extension problem to continuous functions). The construction of the extension needs a locally finite refinement of a certain covering of the open set complement of $\operatorname{dom}(\varphi)$. Since the construction of this refinement [6] uses the axiom of choice (via Zermelo's theorem), it is hopeless to prove the DS-stability of this scheme. Notice that Dugundji's extension scheme is linear.

4.5. When $E = \mathbb{R}^d$, $d \geq 1$, and E' is any real Banach space, there exist *linear* DV-stable and Ω-stable extension schemes. This result, which holds in the more general context of extension to $\mathcal{C}^m$-functions, is essentially due to Glaeser [4]. In fact, the linear scheme described in Whitney's original extension theorem [8] is DV-instable and Ω-instable but a slight modification of Whitney's construction yields a linear, DV-stable, and Ω-stable extension scheme. Both original and modified schemes are DS-instable.

Excluding the univariate case, we do not know of any linear DV+Ω+DS-stable extension scheme to continuous functions.

4.6. Let us fix a modulus of continuity ω and consider the extension scheme $\mathcal{M}$ defined by (10) for partial functions φ satisfying (9). It can be seen that $\mathcal{M}$ is DV-stable with 1 as the constant of DV-stability and that, for any φ satisfying (9), any $A, B \in \Pi(E)$, $A, B \subset \operatorname{dom}(\varphi)$,

$$d'_E(\mathcal{M}(\varphi|_A), \mathcal{M}(\varphi|_B)) \leq 2\omega(\delta(A,B)). \tag{15}$$

Proofs begin as in Theorem 3.3 but Proposition 2.2 is useless. We notice that (15) is not DS-stable and (11) is not Ω-stable. Heuristically, we can say that (15) (resp., (11)) is as far from DS-stability (resp., Ω-stability) as $\overline{\omega}_\varphi$ is far from ω.

5. Link between stable extrapolation schemes and stable extension schemes.

THEOREM 5.1. *Let (E', d') be a complete metric space. Let us assume that $\mathcal{E}$ is a DS-stable $\mathcal{C}^o$-extrapolation scheme. Then $\mathcal{E}$ extends to a unique $\mathcal{C}^o$-extension scheme $\overline{\mathcal{E}}$ which is also DS-stable with the same constant of DS-stability. Moreover, if $\mathcal{E}$ is DV-stable and/or Ω-stable, $\overline{\mathcal{E}}$ also is with the same constants of stability.*

Proof. Let $\varphi \in \Omega(E, E')$ and set $K := \operatorname{dom}(\varphi)$,

$$\Delta(K) := \{A \in \Delta(E) : A \subset K\}, \quad \Pi(K) := \{A \in \Pi(E) : A \subset K\}.$$

Let us choose a sequence $(A_n)_{n \in \mathbb{N}}$ of elements of $\Delta(K)$ such that

$$\lim_{n \to +\infty} \delta(A_n, K) = 0.$$

Such a sequence exists because E is paracompact [6]. Using the DS-stability of $\mathcal{E}$, the sequence $\mathcal{E}(\varphi|_{A_n})(x)$ is, for any $x \in E$, a Cauchy sequence. Let us denote the limit by $\overline{\mathcal{E}}(\varphi)(x)$. Again by DS-stability of $\mathcal{E}$, the sequence $(\mathcal{E}(\varphi|_{A_n}))_{n \in \mathbb{N}}$ converges uniformly to $\overline{\mathcal{E}}(\varphi)$. Therefore $\overline{\mathcal{E}}(\varphi)$ is continuous. By DS-stability of $\mathcal{E}$ once more, $\overline{\mathcal{E}}(\varphi)$ does not depend on the sequence $(A_n)_{n \in \mathbb{N}}$. Choosing, for each $a \in K$, a sequence $(A_n)_{n \in \mathbb{N}}$ such that $a \in A_n$, for any $n \in \mathbb{N}$, we infer that

$$\overline{\mathcal{E}}(\varphi) = \lim_{n \to +\infty} \mathcal{E}(\varphi|_{A_n})(a) = \varphi(a)$$

because $\mathcal{E}$ is an extrapolation scheme. Therefore $\overline{\mathcal{E}}$ is an extension scheme.

Now let $A, B \in \Pi(K)$. Let us choose

$$(A_n)_{n\in\mathbb{N}},\ A_n \in \Delta(K),\ A_n \subset A,\ \lim_{n\to+\infty} \delta(A_n, A) = 0,$$

$$(B_n)_{n\in\mathbb{N}},\ B_n \in \Delta(K),\ B_n \subset B,\ \lim_{n\to+\infty} \delta(B_n, B) = 0.$$

We have

$$\begin{aligned} d'_E(\overline{\mathcal{E}}(\varphi|_A), \overline{\mathcal{E}}(\varphi|_B)) \leq\ & d'_E(\overline{\mathcal{E}}(\varphi|_A), \mathcal{E}(\varphi|_{A_n})) \\ & + d'_E(\mathcal{E}(\varphi|_{A_n}), \mathcal{E}(\varphi|_{B_n})) \\ & + d'_E(\mathcal{E}(\varphi|_{B_n}), \overline{\mathcal{E}}(\varphi|_B)). \end{aligned}$$

Since $\mathcal{E}$ is DS-stable, we have

$$d'_E(\mathcal{E}(\varphi|_{A_n}), \mathcal{E}(\varphi|_{B_n})) \leq C\overline{\omega}_\varphi(\delta(A_n, B_n)) \leq C\overline{\omega}_\varphi(\delta(A, B) + \delta(A, A_n) + \delta(B, B_n)),$$

where C denotes the constant of DS-stability of $\mathcal{E}$. Letting $n \longrightarrow \infty$, using the definition of $\overline{\mathcal{E}}$ and the continuity of $\overline{\omega}_\varphi$, we obtain

$$d'_E(\overline{\mathcal{E}}(\varphi|_A), \overline{\mathcal{E}}(\varphi|_B)) \leq C\overline{\omega}_\varphi(\delta(A, B)).$$

This last inequality means that $\overline{\mathcal{E}}$ is DS-stable and that $\mathcal{E}$ and $\overline{\mathcal{E}}$ have the same constant of DS-stability.

Now, let us assume that $\mathcal{E}$ is Ω-stable with D as constant of Ω-stability and let $\varphi \in \Omega(E, E')$, $x, y \in E$. Set $A := \mathrm{dom}(\varphi)$ and let us choose $A_n \in \Delta(E)$, $A_n \subset A$, $\lim_{n\to\infty} \delta(A_n, A) = 0$. We have

$$\begin{aligned} d'(\overline{\mathcal{E}}(\varphi)(x), \overline{\mathcal{E}}(\varphi)(y)) \leq\ & d'(\overline{\mathcal{E}}(\varphi)(x), \mathcal{E}(\varphi|_{A_n})(x)) \\ & + d'(\mathcal{E}(\varphi|_{A_n})(x), \mathcal{E}(\varphi|_{A_n})(y)) \\ & + d'(\mathcal{E}(\varphi|_{A_n})(y), \overline{\mathcal{E}}(\varphi)(y)). \end{aligned}$$

Since $\mathcal{E}$ is Ω-stable, we have

$$d'(\mathcal{E}(\varphi|_{A_n})(x), \mathcal{E}(\varphi|_{A_n})(y)) \leq D\overline{\omega}_\varphi(d(x, y)).$$

The Ω-stability of $\overline{\mathcal{E}}$ follows by letting $n \longrightarrow \infty$. The proof of DV-stability is similar. □

PROPOSITION 5.2. *Let $\mathcal{E}$ be an Ω-stable and DV-stable extrapolation scheme from a compact metric space (E, d) to a complete metric space (E', d') in which any closed and bounded subset is compact. Then $\mathcal{E}$ extends to an Ω-stable extension scheme with the same constant of Ω-stability.*

Proof. Let $f \in \Omega(E, E')$ and set $A := \mathrm{dom}(f)$. We have to define $\mathcal{E}(f)$ for $A \in \Pi(E)\backslash\Delta(E)$. Since E is compact, elements of $\Delta(E)$ are finite nonempty subsets of E. Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of finite nonempty subsets of A such that $\lim_{n\to\infty} \delta(A_n, A) = 0$. Since $\mathcal{E}$ is Ω-stable and DV-stable, sequence $(\mathcal{E}(f|_{A_n}))_{n\in\mathbb{N}}$ is equicontinuous and equibounded. Therefore, by Ascoli's theorem, there exists a subsequence $(B_n)_{n\in\mathbb{N}}$ of $(A_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} \delta(B_n, A) = 0$ and $(\mathcal{E}(f|_{B_n}))_{n\in\mathbb{N}}$ converges for the supremun norm. It remains to check that the limit of this subsequence is the desired $\mathcal{E}(f)$ with the required property. □

Remark 5.1. The use of Ascoli's theorem does not allow us to say anything about the uniqueness of the extension. In fact, we cannot expect uniqueness from the hypotheses of Proposition 5.2. A counterexample is as follows. For E a compact infinite metric space, let us define an *extrapolation* scheme $\mathcal{E}$: for $\varphi \in \mathcal{D}(E, \mathrm{I\!R})$, $\mathcal{E}(\varphi)$ is $\overline{\mathcal{E}}^{-}(\varphi)$ (resp., $\overline{\mathcal{E}}^{+}(\varphi)$) ($\overline{\mathcal{E}}^{-}$ and $\overline{\mathcal{E}}^{+}$ have been defined in Remark 3.2) if the number of elements of dom(φ) is even (resp., odd). Now, let us define two *extension* schemes $\mathcal{E}_1$ and $\mathcal{E}_2$ by

$$\begin{array}{ll} \mathcal{E}_1(\varphi) = \mathcal{E}_2(\varphi) = \mathcal{E}(\varphi) & \text{for } \varphi \in \mathcal{D}(E, \mathrm{I\!R}), \\ \mathcal{E}_1(\varphi) = \overline{\mathcal{E}}^{-}(\varphi),\ \mathcal{E}_2(\varphi) = \overline{\mathcal{E}}^{+}(\varphi) & \text{for } \varphi \in \Omega(E, \mathrm{I\!R}) \backslash \mathcal{D}(E, \mathrm{I\!R}). \end{array}$$

Using Theorem 3.3 and since DV-stability and Ω-stability involve only *fixed* sets of data sites, $\mathcal{E}$ is a DV-stable and Ω-stable extrapolation scheme and $\mathcal{E}_1$ and $\mathcal{E}_2$ are distinct (see Remark 3.2) DV-stable and Ω-stable extension schemes which both extend $\mathcal{E}$. This remark highlights Theorem 5.1 and, in fact, the very notion of DS-stability.

Acknowledgments. The authors thank the referees for their helpful suggestions and comments on improving the original manuscript.

REFERENCES

[1] J. C. Archer and E. Le Gruyer, *A constructive proof of a Whitney extension theorem in one variable*, J. Approx. Theory, 71 (1992), pp. 312–328.

[2] J. Dieudonné, *Eléments d'analyse*, Fondements de l'Analyse Moderne, Vol. 1, Gauthier–Villars, Paris, 1968.

[3] J. Dugundji, *An extension of Tietze's theorem*, Pacific J. Math., 1 (1951), pp. 353–367.

[4] G. Glaeser, *Etude de quelques algèbres tayloriennes*, J. Anal. Math. Jerusalem, 6 (1958), pp. 1–124.

[5] E. J. Mc Shane, *Extension of range of functions*, Bull. Amer. Math. Soc., 40 (1934), pp. 837–842.

[6] A. H. Stone, *Paracompactness and product spaces*, Bull. Amer. Math. Soc., 54 (1948), pp. 969–977.

[7] F. A. Valentine, *On the extension of a vector function so as to preserve a Lipschitz condition*, Bull. Amer. Math. Soc., 49 (1943), pp. 100–108.

[8] H. Whitney, *Analytic extensions of differentiable functions defined in closed sets*, Trans. Amer. Math. Soc., 36 (1934), pp. 63–89.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 286–304, January 1996

016

CONVOLUTION OPERATORS FOR RADIAL BASIS APPROXIMATION*

JEREMY LEVESLEY[†], YUAN XU[‡], WILL LIGHT[†], AND WARD CHENEY[§]

Abstract. We construct a large class of continuous integrable functions on $\mathbb{R}^n$ to serve as kernels for approximations to the identity. These kernels are associated with convolution operators that produce approximations to arbitrary continuous functions on $\mathbb{R}^n$ by linear combinations of shifted and dilated radial functions.

Key words. radial basis, convolution, quasi-interpolation

AMS subject classifications. 41A30, 41A35, 41A45, 65D10

1. Introduction. A *radial function* on $\mathbb{R}^n$ is any function of the form $x \mapsto \phi(\|x\|)$, in which $\|\cdot\|$ is the Euclidean norm and ϕ is any real-valued function defined on $[0,\infty)$. A *shifted radial function* is a function of the form $x \mapsto \phi(\|x-v\|)$, where v is a specified point in $\mathbb{R}^n$. The radial function is then sometimes described as being *centered* at v.

It was discovered in the 1970s by Roland Hardy [Ha] that such functions were useful in the interpolation of scattered data. For some time their use rested on empirical results. It was observed, for example, that the surfaces produced by the interpolants possessed very good visual qualities (Franke [F]). In addition, the interpolation matrix seemed always to be nonsingular for the common choice of ϕ (the multiquadric). The important papers of Micchelli [M] and Madych and Nelson [MN], as well as the rediscovery of work by Schoenberg [S1]–[S4], showed that the observation about nonsingularity had a very firm theoretical basis.

Once the interpolation problem was better understood, there arose another important question as to the approximating power of shifted radial functions. Two fundamental tools for studying this problem are convolution techniques (when the centers can be arbitrary) and quasi-interpolation or interpolation techniques, which work best when the centers are regularly distributed throughout $\mathbb{R}^n$. The regularly distributed centers are assumed to be located on the scaled integer grid $h\mathbb{Z}^n$, where $h > 0$. In this case, Fourier transform methods can be used, and results which measure the quality of approximation offered by the set

$$(1.1) \qquad \{\phi(\|\tfrac{\cdot}{h} - z\|) : z \in \mathbb{Z}^n\}$$

in terms of some power of h are obtained. Major developments in this field are due to Jackson and Buhmann [Ji], [Bu], and these results are often somewhat surprising. Neither convolution nor quasi-interpolation can be used in a crude way, as most of the common radial functions exhibit growth at infinity, whereas both these techniques need functions which have rapid decay. However, if we *did* have sufficiently rapid decay

*Received by the editors September 9, 1993; accepted for publication (in revised form) April 27, 1994. This research was supported by NATO grant CRG 910885.

[†]Department of Mathematics, University of Leicester, Leicester LE1 7RH, England.

[‡]Department of Mathematics, University of Oregon, Eugene, OR 97403.

[§]Department of Mathematics, University of Texas at Austin, Austin, TX 78712.

at infinity, then the convolution method is as follows. Suppose that ϕ is continuous and positive on $[0,\infty)$. Assume also that

$$\int_0^\infty t^{n-1}|\phi(t)|\,dt < \infty. \tag{1.2}$$

Then for any $f \in C_0(\mathbb{R}^n)$ and any x,

$$\lim_{k\to\infty} c_k \int_{\mathbb{R}^n} \phi(k\|x-y\|)f(y)\,dy = f(x). \tag{1.3}$$

(Here c_k are certain coefficients depending on ϕ only.) This result is a classical theorem on convolution in disguise. Indeed, if we set $G(x) = \phi(\|x\|)$, we find that G is in $L^1(\mathbb{R}^n)$ since (by passing to polar coordinates),

$$\begin{aligned}\int_{\mathbb{R}^n} |G(x)|\,dx &= \int_0^\infty \int_{S_r^{n-1}} |G(v)|\,dS_r^{n-1}(v)\,dr \\ &= \omega_{n-1}\int_0^\infty \phi(r)r^{n-1}\,dr < \infty.\end{aligned} \tag{1.4}$$

The positivity of ϕ ensures that $\alpha \equiv \int G(x)\,dx \neq 0$. Hence $\alpha^{-1}G$ is a suitable kernel, giving rise to the approximate identity $G_k(x) = \alpha^{-1}k^nG(kx)$. Then by the classical theory, $G_k * f \to f$ pointwise for each $f \in C_0(\mathbb{R}^n)$. The convergence is uniform on compact sets. Refer to [SW, p. 11] for further details. After approximating the integral in equation (1.3) by a quadrature formula, one obtains approximations to f of the form

$$x \mapsto \sum_{j=1}^m a_j\phi(\|kx - v_j\|) \quad (k \in \mathbb{R},\ x, v_j \in \mathbb{R}^n).$$

Such functions are a linear combination of shifted, dilated radial functions. The shifts are the v_j, and the dilation factor is $k > 0$. A byproduct of this analysis is that the set of functions

$$x \mapsto \phi(k\|x-y\|) \qquad (k \in \mathbb{N},\ y \in \mathbb{R}^n)$$

is fundamental in $C_0(\mathbb{R}^n)$; i.e., its linear span is dense. In such assertions, the topology of uniform convergence on compact sets is assumed to be in force on the linear space $C_0(\mathbb{R}^n)$.

As we have already pointed out, when one considers the radial functions used in practice one sees that (1.2) is far too stringent an assumption. For example, (1.2) is clearly not satisfied for the multiquadric referred to previously, where $\phi(t) = \sqrt{t^2+c}$, $c > 0$. Because of this, one needs alternative methods of manufacturing kernels $G \in L^1(\mathbb{R}^n)$. The technique of Jackson was to define G as a finite linear combination of shifts of the radial function. Thus Jackson sought coefficients $a_1, \ldots, a_m$ in $\mathbb{R}$ and centers $u_1, \ldots, u_m$ in $\mathbb{R}^n$ such that the function $G \in C(\mathbb{R}^n)$ defined by

$$G(x) = \sum_{j=1}^m a_j\phi(\|x - u_j\|)$$

satisfies $G \in L^1(\mathbb{R}^n)$ and $\int_{\mathbb{R}^n} G \neq 0$. One of the surprises of the theory is that this apparently easy goal is often unattainable. In a manner which depends critically on the parity of n, one often finds that $G \in L^1(\mathbb{R}^n)$ implies $\int_{\mathbb{R}^n} G = 0$.

Our procedure for creating a useful kernel from a given function ϕ is more general than that of Jackson. First, we average $\phi(\|x-tv\|)$ as v ranges over the unit sphere S^{n-1} in $\mathbb{R}^n$

$$g(x,t)=\frac{1}{\omega_{n-1}}\int_{S^{n-1}}\phi(\|x-tv\|)\,dS^{n-1}(v)\qquad(x\in\mathbb{R}^n,\ t\in\mathbb{R}).\tag{1.5}$$

In this equation, $dS^{n-1}(v)$ represents the rotationally invariant measure on S^{n-1} that is consistent with Lebesgue measure in $\mathbb{R}^n$ (via the standard polar coordinate system). The symbol ω_{n-1} represents the measure of S^{n-1}. Equation (1.5) appears frequently in the literature. See, for example, [Jo, p. 2], [SW, p. 38], [GS, p. 78]. It is readily seen that $g(x,t)$ is a radial function of x and an even function of t.

In the second step, we construct our kernel by putting

$$G(x)=\int_{\mathbb{R}}g(x,t)\,d\mu(t)\qquad(x\in\mathbb{R}^n),\tag{1.6}$$

where μ is a finite, signed, regular, Borel measure having compact support. By manipulating this measure we hope to produce a kernel G such that $G\in L^1(\mathbb{R}^n)$ and $\int_{\mathbb{R}^n}G\neq 0$.

Our subsequent analysis will have the following features:

1. Very weak assumptions are made on ϕ. Roughly speaking, we assume that $\phi=\phi_1+\phi_2$ where $x\mapsto\phi_1(\|x\|)$ is in $L^1(\mathbb{R}^n)$ and ϕ_2 has a power series form for large values of its argument. See Lemma 3.1 for details.
2. Despite the significant increase in generality over Jackson's approach, we still find that $G\in L^1(\mathbb{R}^n)$ often implies $\int_{\mathbb{R}^n}G=0$. Again, the parity of n plays a role, indicating that the reason for this phenomenon is the *compact support* of the measure μ. See Theorem 4.7 for example.
3. We give very general conditions for $\int_{\mathbb{R}^n}G\neq 0$. We suspect these conditions are necessary.

Our analysis needs quite a few preliminary results. Some of these are technical facts about hypergeometric functions. Such results are collected in §2. In §3 we examine the possibility of the kernel constructed being in $L^1(\mathbb{R}^n)$. Finally in §4 we study $\int_{\mathbb{R}^n}G$. We conclude this section with some examples of cases when kernels with nonzero integrals can be constructed.

2. The function *h*.

LEMMA 2.1. *If ϕ and g are as described in §1, then, with $r=\|x\|$ and $\lambda=(n-3)/2$, we have*

$$g(x,t)=\omega_{n-2}\omega_{n-1}^{-1}\int_{-1}^{1}\phi\big[(r^2+t^2-2rst)^{1/2}\big](1-s^2)^{\lambda}\,ds.\tag{2.1}$$

Proof. Fix x,r, and t. Define

$$\psi(s)=\phi\big[(r^2+t^2-2ts)^{1/2}\big]\qquad(-1\le s\le 1).$$

By an equation in [Jo, p. 8] or equation (2.1) in [XLC] we have

$$\begin{aligned}g(x,t)&=\omega_{n-1}^{-1}\int_{S^{n-1}}\Big[\big\{\|x\|^2-2t\langle x,v\rangle+t^2\|v\|^2\big\}^{1/2}\Big]\,dS^{n-1}(v)\\&=\omega_{n-1}^{-1}\int_{S^{n-1}}\psi\big(\langle x,v\rangle\big)\,dS^{n-1}(v)\\&=\omega_{n-2}\omega_{n-1}^{-1}\int_{-1}^{1}\psi(rs)(1-s^2)^{\lambda}\,ds.\qquad\square\end{aligned}$$

A central role in this paper is played by the function

$$h(z) = h(\beta, \lambda, z) = \int_{-1}^{1} (1 + z^2 - 2zs)^{\beta}(1 - s^2)^{\lambda}\, ds \qquad (|z| < 1). \tag{2.2}$$

Here $\beta \in \mathbb{R}$ and $\lambda > -1$. In this section, we shall prove that h is a hypergeometric function and use that information to study the coefficients in its Maclaurin series. We adopt the following standard notation for the Gauss hypergeometric function:

$$F(a, b, c, z) = \sum_{k=0}^{\infty} \frac{(a, k)(b, k)}{(c, k)k!}\, z^k \qquad (|z| < 1). \tag{2.3}$$

In this equation, the Pochhammer symbols are defined by

$$(a, k) = a(a + 1)(a + 2) \cdots (a + k - 1) \qquad (k \geq 1). \tag{2.4}$$

By convention, $(a, 0) = 1$. See [AS, Chap. 15] for information about the hypergeometric function.

LEMMA 2.2. *The function h is even and analytic in the open unit disk of the complex plane.*

Proof. With the substitution $t = -s$ in equation (2.2) we obtain

$$\begin{aligned} h(-z) &= \int_{-1}^{1} (1 + z^2 + 2zs)^{\beta}(1 - s^2)^{\lambda}\, ds \\ &= \int_{-1}^{1} (1 + z^2 - 2zt)^{\beta}(1 - t^2)^{\lambda}\, dt = h(z). \end{aligned}$$

Hence h is even. Since $\lambda > -1$, the factor $(1 - s^2)^{\lambda}$ is integrable. Furthermore, a singularity can occur in the integrand only when $z^2 - 2zs + 1 = 0$, i.e., when $z = s \pm i\sqrt{1 - s^2}$, $s \in [-1, 1]$. These singular points lie on the unit circle, and hence h is analytic in the open disk. □

THEOREM 2.3. *If $\beta \in \mathbb{R}$ and $\lambda > -1$, then for $|z| < 1$,*

$$h(\beta, \lambda, z) = \frac{\Gamma(\lambda + 1)\Gamma(\frac{1}{2})}{\Gamma(\lambda + \frac{3}{2})} F(-\beta, -\lambda - \beta - \tfrac{1}{2}, \lambda + \tfrac{3}{2}, z^2). \tag{2.5}$$

Proof. Make the change of variable $s = 2t - 1$ in equation (2.2) to obtain

$$\begin{aligned} h(\beta, \lambda, z) &= 2 \int_0^1 \left[1 + z^2 - 2z(2t - 1)\right]^{\beta} (2 - 2t)^{\lambda}(2t)^{\lambda}\, dt \\ &= 2^{2\lambda+1} \int_0^1 \left[(1 + z)^2 - 4zt\right]^{\beta} t^{\lambda}(1 - t)^{\lambda}\, dt \qquad (2.6) \\ &= 2^{2\lambda+1}(1 + z)^{2\beta} \int_0^1 \left[1 - \frac{4zt}{(1 + z)^2}\right]^{\beta} t^{\lambda}(1 - t)^{\lambda}\, dt. \end{aligned}$$

We shall now employ several formulæ from [AS]. Each is valid in some neighborhood of the origin, and the result of using them will be a representation of h that is valid in a neighborhood of the origin. Formula 15.3.1 from [AS] is

$$F(a, b, c, z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c - b)} \int_0^1 t^{b-1}(1 - t)^{c-b-1}(1 - tz)^{-a}\, dt. \tag{2.7}$$

Employing equations (2.6) and (2.7) we have

$$h(\beta,\lambda,z)=\frac{2^{2\lambda+1}[\Gamma(\lambda+1)]^2}{\Gamma(2\lambda+2)}(1+z)^{2\beta}F\left(-\beta,\lambda+1,2\lambda+2,\frac{4z}{(1+z)^2}\right). \tag{2.8}$$

Formula 6.1.18 from [AS] is

$$\Gamma(z)\Gamma\left(z+\tfrac{1}{2}\right)=2^{1-2z}\sqrt{\pi}\,\Gamma(2z). \tag{2.9}$$

This leads to

$$\frac{2^{2\lambda+1}[\Gamma(\lambda+1)]^2}{\Gamma(2\lambda+2)}=\frac{\Gamma(\lambda+1)\Gamma(\frac{1}{2})}{\Gamma(\lambda+\frac{3}{2})}. \tag{2.10}$$

Formula 15.3.27 from [AS] is

$$F(a,b,a-b+1,z^2)=(1+z)^{-2a}F\left(a,a-b+\tfrac{1}{2},2a-2b+1,\frac{4z}{(1+z)^2}\right). \tag{2.11}$$

In (2.11) we set $a=-\beta$ and $b=-\lambda-\beta-\frac{1}{2}$. Using equation (2.8), we are led to equation (2.5), valid in some neighborhood of the origin. Since h is analytic in the unit disk, equation (2.5) is valid in the unit disk. □

In the future, we denote $[\Gamma(\lambda+1)\Gamma(\frac{1}{2})]/[\Gamma(\lambda+\frac{3}{2})]$ by γ. It is not zero because the Gamma function has no zeros, and its poles are at $0,-1,-2,\dots$, whereas $\lambda+\frac{3}{2}>\frac{1}{2}$.

LEMMA 2.4. *Let $\beta\in\mathbb{R}$ and let $\lambda>-1$. Let*

$$h(\beta,\lambda,z)=\int_{-1}^{1}(1+z^2-2sz)^{\beta}(1-s^2)^{\lambda}\,ds=\sum_{j=0}^{\infty}c_j(\beta)z^{2j}.$$

The condition $c_j(\beta)=0$ occurs if and only if $j>\beta\in\mathbb{Z}_+$ or $j>\lambda+\beta+\frac{1}{2}\in\mathbb{Z}_+$.

Proof. By Theorem 2.3,

$$c_j(\beta)=\gamma\,\frac{(-\beta,j)(-\lambda-\beta-\frac{1}{2},j)}{(\lambda+\frac{3}{2},j)j!}.$$

Since $\gamma\neq 0$, we see that $c_j(\beta)=0$ if and only if $(-\beta,j)=0$ or $(-\lambda-\beta-\frac{1}{2},j)=0$. Since

$$(-\beta,j)=-\beta(-\beta+1)(-\beta+2)\cdots(-\beta+j-1),$$

we have $(-\beta,j)=0$ if and only if $\beta\in\{0,1,2,\dots,j-1\}$. Equivalently, $j>\beta\in\mathbb{Z}_+$. The other Pochhammer symbol vanishes if and only if $j>\lambda+\beta+\frac{1}{2}\in\mathbb{Z}_+$. □

LEMMA 2.5. *Let $F(a,b,c,x)=\sum_{k=0}^{\infty}A_kx^k$, where $-c\notin\mathbb{Z}_+$. If $b\neq 1$ then*

$$(b-1)^{-1}F(a,b-1,c,x)=\sum_{k=0}^{\infty}A_k(k+b-1)^{-1}x^k.$$

If $1-b\in\mathbb{N}$, the singularity when $k=1-b$ is removable.

Proof. The coefficients in $(b-1)^{-1}F(a,b-1,c,x)$ are

$$\frac{(a,k)(b-1,k)}{(c,k)k!(b-1)}=\frac{(a,k)(b,k-1)}{(c,k)k!}=\frac{(a,k)(b,k)}{(c,k)k!(b+k-1)}=\frac{A_k}{k+b-1}.$$

Since $(b,k-1) = b(b+1)(b+2)\cdots(b+k-2)$, the zeros of $(b,k-1)$ are $b = 0,-1,-2,\ldots,2-k$. Thus if $1-b \in \mathbb{N}$, we have $(b,k-1) = 0$ for $k \geq 2-b$. The last nonzero term in the series will correspond to $k = 1-b$. This term is

$$\frac{(a,k)(b,k-1)}{(c,k)k!}x^k = \frac{(a,k)(1-k,k-1)}{(c,k)k!}x^k = \frac{(a,k)(-1)^{k-1}}{(c,k)k}x^k = \frac{(a,1-b)(-1)^b}{(c,1-b)(1-b)}x^{1-b}.$$

The proof is thus completed. □

LEMMA 2.6. *Define A_k as in Lemma* 2.5. *If $-c \notin \mathbb{Z}_+$ then*

$$c^{-1}F(a,b,c+1,x) = \sum_{k=0}^{\infty} \frac{A_k}{k+c}x^k \qquad (|x|<1). \tag{2.12}$$

Proof. The hypothesis on c ensures that $F(a,b,c,x)$ is well defined. The equality asserted rests upon this identity between the coefficients of x^k on the two sides

$$\frac{(a,k)(b,k)}{(c+1,k)k!c} = \frac{(a,k)(b,k)}{(c,k)k!(k+c)}. \qquad \Box$$

LEMMA 2.7. *Consider the hypergeometric function*

$$F(a,b,c,z) = \sum_{k=0}^{\infty} A_k z^k \qquad (|z|<1). \tag{2.13}$$

If the real part of $(c-a-b)$ is positive, then the series converges absolutely on the circle $|z| = 1$, and Gauss' formula is valid

$$\lim_{x\uparrow 1} F(a,b,c,x) = \frac{\Gamma(c-a-b)\Gamma(c)}{\Gamma(c-a)\Gamma(c-b)}. \tag{2.14}$$

Proof. See [AS, p. 556]. □

LEMMA 2.8. *Let $n \geq 2$, $\alpha > -n$, $\beta = \alpha/2$, $\lambda = (n-3)/2$. Define $c_k(\beta)$ as in Lemma* 2.4. *Then these series converge absolutely*

$$\sum_{k=0}^{\infty} \frac{c_k(\beta)}{n+2k}, \qquad \sum_{k=0}^{\infty} \frac{c_k(\beta)}{n+\alpha-2k}. \tag{2.15}$$

Proof. By Lemmas 2.4 and 2.3, using $c_k = c_k(\beta)$, we have

$$\sum c_k z^{2k} = h(\beta,\lambda,z) = \frac{\Gamma(\lambda+1)\Gamma(\frac{1}{2})}{\Gamma(\lambda+\frac{3}{2})}F(-\beta,-\lambda-\beta-\tfrac{1}{2},\lambda+\tfrac{3}{2},z^2).$$

Let $a = -\beta$, $b = -\lambda-\beta-\frac{1}{2}$, $c = \lambda+\frac{3}{2}$, and $x = z^2$. Then

$$\sum c_k x^k = \gamma F(a,b,c,x).$$

We note that $c \notin -\mathbb{Z}_+$. Hence Lemma 2.6 applies, and we have

$$\sum \frac{c_k}{k+c}x^k = c^{-1}F(a,b,c+1,x).$$

We note also that $b \neq 1$. Hence Lemma 2.5 applies, and we have

$$\sum \frac{c_k}{k+b-1} x^k = (b-1)^{-1} F(a, b-1, c, x).$$

Both of these last two series converge absolutely for $|x| = 1$ by Lemma 2.7. Hence the series $\sum c_k (k+c)^{-1} \equiv 2 \sum c_k (n+2k)^{-1}$ converges. Similarly, the series $\sum c_k (n+\alpha - 2k)^{-1}$ converges. □

LEMMA 2.9. *Define A_k by equation* (2.13), *and assume that* $0 < c-a = 1-b$ *and that* $-c \notin \mathbb{Z}_+$. *Then*

$$\sum_{k=0}^{\infty} \frac{A_k}{k+c} = \sum_{k=0}^{\infty} \frac{A_k}{1-b-k}. \tag{2.16}$$

Proof. Note that $b \neq 1$. Hence, by Lemma 2.5,

$$(1-b)^{-1} F(a, b-1, c, z) = \sum_{k=0}^{\infty} A_k (1-b-k)^{-1} z^k \qquad (|z| < 1). \tag{2.17}$$

Since $c - a - (b-1) = 2(1-b) > 0$, Lemma 2.7 applies, and we conclude that the series in equation (2.17) converges absolutely when $|z| = 1$. In particular, it converges at $z = 1$. Hence by Abel's Theorem [W, p. 320],

$$\lim_{x \uparrow 1} (1-b)^{-1} F(a, b-1, c, x) = \sum_{k=0}^{\infty} A_k (1-b-k)^{-1}.$$

Using Gauss' formula in Lemma 2.7, we obtain

$$\frac{\Gamma(c+1-a-b)\Gamma(c)}{(1-b)\Gamma(c-a)\Gamma(c-b+1)} = \sum_{k=0}^{\infty} A_k (1-b-k)^{-1}. \tag{2.18}$$

In exactly the same way, starting with Lemma 2.6 we obtain

$$\frac{\Gamma(c+1-a-b)\Gamma(c+1)}{c\Gamma(c+1-a)\Gamma(c+1-b)} = \sum_{k=0}^{\infty} A_k (k+c)^{-1}. \tag{2.19}$$

By using the identity $x\Gamma(x) = \Gamma(x+1)$ we see that the left sides in equations (2.19) and (2.18) are equal. □

LEMMA 2.10. *Let* $n \geq 2$, $\lambda = (n-3)/2$, *and* $\beta > -n/2$. *Define* $c_j(\beta)$ *as in Lemma* 2.4. *Then*

$$\sum_{j=0}^{\infty} c_j(\beta)(n+2j)^{-1} = \sum_{j=0}^{\infty} c_j(\beta)(n+2\beta-2j)^{-1}.$$

Proof. Define a, b, c, and x as in the proof of Lemma 2.8. Then the assertion of Lemma 2.8 is valid. The hypotheses of Lemma 2.9 are now fulfilled since

$$\begin{aligned} c - a &= \lambda + \frac{3}{2} + \beta = \frac{n}{2} + \beta > 0, \\ 1 - b &= 1 + \lambda + \beta + \frac{1}{2} = \frac{n}{2} + \beta = c - a, \\ -c &= -\frac{n}{2} \notin \mathbb{Z}_+. \end{aligned}$$

By Lemma 2.9,

$$\sum_{k=0}^{\infty} A_k(k+c)^{-1} = \sum_{k=0}^{\infty} A_k(1-b-k)^{-1}.$$

Here A_k are the Taylor coefficients of $F(a,b,c,x)$. Hence $A_k = \frac{1}{\gamma}c_k(\beta)$. Inserting the values of c and b in the preceding series, we obtain the equation to be proved. □

LEMMA 2.11. *Let $n \geq 2$ and $\alpha > -n$. Define $c_k(\beta)$ as in Lemma 2.4. Define*

$$R(n,\alpha) = \sum_{\substack{j=0 \\ 2j \neq n+\alpha}}^{\infty} c_j\left(\frac{\alpha}{2}\right)\left[\frac{1}{n+2j} - \frac{1}{n+\alpha-2j}\right].$$

For $R(n,\alpha)$ to be nonzero, it is necessary and sufficient that $n+\alpha \in 2\mathbb{N}$ and $\alpha \notin 2\mathbb{Z}_+$.

Proof. If $n+\alpha \notin 2\mathbb{Z}_+$, then in the sum defining $R(n,\alpha)$ the condition $2j \neq n+\alpha$ is vacuous (and can be omitted). In this case, $R(n,\alpha) = 0$ by Lemma 2.10.

Assume, therefore, that $n+\alpha = 2m \in 2\mathbb{Z}_+$. Since $n+\alpha > 0$ this means $n+\alpha \in 2\mathbb{N}$. By Lemma 2.10 we see that

$$-R(n,\alpha) = c_m\left(\frac{\alpha}{2}\right)\left[\frac{1}{n+2m} - \frac{1}{n+\alpha-2m}\right].$$

This is precisely the term in which we must cope with the removable singularity mentioned in Lemma 2.5. As in Lemma 2.8, let $a = -\alpha/2$, $b = -\lambda - \alpha/2 - 1/2$, and $c = n/2$. Then we have

$$c_m\left(\frac{\alpha}{2}\right) = \frac{(a,m)(b,m)}{(c,m)m!} = \frac{(a,m)}{(c,m)m!}(b,m-1)(b+m-1).$$

Since $2m - n - \alpha = 2(b+m-1)$ we have

$$-R(n,\alpha) = \frac{(a,m)}{(c,m)m!}\,\frac{(b,m-1)}{2}.$$

Therefore $R(n,\alpha) = 0$ if and only if $(a,m)(b,m-1) = 0$. Recall that a Pochhammer symbol (x,k) vanishes if and only if $x \in \mathbb{Z}$ and $1-k \leq x \leq 0$. It follows that $(b,m-1) \neq 0$, because

$$b = (-n-\alpha+2)/2 = -m+1 < 1-(m-1).$$

Thus we conclude that $R(n,\alpha) = 0$ if and only if $(a,m) = 0$. This is equivalent in turn to

(i) $a \in \mathbb{Z}$ and $1-m \leq a \leq 0$,
(ii) $\alpha \in 2\mathbb{Z}$ and $1-m \leq -\alpha/2 \leq 0$,
(iii) $\alpha \in 2\mathbb{Z}$ and $0 \leq \alpha \leq n+\alpha-2$,
(iv) $\alpha \in 2\mathbb{Z}_+$. □

3. The integrability of G. Recall from §1 that if a function ϕ has been prescribed, we define a function G on $\mathbb{R}^n$ by the formula

$$G(x) = \frac{\omega_{n-2}}{\omega_{n-1}}\int_{\mathbb{R}}\int_{-1}^{1}\phi\Big(\big[\|x\|^2+t^2-2\|x\|st\big]^{1/2}\Big)(1-s^2)^{\lambda}\,ds\,d\mu(t). \tag{3.1}$$

Here μ is a regular, signed, Borel measure having finite total variation and compact support. Let a be chosen so that the interval $[-a, a]$ contains the support of μ. To ensure the local integrability of G, we assume that $\phi \in C(0,\infty)$ and that $\phi(r) = O(r^\theta)$ as $r \to 0$, for some $\theta > -n$.

Our analysis involves the moments of the measure μ, defined by

$$m_j = \int_{\mathbb{R}} t^j \, d\mu(t) \qquad (j = 0, 1, 2, \ldots).$$

We also require the function $h(z) = h(\beta, \lambda, z)$ from equation (2.2), in which $\lambda = (n-3)/2$.

LEMMA 3.1. *Let ϕ be a function as described above that is expressible for sufficiently large values of r in the form*

$$\phi(r) = \sum_{k=0}^{[n+\alpha]} a_k r^{\alpha-k} + \phi_1(r), \tag{3.2}$$

where $a_0 \neq 0$, $\alpha \in \mathbb{R}$, and $\phi_1(\|x\|) \in L^1(\mathbb{R}^n)$. Let h, μ, G, m_j, a, and c_j be as above. For G to be integrable it is necessary and sufficient that

$$\sum_{k+2j=\nu} a_k c_j \left(\frac{\alpha-k}{2}\right) m_{2j} = 0 \qquad \big(0 \leq \nu \leq n + [\alpha]\big). \tag{3.3}$$

Proof. The process that generates G from ϕ is linear and produces an integrable function from ϕ_1. Hence, without loss of generality, we shall assume that for $r > A$,

$$\phi(r) = \sum_{k=0}^{[n+\alpha]} a_k r^{\alpha-k}.$$

Since G is a radial function

$$\omega_{n-2}^{-1} \int_{\rho \leq \|x\| \leq M} |G(x)| \, dx = \frac{\omega_{n-1}}{\omega_{n-2}} \int_\rho^M r^{n-1} |G_0(r)| \, dr$$

$$= \int_\rho^M r^{n-1} \left| \int_{-a}^{a} \phi\big([r^2 + t^2 - 2rst]^{1/2}\big)(1-s^2)^\lambda \, ds \, d\mu(t) \right| dr$$

$$= \int_\rho^M r^{n-1} \left| \sum_{k=0}^{[n+\alpha]} a_k \int_{-a}^{a} \int_{-1}^{1} (r^2 + t^2 - 2rst)^{(\alpha-k)/2} (1-s^2)^\lambda \, ds \, d\mu(t) \right| dr$$

$$= \int_\rho^M r^{n-1} \left| \sum_{k=0}^{[n+\alpha]} a_k r^{\alpha-k} \int_{-a}^{a} h\left(\frac{\alpha-k}{2}, \lambda, \frac{t}{r}\right) d\mu(t) \right| dr$$

$$= \int_\rho^M \left| \sum_{k=0}^{[n+\alpha]} a_k r^{n-1+\alpha-k} \int_{-a}^{a} \sum_{j=0}^{\infty} c_j \left(\frac{\alpha-k}{2}\right) \left(\frac{t}{r}\right)^{2j} d\mu(t) \right| dr$$

$$= \int_\rho^M \left| \sum_{k=0}^{[n+\alpha]} \sum_{j=0}^{\infty} a_k c_j \left(\frac{\alpha-k}{2}\right) m_{2j} r^{n-1+\alpha-k-2j} \right| dr$$

$$= \int_\rho^M \left| \sum_{\nu=0}^{\infty} r^{n+\alpha-1-\nu} \sum_{k+2j=\nu} a_k c_j \left(\frac{\alpha-k}{2}\right) m_{2j} \right| dr.$$

It is straightforward to prove that

$$\left|c_j\left(\frac{\alpha-k}{2}\right)\right| \le \text{const.}\, k^{2j} \le \text{const.}[n+\alpha+1]^{2j}.$$

Therefore, the double series in the penultimate line of the above calculation is absolutely convergent if ρ is sufficiently large. This justifies the rearrangement of series in the ultimate line.

Our calculation shows that $G \in L^1(\mathbb{R}^n)$ if and only if the coefficient of $r^{n+\alpha-\nu}$ is zero whenever $n+\alpha-1-\nu \ge -1$. This leads to equation (3.3). □

We shall require in our analysis the following result in linear algebra.

LEMMA 3.2. *Define linear functionals $L_\nu(x) = \sum_{k+2j=\nu} a_{kj}x_j$, in which $0 \le \nu \le N$, $k \ge 0$, $j \ge 0$, $a_{00} \ne 0$, and $a_{kj} = 0 \Rightarrow a_{ki} = 0$ for $i \ge j$. Let s be the largest integer for which $\sum_{k=0}^{N-2s} |a_{ks}| > 0$. The following properties of the x-vector are equivalent: (1) $L_\nu(x) = 0$ for $0 \le \nu \le N$; (2) $x_j = 0$ for $0 \le j \le s$.*

Proof. Assume that (2) is true. Then for each ν we have $L_\nu(x) = \sum_{k+2j=\nu,\, j>s} a_{kj}x_j$. By the definition of s, $a_{kj} = 0$ if $j > s$ and $0 \le k \le N-2j$. Hence $L_\nu(x) = 0$.

Now assume that (2) is false. Let α be the first integer for which $x_\alpha \ne 0$. Thus $0 \le \alpha \le s$. By the definition of s, $\sum_{k=0}^{N-2s} |a_{ks}| > 0$. Hence there is an index $\mu \in [0, N-2s]$ for which $a_{\mu s} \ne 0$. By the hypothesis on the array (a_{ij}), we have $a_{\mu\alpha} \ne 0$. Let β be the smallest integer for which $a_{\beta\alpha} \ne 0$. Thus $\beta \le \mu \le N-2s$. Put $\nu = \beta + 2\alpha$. Thus $\nu \le (N-2s)+2s = N$. Also $a_{\nu-2\alpha,\alpha} = a_{\beta\alpha} \ne 0$. If $j > \alpha$, then $\nu - 2j < \nu - 2\alpha = \beta$, and by the definition of β, $a_{\nu-2j,\alpha} = 0$. Since $j > \alpha$, $a_{\nu-2j,j} = 0$. If $j < \alpha$ then by the definition of α, $x_j = 0$. Consequently,

$$L_\nu(x) = \sum_{k+2j=\nu} a_{kj}x_j = \sum_j a_{\nu-2j,j}x_j = a_{\nu-2\alpha,\alpha}x_\alpha = a_{\beta\alpha}x_\alpha \ne 0.$$

Thus (1) is false. □

THEOREM 3.3. *For G to belong to $L^1(\mathbb{R}^n)$ it is necessary and sufficient that $m_{2j} = 0$ for $0 \le j \le s$, where*

$$s = \max\left\{ j : \sum_{k=0}^{n+[\alpha]-2j} \left| a_k c_j\left(\frac{\alpha-k}{2}\right)\right| > 0 \right\}.$$

Proof. We intend to use Lemmas 3.1 and 3.2. Define

$$a_{kj} = a_k c_j\left(\frac{\alpha-k}{2}\right), \quad N = n+[\alpha], \quad \text{and} \quad x_j = m_{2j}.$$

We verify the hypotheses of Lemma 3.2 in the present case. We have $a_{00} = a_0c_0(\alpha/2) \ne 0$ because a_0 has explicitly been assumed to be nonzero, and $c_0(\alpha/2) \ne 0$ by Lemma 2.4. To verify the other hypothesis, suppose that it fails. Let $a_{kj} = 0$, $i > j$, and $a_{ki} \ne 0$. Then $a_k \ne 0$ and $c_j(\frac{\alpha-k}{2}) = 0$. By Lemma 2.4, $c_k(\frac{\alpha-k}{2}) = 0$ also, contradicting the assumption $a_{ki} \ne 0$. By applying Lemma 3.2 to the system of equations (3.3) we obtain the desired conclusion. □

An alternative version of Theorem 3.3 can be obtained by establishing a different formula for the parameter s. Having fixed n and α as above, we define

$$p_k = \max\left\{ j \in \mathbb{Z}_+ : j \le (n+\alpha-k)/2 \text{ and } c_j\left(\frac{\alpha-k}{2}\right) \ne 0\right\}, \tag{3.4}$$

$$Q = \max\{p_k : 0 \le k \le n+\alpha \text{ and } a_k \ne 0\}. \tag{3.5}$$

LEMMA 3.4. $s = Q$.

Proof. By the definition of Q, there is an integer k such that $a_k \neq 0$ and $p_k = Q$. By the definition of p_k, $c_{p_k}(\frac{\alpha-k}{2}) \neq 0$. Hence $a_{k,p_k} \neq 0$. By the definition of s, $a_{kj} = 0$ for $j > s$. Hence $p_k \leq s$ and $Q \leq s$. To prove that $s \leq Q$, we note first that by the definition of s, $\sum_{k=0}^{n+\alpha-2s} |a_{ks}| > 0$. Select an index $k \in [0, n+\alpha-2s]$ such that $a_{ks} \neq 0$. It follows that $a_k \neq 0$ and $c_s(\frac{\alpha-k}{2}) \neq 0$. Since $s \leq (n+\alpha-k)/2$, the definition of p_k yields immediately $p_k \geq s$. Since $k \leq n+\alpha-2s \leq n+\alpha$ and $a_k \neq 0$, the definition of Q yields $Q \geq p_k$. Hence $Q \geq s$. □

LEMMA 3.5. *Define* $p(\beta) = \max\{j \in \mathbb{Z}_+ : j \leq \frac{n}{2} + \beta$ *and* $c_j(\beta) \neq 0\}$. *Then*

$$p(\beta) = \begin{cases} \beta & \text{if } \beta \in \mathbb{Z}_+, \\ \frac{n}{2} + \beta - 1 & \text{if } \beta \notin \mathbb{Z}_+ \text{ and } \frac{n}{2} + \beta \in \mathbb{N}, \\ \left[\frac{n}{2} + \beta\right] & \text{if } \beta \notin \mathbb{Z}_+ \text{ and } \frac{n}{2} + \beta \notin \mathbb{N}. \end{cases}$$

Proof. As a consequence of Lemma 2.4 we know that $c_j(\beta) = 0$ if and only if

(1) either $(j > \beta \in \mathbb{Z}_+)$ or $(j > \frac{n}{2} + \beta - 1 \in \mathbb{Z}_+)$.

Consequently, $c_j(\beta) \neq 0$ if and only if

(2) $\sim (j > \beta \in \mathbb{Z}_+)$ and $\sim (j > \frac{n}{2} + \beta - 1 \in \mathbb{Z}_+)$,

in which $\sim$ signifies logical negation. Thus we have the following formula for $p(\beta)$:

(3) $p(\beta) = \max\{j \in \mathbb{Z}_+ : (j \leq \frac{n}{2} + \beta)$ and $\sim (j > \beta \in \mathbb{Z}_+)$ and $\sim (j > \frac{n}{2} + \beta - 1 \in \mathbb{Z}_+)\}$.

If $\beta \in \mathbb{Z}_+$, then from (3) we have

$$\begin{aligned} p(\beta) &= \max\left\{j \in \mathbb{Z}_+ : \left(j \leq \tfrac{n}{2} + \beta\right), \ (j \leq \beta), \left(j \leq \tfrac{n}{2} + \beta - 1 \text{ or } \tfrac{n}{2} + \beta - 1 \notin \mathbb{Z}_+\right)\right\} \\ &= \max\{j \in \mathbb{Z}_+ : j \leq \beta\} = \beta. \end{aligned}$$

The other two cases are treated similarly. □

The point of this analysis is that we can now reformulate Theorem 3.3 in a way which is usually easier to apply.

COROLLARY 3.6. *The quantities* p_k *defined by equation* (3.4) *are given explicitly by the formula*

$$p_k = \begin{cases} (\alpha - k)/2 & \text{if } \alpha - k \in 2\mathbb{Z}_+, \\ (n + \alpha - k - 2)/2 & \text{if } \alpha - k \notin 2\mathbb{Z}_+ \text{ and } n + \alpha - k \in 2\mathbb{N}, \\ [(n + \alpha - k)/2] & \text{if } \alpha - k \notin 2\mathbb{Z}_+ \text{ and } n + \alpha - k \notin 2\mathbb{N}. \end{cases}$$

For G *to belong to* $L^1(\mathbb{R}^n)$, *it is necessary and sufficient that* $m_{2j} = 0$ *for* $0 \leq j \leq Q$, *where* Q *is given in equation* (3.5).

4. The integral of G. As discussed in §1, the two conditions required of a mapping $G : \mathbb{R}^n \to \mathbb{R}$ for it to be a suitable kernel are that $G \in L^1(\mathbb{R}^n)$ and that $\int G(x)\,dx \neq 0$. In §3, we elaborated on the first of these conditions. Now we address the second. The main results are Theorems 4.7, 4.10, and 4.12.

We recall the manner in which a kernel G has been constructed from a function $\phi \in C(0, \infty)$

$$G(x) = \omega_{n-1}^{-1} \int_{\mathbb{R}} \int_{S^{n-1}} \phi(\|x - tv\|)\, dS^{n-1}(v)\, d\mu(t). \tag{4.1}$$

Here μ is a regular, signed, Borel measure on $\mathbb{R}$ having compact support and finite total variation.

The first result indicates that there is a delicate balance between the two desiderata that $G \in L^1$ and $\int G \neq 0$.

THEOREM 4.1. *Suppose that the function $x \mapsto \phi(\|x\|)$ is in $L^1(\mathbb{R}^n)$. Then $G \in L^1(\mathbb{R}^n)$. If $\int_{\mathbf{R}} d\mu(t) = 0$, then $\int_{\mathbf{R}^n} G(x)\,dx = 0$.*

Proof. The function G is obtained by applying two averaging processes to the L^1 function $x \mapsto \phi(\|x\|)$, and consequently $G \in L^1$. By the Fubini theorem we have

$$\begin{aligned} \int_{\mathbf{R}^n} G(x)\,dx &= \omega_{n-1}^{-1} \int_{\mathbf{R}} \int_{S^{n-1}} \int_{\mathbf{R}^n} \phi(\|x - tv\|)\,dx\,dS^{n-1}(v)\,d\mu(t) \\ &= \omega_{n-1}^{-1} \int_{\mathbf{R}} \int_{S^{n-1}} \int_{\mathbf{R}^n} \phi(\|x\|)\,dx\,dS^{n-1}(v)\,d\mu(t) \\ &= \int_{\mathbf{R}^n} \phi(\|x\|)\,dx \int_{\mathbf{R}} d\mu(t). \end{aligned} \tag{4.2}$$

This calculation proves the second assertion of the theorem. □

Theorem 4.1 indicates that if the function ϕ is "too nice," i.e., $\int_{\mathbf{R}^n} |\phi(\|x\|)|\,dx < \infty$, then our procedure may lead to a function G having a zero integral.

With the aid of Lemma 2.1 we have, using $G_0(\|x\|) = G(x)$ and $g_0(\|x\|, t) = g(x,t)$,

(4.3)

$$\begin{aligned} \int_{\mathbf{R}^n} G(x)\,dx &= \omega_{n-1} \int_0^\infty r^{n-1} G_0(r)\,dr \\ &= \omega_{n-1} \int_0^\infty r^{n-1} \int_{\mathbf{R}} g_0(r,t)\,d\mu(t)\,dr \\ &= \omega_{n-2} \int_0^\infty r^{n-1} \int_{\mathbf{R}} \int_{-1}^1 \phi\big([r^2 + t^2 - 2rst]^{1/2}\big)(1-s^2)^\lambda\,ds\,d\mu(t)\,dr \\ &= \omega_{n-2} \lim_{M\to\infty} \int_{\mathbf{R}} I(M,t)\,d\mu(t), \end{aligned}$$

in which we have introduced the abbreviations

$$I(M,t) = \int_0^M \int_{-1}^1 k(r,s,t)\,ds\,dr, \tag{4.4}$$

$$k(r,s,t) = r^{n-1}\phi\big([r^2 + t^2 - 2rst]^{1/2}\big)(1-s^2)^\lambda. \tag{4.5}$$

LEMMA 4.2. *Let $0 < t \le a < M$, where a is chosen so that* supp$(\mu) \subset [-a,a]$, *and M is fixed. Let $f(r,t) = (r^2 + t^2 - M^2)/(2rt)$. Let k be as above. Then the expression*

$$J = \left[\int_0^{M+t} \int_{-1}^1 - \int_{M-t}^{M+t} \int_{-1}^{f(r,t)}\right] k(r,s,t)\,ds\,dr \tag{4.6}$$

is independent of t.

Proof. Put $e(r,t) = \max\{-1, f(r,t)\}$. Then

$$J = \int_0^{M+t} \int_{e(r,t)}^1 r^{n-1}\phi\big([r^2+t^2-2rst]^{1/2}\big)(1-s^2)^\lambda \, ds\, dr. \tag{4.7}$$

Change variables from s to v by setting $v = (r^2+t^2-2rst)^{1/2}$;

$$J = \int_0^{M+t} \int_{|r-t|}^{\min(r+t,M)} r2^{-2\lambda}t^{2-n}\phi(v)\big[v^2-(r-t)^2\big]^\lambda \big[(r+t)^2-v^2\big]^\lambda v\, dv\, dr. \tag{4.8}$$

Now reverse the order of integration;

$$J = \int_0^{M} \int_{|v-t|}^{v+t} r2^{-2\lambda}t^{2-n}\phi(v)\big[(v+t)^2-r^2\big]^\lambda \big[r^2-(v-t)^2\big]^\lambda v\, dr\, dv. \tag{4.9}$$

Finally, change variables from r to w by $r = (v^2+t^2-2vtw)^{1/2}$;

$$J = \int_0^M \phi(v)v^{n-1}\, dv \int_{-1}^1 (1-w^2)^\lambda \, dw. \tag{4.10}$$

In this argument the following identity is useful:

$$\big[v^2-(r-t)^2\big]\big[(r+t)^2-v^2\big] = \big[(v+t)^2-r^2\big]\big[r^2-(v-t)^2\big]. \qquad \Box$$

LEMMA 4.3. *Let $-a \le t \le a$ and $M > 4a$. Let $I(M,t)$ and J $(= J(M))$ be as defined above. Then $I(M,t) - J(M)$ does not depend on the values of ϕ in $[0, M/2]$.*

Proof. Note that $I(M,t)$ is an even function of t Lemma 2.2. It is therefore sufficient to consider $t > 0$. We have, as in the preceding proof,

$$\begin{aligned}
I(M,t) &= \int_0^M \int_{-1}^1 k(r,s,t)\, ds\, dr \\
&= \left(\int_0^{M-t} + \int_{M-t}^{M+t} - \int_M^{M+t}\right)\int_{-1}^1 k(r,s,t)\, ds\, dr \\
&= \left(\int_0^{M-t}\int_{-1}^1 + \int_{M-t}^{M+t}\int_{f(r,t)}^1 + \int_{M-t}^{M+t}\int_{-1}^{f(r,t)} - \int_M^{M+t}\int_{-1}^1\right) k(r,s,t)\, ds\, dr \\
&= \left(\int_0^{M+t}\int_{e(r,t)}^1 + \int_{M-t}^{M+t}\int_{-1}^{f(r,t)} - \int_M^{M+t}\int_{-1}^1\right) k(r,s,t)\, ds\, dr \\
&= J(M) + \left(\int_{M-t}^{M+t}\int_{-1}^{f(r,t)} - \int_M^{M+t}\int_{-1}^1\right) k(r,s,t)\, ds\, dr.
\end{aligned}$$

In both integrals on this last line, we have $0 \le t \le a$, $-1 \le s \le 1$, and $r \ge M - t$. Hence the argument of ϕ satisfies

$$\begin{aligned}
(r^2+t^2-2rst)^{1/2} &\ge (r^2+t^2-2rt)^{1/2} = |r-t| \ge r-t \ge M-2t \\
&\ge M-2a > M - M/2 = M/2. \qquad \Box
\end{aligned}$$

LEMMA 4.4. *Adopt all the preceding notation, and assume that $m_0 = 0$. If the function G in equation (4.1) belongs to $L^1(\mathbb{R}^n)$ then its integral does not depend on the values of ϕ in any bounded subset of $(0, \infty)$.*

Proof. Let $I(M,t)$ and $J(M)$ be as defined in equations (4.4) and (4.6). Let $\operatorname{supp}(\mu) \subset [-a, a]$. Since $\mu(\mathbb{R}) = 0$, we have, from equation (4.3),

$$\int_{\mathbb{R}^n} G(x)\,dx = \lim_{M\to\infty} \omega_{n-2} \int_{-a}^{a} I(M,t)\,d\mu(t)$$

$$= \lim_{M\to\infty} \omega_{n-2} \int_{-a}^{a} \left[I(M,t) - J(M)\right] d\mu(t).$$

By Lemma 4.3, if $M > 4a$, then $I(M,t) - J(M)$ does not depend on the values of ϕ in $(0, M/2]$. For sufficiently large M, this last interval contains any given bounded subset of $(0, \infty)$. □

LEMMA 4.5. *Let $n \geq 2$, $\alpha > -n$, $\phi(t) = t^\alpha$ for $t > 0$. Define $I(M,t)$ by equation (4.4) and let $c_j = c_j(\alpha/2)$. Then*

$$I(M,t) = |t|^{n+\alpha} \sum_{\substack{j=0 \\ 2j \neq n+\alpha}}^{\infty} c_j \left\{ \frac{1}{n+2j} - \frac{1}{n+\alpha-2j} \right\}$$

$$+ M^{n+\alpha} \sum_{\substack{j=0 \\ 2j \neq n-\alpha}}^{\infty} \frac{c_j}{n+\alpha-2j} \left(\frac{t}{M}\right)^{2j}.$$

Proof. Recall (from the proof of Lemma 4.3) that $I(M,t)$ is an even function of t. Write

$$I(M,t) = \int_0^M \int_{-1}^1 k(r,s,t)\,ds\,dr.$$

The integrand can have singularities at $r = 0$, $t = r$, and $s = \pm 1$. When $t = 0$, the integrand contains the factor $r^{n-1} r^\alpha$. This is integrable since $n - 1 + \alpha > -1$. Since $n \geq 2$, we have $\lambda \geq -1/2$; thus $(1 - s^2)^\lambda$ is integrable. To treat the singularity at $r = t$, we write

$$I(M,t) = \lim_{\varepsilon,\delta\downarrow 0} \left(\int_0^{t-\varepsilon} + \int_{t+\delta}^{M} \right) \int_{-1}^1 k(r,s,t)\,ds\,dr.$$

By using Lemma 2.4, we can compute the first summand on the right as follows:

$$\lim_{\varepsilon\downarrow 0} \int_0^{t-\varepsilon} \int_{-1}^1 r^{n-1} t^\alpha \left[\left(\frac{r}{t}\right)^2 + 1 - 2s\left(\frac{r}{t}\right) \right]^{\alpha/2} (1 - s^2)^\lambda\,ds\,dr$$

$$= \lim_{\varepsilon\downarrow 0} \int_0^{t-\varepsilon} h\left(\frac{\alpha}{2}, \lambda, \frac{r}{t}\right) r^{n-1} t^\alpha\,dr$$

$$= \lim_{\varepsilon\downarrow 0} t^\alpha \int_0^{t-\varepsilon} r^{n-1} \sum_{j=0}^{\infty} c_j \left(\frac{r}{t}\right)^{2j} dr$$

$$= \lim_{\varepsilon\downarrow 0} t^{n+\alpha} \sum_{j=0}^{\infty} \frac{c_j}{n+2j} \left(\frac{t-\varepsilon}{t}\right)^{n+2j}$$

$$= t^{n+\alpha} \lim_{x\uparrow 1} \sum_{j=0}^{\infty} \frac{c_j}{n+2j}\, x^{n+2j}.$$

By Lemma 2.7, the series $\sum c_j/(n+2j)$ converges. By Abel's theorem [W, p. 320], the evaluation of the limit above yields

$$t^{n+\alpha}\sum_{j=0}^{\infty}\frac{c_j}{n+2j}.$$

A similar analysis, appealing to Lemma 2.8 and Abel's theorem, gives

$$\lim_{\delta\downarrow 0}\int_{t+\delta}^{M}\int_{-1}^{1}k(r,s,t)\,ds\,dr=\sum_{\substack{j=0\\ 2j\neq n+\alpha}}^{\infty}\frac{c_j}{n+\alpha-2j}\left[M^{n+\alpha}\Big(\frac{t}{M}\Big)^{2j}-t^{n+\alpha}\right].$$

Notice that by Lemma 2.4, $c_j=0$ when $2j=n+\alpha$. □

LEMMA 4.6. *Let f and k be as in Lemma 4.2 and equation (4.5). If $t>0$ and $\phi(t)=t^{-n}$, then*

$$\lim_{M\to\infty}\int_{M}^{M+t}\int_{-1}^{1}k(r,s,t)\,ds\,dr=0,$$

$$\lim_{M\to\infty}\int_{M-t}^{M+t}\int_{-1}^{f(r,t)}k(r,s,t)\,ds\,dr=0.$$

Proof. We assume that $M>3t$. Then it is easily verified that $-1\le f(r,t)\le 1$. Also, in the integrals above, we have $s\le 1$, $r\ge M-t\ge 2t>0$. Hence

$$1+(t/r)^2-2s(t/r)\ge 1+(t/r)^2-2(t/r)=(1-t/r)^2\ge 1/4.$$

If we write

$$\begin{aligned}k(r,s,t)&=r^{n-1}(r^2+t^2-2srt)^{-n/2}(1-s^2)^{\lambda}\\&=r^{-1}\big[1+(t/r)^2-2s(t/r)\big]^{-n/2}(1-s^2)^{\lambda},\end{aligned}$$

then we see that

$$0\le k(r,s,t)\le r^{-1}(1/4)^{-n/2}(1-s^2)^{\lambda}.$$

Hence, an upper bound for both integrals is

$$\int_{M-t}^{M+t}\int_{-1}^{1}k(r,s,t)\,ds\,dr\le 2^n\int_{M-t}^{M+t}r^{-1}\,dr\int_{-1}^{1}(1-s^2)^{\lambda}\,ds\to 0.\qquad\square$$

THEOREM 4.7. *Let ϕ be an element of $C(0,\infty)$ that can be represented as*

$$\phi(r)=\sum_{k=0}^{[n+\alpha]}a_k r^{\alpha-k}+\phi_1(r),$$

where $a_0\neq 0$, $\alpha\in\mathbb{R}\backslash\mathbb{Z}$, and $\phi_1(\|\cdot\|)\in L^1(\mathbb{R}^n)$. Assume also that $G\in L^1(\mathbb{R}^n)$ and $Q\ge 0$ (definition in equation (3.5)). Then $\int G=0$.

Proof. Since $G\in L^1$, Theorem 3.3 implies that $m_{2j}=0$ for $0\le j\le s$. By Lemma 3.4, $s=Q$. Hence $m_0=0$. By the definition of Q in equation (3.5), $n+\alpha\ge 0$. Since $\alpha\notin\mathbb{Z}$, $n+\alpha>0$. Let $\theta=\phi-\phi_1$, and set $G=F+H$, where

$$H(x)=\omega_{n-1}^{-1}\int_{-a}^{a}\int_{S^{n-1}}\theta(\|x-tv\|)\,dS^{n-1}(v)\,d\mu(t),$$

$$F(x)=\omega_{n-1}^{-1}\int_{-a}^{a}\int_{S^{n-1}}\phi_1(\|x-tv\|)\,dS^{n-1}(v)\,d\mu(t).$$

Since $m_0 = 0$, it follows from Theorem 4.1 that $\int F = 0$. Now consider $\int H$. Recall that $\operatorname{supp}(\mu) \subset [-a, a]$. Choose L so that $L > 3a$. By Lemma 2.1, equation (4.4), and Lemma 4.5, we have, for $M > L$,

$$\begin{aligned}
&\frac{\omega_{n-1}}{\omega_{n-2}} \int_{\|x\| \le M} H(x)\, dx \\
&= \int_{-a}^{a} \int_{0}^{M} r^{n-1} \int_{-1}^{1} \theta\big([r^2 + t^2 - 2rst]^{1/2}\big)(1 - s^2)^{\lambda}\, ds\, dr\, d\mu(t) \\
&= \sum_{k=0}^{[n+\alpha]} a_k \int_{-a}^{a} \int_{0}^{M} r^{n-1} \int_{-1}^{1} (r^2 + t^2 - 2rst)^{(\alpha-k)/2}(1 - s^2)^{\lambda}\, ds\, dr\, d\mu(t) \\
&= \sum_{k=0}^{[n+\alpha]} a_k \biggl\{ \int_{-a}^{a} |t|^{n+\alpha-k}\, d\mu(t) \sum_{j=0}^{\infty} c_j\Big(\frac{\alpha-k}{2}\Big)\Big[\frac{1}{n+2j} - \frac{1}{n+\alpha-k-2j}\Big] \\
&\qquad + M^{n+\alpha-k} \sum_{j=0}^{\infty} \frac{c_j\big(\frac{\alpha-k}{2}\big)}{n+\alpha-k-2j} \int_{-a}^{a} \Big(\frac{t}{M}\Big)^{2j} d\mu(t) \biggr\}.
\end{aligned}$$

By Lemma 2.10 we have

$$\sum_{j=0}^{\infty} c_j\Big(\frac{\alpha-k}{2}\Big)\Big[\frac{1}{n+2j} - \frac{1}{n+\alpha-k-2j}\Big] = 0 \qquad (0 \le k \le n+\alpha).$$

Since $H \in L^1(\mathbb{R}^n)$ the expression

$$\sum_{k=0}^{[n+\alpha]} a_k M^{n+\alpha-k} \sum_{j=0}^{\infty} \frac{c_j\big(\frac{\alpha-k}{2}\big)}{n+\alpha-k-2j}\, M^{-2j} \int_{-a}^{a} t^{2j}\, d\mu(t)$$

cannot contain any positive power of M. Hence the highest term present has a negative exponent since α is not an integer. Thus the expression under consideration converges to 0 as M tends to ∞. It follows that

$$\int_{\mathbb{R}^n} H = \lim_{M \to \infty} \int_{\|x\| \le M} H = 0. \qquad \Box$$

LEMMA 4.8. *Adopt the hypotheses of Theorem 4.7, except that now $\alpha \in \mathbb{Z}$. Define $m_\beta = \int |t|^\beta\, d\mu(t)$ and*

$$R(n, \beta) = \sum_{\substack{j=0 \\ 2j \ne n+\beta}}^{\infty} c_j\Big(\frac{\beta}{2}\Big)\Big(\frac{1}{n+2j} - \frac{1}{n+\beta-2j}\Big).$$

Then there is an index ν such that

$$\int G(x)\, dx = \omega_{n-2} a_\nu m_{n+\alpha-\nu} R(n, \alpha - \nu).$$

Proof. Proceed as in the preceding proof, and define

$$\theta_1(r) = \theta(r) - a_{n+\alpha} r^{-n} \qquad (r > 0).$$

Then we have $G = H + F + K$, where

$$H(x) = \omega_{n-1}^{-1} \int_{-a}^{a} \int_{S^{n-1}} \theta_1(\|x - tu\|)\, dS^{n-1}(u)\, d\mu(t),$$

$$F(x) = \omega_{n-1}^{-1} \int_{-a}^{a} \int_{S^{n-1}} \phi_1(\|x - tu\|)\, dS^{n-1}(u)\, d\mu(t),$$

$$K(x) = \omega_{n-1}^{-1} \int_{-a}^{a} \int_{S^{n-1}} a_{n+\alpha} \|x - tu\|^{-n}\, dS^{n-1}(u)\, d\mu(t).$$

Then, by Theorem 4.1, $\int F = 0$, and by Lemma 4.6, $\int K = 0$. Note that when $\alpha = -n$, $H = 0$, and so we shall henceforward consider only the case $\alpha > -n$. Now,

$$\begin{aligned}
I(M) &= \int_{\|x\| \le M} H(x)\, dx \\
&= \omega_{n-2} \int_0^M \int_{-a}^{a} \int_{-1}^{1} \theta_1\big([r^2 + s^2 - 2rst]^{1/2}\big)(1 - s^2)^{\lambda}\, ds\, d\mu(t)\, dr \\
&= \omega_{n-2} \int_0^M \int_{-a}^{a} \sum_{k=0}^{n+\alpha-1} a_k \int_{-1}^{1} (r^2 + t^2 - 2rst)^{\frac{\alpha-k}{2}} (1 - s^2)^{\lambda}\, ds\, d\mu(t)\, dr \\
&= \omega_{n-2} \int_{-a}^{a} \sum_{k=0}^{n+\alpha-1} a_k |t|^{n+\alpha-k} \sum_{\substack{j=0 \\ 2j \ne n+\alpha-k}}^{\infty} c_j\Big(\frac{\alpha-k}{2}\Big)\Big[\frac{1}{n+2j} - \frac{1}{n+\alpha-k-2j}\Big]\, d\mu(t) \\
&\quad + \omega_{n-2} \int_{-a}^{a} \sum_{k=0}^{n+\alpha-1} a_k M^{n+\alpha-k} \sum_{\substack{j=0 \\ 2j \ne n+\alpha-k}}^{\infty} \frac{c_j(\frac{\alpha-k}{2})}{n+\alpha-k-2j}\Big(\frac{t}{M}\Big)^{2j}\, d\mu(t) \\
&= \omega_{n-2} \sum_{k=0}^{n+\alpha-1} a_k m_{n+\alpha-k} R(n, \alpha - k) \\
&\quad + \omega_{n-2} \sum_{k=0}^{n+\alpha-1} a_k M^{n+\alpha-k} \sum_{\substack{j=0 \\ 2j \ne n+\alpha-k}}^{\infty} \frac{c_j(\frac{\alpha-k}{2})}{n+\alpha-k-2j} M^{-2j} m_{2j}.
\end{aligned}$$

The penultimate step above utilizes Lemma 4.5.

Since $G \in L^1$ and

$$\int G = \int H = \lim_{M \to \infty} I(M),$$

we see that in the final equation for $I(M)$, any term $M^{n+\alpha-k-2j}$ with positive exponent must have zero coefficient. The term with 0 exponent is missing because

$2j \neq n+\alpha-k$. Hence only negative exponents occur, and all terms involving M are either zero or converge to 0. Thus

$$\int G = \omega_{n-1} \sum_{k=0}^{n+\alpha-1} a_k m_{n+\alpha-k} R(n, \alpha-k).$$

It remains to be established that this sum contains at most one nonzero summand. Define

$$\nu = \min\{k : a_k R(n, \alpha-k) \neq 0\}.$$

If there is no index k such that $a_k R(n,\alpha-k) \neq 0$, then $\int G = 0$ and the proof ends by setting $\nu = 0$. We assert that only the summand corresponding to $k=\nu$ can be nonzero. Since $R(n,\alpha-\nu) \neq 0$, either $\alpha-\nu$ is odd or $\alpha-\nu$ is even and negative, by Lemma 2.10. In any case, $(\alpha-\nu)/2 \notin \mathbb{Z}_+$. Hence, by Corollary 3.6, $p_\nu \geq (n+\alpha-\nu-2)/2$. By equation (3.5), $Q \geq (n+\alpha-\nu-2)/2$. By Theorem 3.3 and Lemma 3.4, we have $m_{2j} = 0$ for $0 \leq 2j \leq n+\alpha-\nu-2$. Now let $0 \leq k < \nu$. The corresponding summand is 0 by the definition of ν. Next, let $\nu < k \leq n+\alpha-1$. Then $1 \leq n+\alpha-k < n+\alpha-\nu$. If $n+\alpha-k$ is even then $m_{n+\alpha-k} = 0$. If $n+\alpha-k$ is odd, then n and $\alpha-k$ are of opposite parity, and $R(n,\alpha-k) = 0$ by Lemma 2.9. Thus in either case, the term $a_k R(n,\alpha-k) m_{n+\alpha-k}$ is 0. □

We now apply the results of §§3 and 4 to a number of examples. When we do not make any explicit statement about ϕ, but rather refer to the parameters n and α, ϕ is assumed to have the form given in Theorem 4.7. An important feature of Theorem 4.7 is that we only make assumptions about the behaviour of ϕ "at infinity," and so many of our results need no information about the behaviour of ϕ in any bounded neighbourhood of zero. This is a substantial generalisation of the work of Jackson [Ji]. We begin with an easy case in which we can be completely prescriptive.

Example 4.9. Let $\phi \in C[0,\infty)$, $\phi \geq 0$, and $\phi(t) = t^\alpha$ for large t, where $\alpha < -n$. Then $G \in L^1(\mathbb{R}^n)$ by Lemma 3.1. To have $\int G(x)\,dx \neq 0$ it is necessary and sufficient that $m_0 \neq 0$. (See the proof of Theorem 4.1 for this fact.)

The next two results encompass some of the examples mentioned by Jackson [Ji].

THEOREM 4.10. *Let n and α be odd, $\alpha+n > 0$. In order that $G \in L^1(\mathbb{R}^n)$ and $\int G(x)\,dx \neq 0$ it is necessary and sufficient that $m_{2j} = 0$ for $0 \leq 2j \leq n+\alpha-2$ and $m_{n+\alpha} \neq 0$.*

Proof. By Lemma 2.11, $R(n,\alpha) \neq 0$. By hypothesis, $a_0 \neq 0$. Hence by the definition of ν (in the proof of Lemma 4.8), $\nu = 0$. By Lemma 4.8,

$$\int G(x)\,dx = \frac{\omega_{n-2}}{\omega_{n-1}} a_0 m_{n+\alpha} R(n,\alpha).$$

Since $\alpha \notin 2\mathbb{Z}_+$, and $\alpha+n \in 2\mathbb{N}$, we see by Corollary 3.6 that $p_0 = (n+\alpha-2)/2$. Also by Corollary 3.6 we have $p_k \leq p_0$. Hence (by equation (3.5)), $Q = p_0$. In order that $G \in L^1$, it is necessary and sufficient that $m_{2j} = 0$ for $0 \leq 2j \leq 2Q$. Hence the condition $m_{n+\alpha} \neq 0$ is equivalent to $\int G(x)\,dx \neq 0$. □

Example 4.11. For the multiquadrics mentioned in the introduction, we let $\phi(t) = (a^2+t^2)^{1/2}$. For this function the asymptotic series has $\alpha = 1$. Hence in odd dimensions we can obtain a suitable kernel and give a constructive proof of the fundamentality of the set $\{x \mapsto (a^2 + \|\lambda x + y\|^2)^{1/2} : \lambda \in \mathbb{R} \text{ and } y \in \mathbb{R}^n\}$. Similar remarks pertain to $(a^2+t^2)^{k/2}$, k odd, $k > -n$.

We conclude with a result and an application of this result which leads to kernels of a form not previously considered by other authors.

THEOREM 4.12. . *Let n and α be even, $0 > \alpha > -n$. In order that $G \in L^1(\mathbb{R}^n)$ and $\int G \neq 0$ it is necessary and sufficient that $m_{2j} = 0$ for $0 \leq 2j \leq n + \alpha - 2$ and $m_{n+\alpha} \neq 0$.*

Proof. Proceed as in the proof of Theorem 4.10. Since $n + \alpha \in 2\mathbb{N}$ and $\alpha \notin 2\mathbb{Z}_+$, we have $R(n, \alpha) \neq 0$ by Lemma 2.11. Again we find that $\int G \neq 0$ if and only if $m_{n+\alpha} \neq 0$. In order that $G \in L^1$ we must have $m_{2j} = 0$ for $0 \leq 2j \leq 2Q$. Obviously we require $Q < n + \alpha$, so that we can stipulate $m_{n+\alpha} \neq 0$ as a condition on μ. By Corollary 3.6, $p_0 = (n + \alpha - 2)/2$. Also $p_1 = [(n + \alpha - 1)/2] = (n + \alpha - 2)/2$, and in general $p_k \leq p_0$. Hence $Q = p_0 < (n + \alpha)/2$. □

Example 4.13. In dimensions $n = 4, 6, 8, \ldots$ we can let $\phi(t) = (a^2 + t^2)^{-1}$. For this example, $\alpha = -2$. According to Theorem 4.12, a suitable kernel can be formed, and we can prove constructively the fundamentality of the set of functions $x \mapsto (a^2 + \|\lambda x - y\|^2)^{-1}$.

REFERENCES

[AS] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, DC, 1964.

[Bu] M. D. BUHMANN, *Multivariate interpolation in odd-dimensional Euclidean spaces using multi-quadrics*, Constr. Approx. 6 (1990), pp. 21–34.

[F] R. FRANKE, *Scattered data interpolation: Tests of some methods*, Math. Comp. 38 (1982), pp. 381–200.

[GS] I. M. GELFAND AND G. E. SHILOV, *Generalized Functions*, vol. I, Academic Press, New York, 1964.

[Ha] R. L. HARDY, *Theory and applications of the multiquadric biharmonic method*, Comput. Math. Appl., 19 (1990), pp. 163–208. (This paper has a bibliography of 109 items.)

[Ji] I. R. H. JACKSON, *Convergence properties of radial basis functions*, Constr. Approx. 4 (1988), pp. 243–264.

[Jo] FRITZ JOHN, *Plane Waves and Spherical Means Applied to Partial Differential Equations*, Interscience, New York, 1955.

[M] C. A. MICCHELLI, *Interpolation of scattered data: Distance matrices and conditionally positive definite functions*, Constr. Approx. 2 (1986), pp. 11–22.

[MN] W. R. MADYCH AND S. A. NELSON, *Multivariate interpolation: A variational approach*, preprint, 1983.

[S1] I. J. SCHOENBERG, *Metric spaces and completely monotone functions*, Ann. of Math. 39 (1938), pp. 811–841.

[S2] ———, *On certain metric spaces arising from Euclidean spaces...*, Ann. of Math. 36 (1937), pp. 787–793.

[S3] ———, *Metric spaces and positive definite functions*, Trans. Amer. Math. Soc., 44 (1938), pp. 811–841.

[S4] ———, *Remarks to Maurice Frechet's article...*, Ann. of Math., 36 (1935), pp. 724–732.

[SW] E. M. STEIN AND G. WEISS, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.

[W] D. V. WIDDER, *Advanced Calculus*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ, 1961.

[XLC] Y. XU, W. A. LIGHT, AND E. W. CHENEY, *Constructive methods of approximation by ridge functions and radial functions*, Numer. Algorithms, 4 (1993), pp. 205–223.

ORTHOGONAL WAVELETS ON THE CANTOR DYADIC GROUP*

W. CHRISTOPHER LANG†

Abstract. Based upon the shift operator as a dilation operator, multiresolution analyses are built on the Cantor dyadic group. A regularity condition is given for wavelets and sufficient conditions are given on scaling filters for regular orthonormal wavelets to occur. Examples of wavelets given include the Haar functions and certain lacunary Walsh function series analogous to the compactly supported wavelets of I. Daubechies.

Key words. orthogonal wavelets, locally compact Abelian groups, Cantor dyadic group

AMS subject classifications. 42C05, 43A70

1. Introduction.

We wish to construct orthogonal wavelets on locally compact Abelian (LCA) groups, resembling the familiar constructions of Y. Meyer [9] or S. Mallat [8] on n-dimensional Euclidean space. We are interested in general groups, not necessarily with Lie structure. Very general constructions have already been given of wavelets on stratified Lie groups based on spline approximations [7] and of wavelet bases for Hilbert spaces with unitary operators representing dilation and translation [1]. In [8] and [9], orthogonal wavelets are built in terms of scaling filters. Conditions are given there on scaling filters for multiresolution analyses (MRAs) to be generated and for the orthogonality of wavelets. We shall define translation, dilation, and MRAs in the setting of LCA groups. On the locally compact Cantor dyadic group, we shall give a definition of the regularity of a wavelet and give sufficient conditions on scaling filters for regular MRAs and orthogonal wavelets to occur. Our main result, Theorem 2.2, resembles [8, Thm. 2].

In §2, we describe the locally compact Cantor dyadic group. Our wavelets on this group will include wavelets corresponding to the Haar wavelets on the line; 2^nth partial sums of Walsh series are nth resolutions (approximations or resolutions at scale 2^{-n}) in these wavelets. Another example of wavelets on the dyadic group consists of certain lacunary Walsh series which are analogous to the continuous compactly supported wavelets of I. Daubechies [2] in that they are generated by scaling filters consisting of (Walsh-) trigonometric polynomials. See §3 below.

We now give our axioms for dilation and translation. These are similar to those of [1], specialized to the case of the Hilbert space $L^2(G)$. In the case of $G = R$, the dilation is the usual $x \to 2x$, and the translation is the usual translation by integer units. There is nothing special about requiring the dilation to be dyadic, but our construction will be general enough to accomodate wavelets on the locally compact Cantor dyadic group as well as wavelets on the plane R^2.

We let G be an LCA group with automorphisms ρ and σ (inverse to each other), Haar measure m, and a subgroup Λ. We assume
(1.1.1) Λ is countably infinite and closed,
(1.1.2) G/Λ is compact,
(1.1.3) $\rho(\Lambda) \subseteq \Lambda$,

* Received by the editors April 26, 1993; accepted for publication (in revised form) May 11, 1994.

† Department of Mathematics, Indiana University Southeast, New Albany, IN 47150 (clang@iusmail.ius.indiana.edu).

(1.1.4) $m(\rho(E)) = 2m(E)$ for all Borel $E \subseteq G$ (so $m(\sigma(E)) = \frac{1}{2}m(E)$),
(1.1.5) $\Lambda/\rho(\Lambda) \approx Z/(2)$.

We will call Λ a *lattice subgroup* of G, and ρ the *dilation operator*. (A simple example is provided by $G = R$, $\Lambda = Z$, and $\rho(x) = 2x$.) We also define the notation for translation $\tau_y f(x) = f(x+y)$ for functions f on G.

We write ρ^j for the jth composition of ρ: if $j < 0$, we set $\rho^j = \sigma^j$. Then let $\Lambda_j = \rho^j(\Lambda)$. It follows from (1.1.5) that $\Lambda_j/\Lambda_{j+1} \approx Z/(2)$ for all j.

In this setting, we define a *multiresolution analysis* of $L^2(G)$ to be a sequence $(V_j)_{j\in Z}$ of closed subspaces of $L^2(G)$ such that
(1.2.1) $V_j \subseteq V_{j+1}$ for all j in Z,
(1.2.2) $\cup V_j$ is dense in $L^2(G)$ and $\cap V_j = \{0\}$,
(1.2.3) $f \in V_j$ if and only if $\rho f = f \circ \rho \in V_{j+1}$ for all $j \in Z$,
(1.2.4) $f \in V_0$ if and only if $\tau_n f \in V_0$ for all n in Λ; $f \in V_0$ if and only if $\rho^j \tau_n f \in V_j$,
(1.2.5) there is a function $g \in L^2(G)$ such that $\{\tau_n g : n \in \Lambda\}$ is a Riesz basis of V_0.

We say that a function f on G is Λ-*periodic* if $\tau_n f = f$ for all $n \in \Lambda$.

See [6] for basic information about sampling of functions and approximations on LCA groups.

We now consider resolution (or scale) filters. For $\phi \in L^2(G)$, we may define V_j to be the L^2-closure of the linear span of $\{\phi(\rho^j(x) + n) : n \in \Lambda\}$, $j \in Z$. We require that $V_j \subseteq V_{j+1}$ for all $j \in Z$. In particular, we require $V_0 \subseteq V_1$. This entails

$$\phi(x) = \sum_{n\in\Lambda} c_n \phi(\rho(x) + n) \tag{1.3}$$

for some choice of coefficients $\{c_n\}$. Under the Fourier transform, we have

$$\hat{\phi}(\omega) = m_0(\omega)\hat{\phi}(\sigma(\omega)). \tag{1.4}$$

We seek conditions on our "resolution filter" $m_0(\omega)$ such that $\{\tau_n\phi : n \in \Lambda\}$ is an orthonormal set and such that

$$\hat{\phi}(\omega) = m_0(\omega)m_0(\sigma(\omega))m_0(\sigma^2(\omega))\cdots$$

converges to an $L^2(\hat{G})$-function $\hat{\phi}$ such that ϕ (the inverse Fourier transform) produces an MRA. Thus, we seek an analogue of Theorem 2 of [8]. Theorem 2.2 below is an analogue of that theorem for the Cantor dyadic group.

2. The Cantor dyadic group. The Cantor dyadic group is $D = \prod_{n=-\infty}^{-1} Z/(2)$ under the cartesian product topology, where $Z/(2)$ is the integers modulo 2, that is, $\{0,1\}$ with addition modulo 2. We will write an element of D as $(x_j)_{j<0}$, where $x_j \in \{0,1\}$ for each j. We make the index j range over the negative integers, so that we may think of $x = (x_j)_{j<0}$ as $x = 0.x_{-1}x_{-2}x_{-3}\cdots$, a binary fraction expansion. We may thus identify D with the unit interval as a measure space by the map $x \to |x| : D \to [0,1]$, where $|x| = \sum_{j<0} x_j 2^j$. This induces the Haar measure on D.

We will actually work on the locally compact version of the Cantor dyadic group, namely

$$\begin{aligned} G &= \prod_{n=-\infty}^{\infty}{}^{*} Z/(2) \\ &= \{(x_j)_{j\in Z} : x_j \in \{0,1\} \text{ for all } j \text{ and } x_j = 0 \text{ for all } j > n, \text{ for some } n \in Z\}. \end{aligned}$$

(See Chapter 6 of [5].) So we may think of $x = x_n x_{n-1}\cdots x_1 x_0.x_{-1}x_{-2}\cdots$, and we identify G with $[0,\infty)$ as a measure space by $x \to |x| : G \to [0,\infty)$, where $|x| = \sum_{j\in Z} x_j 2^j$. Again, this induces the Haar measure on G.

Note that D is a subgroup of G, so $D = \{x \in G : x_j = 0 \text{ for } j \geq 0\}$.

Let $\Lambda \subseteq G$ be the lattice subgroup $\Lambda = \{x \in G : x_j = 0 \text{ for } j < 0\}$. Thus Λ is countable and closed and $G/\Lambda = D$ is compact. Under the map $x \to |x|$, Λ is identified with the nonnegative integers in the half-line.

We now define the dilation $\rho : G \to G$ by $\rho(x)_j = x_{j-1}$ for $x \in G$. We also define $\sigma : G \to G$ to be $\sigma(x)_j = x_{j+1}$ for $x \in G$. So ρ and σ are topological automorphisms inverse to each other. Let $\Lambda_j = \rho^j(\Lambda)$ for $j > 0$, let $\Lambda_0 = \Lambda$, and let $\Lambda_j = \sigma^j(\Lambda)$ for $j < 0$. It is easy to see that Λ_j/Λ_{j+1} is isomorphic to $Z/(2)$ for all j in Z. Also, if $E \subseteq G$ is Borel, then $m(\rho(E)) = 2m(E)$, where m is the Haar measure.

The Pontryagin dual group $\hat{G}$ of G is topologically isomorphic to G. We may write

$$\hat{G} = \{\omega = (\omega_j)_{j\in Z} : \omega_j \in \{0,1\} \text{ for all } j \text{ and } \omega_j = 0 \text{ for all } j > n \text{ for some } n\}$$

and define $\omega(x) = \prod_{j\in Z}(-1)^{\omega_{-1-j}x_j}$ for $x = (x_j)_j$ in G. This gives $\omega \in G$ as a group character on G. Note the minus sign in the exponent in the subscript of ω, so that the product is actually finite.

We may consider dilation on $\hat{G}$ just as on G; let $\sigma(\omega)_j = \omega_{j+1}$ and $\rho(\omega)_j = \omega_{j-1}$ for $\omega = (\omega_j)_j$ in $\hat{G}$. We find that the character $\rho(\omega)$ evaluated at $x = (x_j)_j$ is equal to ω evaluated at $\rho(x)$, i.e., $\omega \circ \rho = \rho(\omega)$. This is analogous to the situation for characters on the real line; $e^{i(2x)y} = e^{ix(2y)}$. We also find that the integers Λ in $\hat{G}$ (defined just as they are on G) correspond exactly to the set Γ of Λ-periodic characters on G. The Λ-periodic characters on G, where G is identified with the real half-line under $x \to |x|$, turn out to be exactly the Walsh functions.

We will use the notation $W_\omega(x) = \omega(x)$ for $\omega \in \hat{G}$ and $x \in G$. If $\omega \in \Gamma$, then $|\omega|$ is an (ordinary) integer and W_ω is identified with an ordinary Walsh function on the real line. Here we will use the Paley enumeration; if we let $r_k(x) = \text{sign}(\sin(2\pi 2^k x))$ be the kth Rademacher function, then the Walsh function W_n is the product of the functions r_k such that the kth bit in the binary expansion of n is a 1. (Note that the least significant bit of the binary expansion of an integer occupies position zero. Thus W_1 is the function constantly equal to 1 on the interval $0 \leq x < 1/2$ and constantly equal to -1 on $1/2 < x < 1$.)

We will also use the notation $W_x(\omega) = \omega(x)$ for $x \in G$ and $\omega \in \hat{G}$ to write characters on $\hat{G}$; by Pontryagin duality, all characters on $\hat{G}$ arise in this way.

It will be convenient to write for both G and $\hat{G}$ specific elements as binary expansions. Thus, for example, by $x = 11.1$ in G, we mean the element $x = (x_j)_j$ such that $x_j = 1$ for $j = 1, 0, -1$ and $x_j = 0$ otherwise.

See [3], [10], or [11] for development of the harmonic analysis of these groups. In [11], the locally compact group G is an example of a local field, the 2-series field; in [10, Chap. 9], G is treated under the name "dyadic field."

For the Cantor dyadic group, we give a regularity condition similar to that of [8]. (Also compare Definition 2 of §II.2 in [9]; we of course do not give a definition involving bounds on the derivative since the Cantor dyadic group has no derivative per se.)

DEFINITION 2.1. *$g \in L^2(G)$ is said to be* regular *if $|g(x)| \leq c(1+|x|)^{-2}$ for all $x \in G$, for some constant $c > 0$. (Recall $|x| = \sum_{j\in Z} x_j 2^j$ for $x \in G$.)*

We also define $\Delta_k = \{\omega \in \hat{G} : \omega_j = 0 \text{ for } j \geq k\}$. (This is analogous to the interval $-2^k\pi \leq \omega \leq 2^k\pi$ in Lemma 2 of [8].)

We may now state our analogue of Theorem 2 of [8].

THEOREM 2.2. *Let $m_0(\omega) = \sum_{n\in\Lambda} h_n W_n(\omega)$ be such that*
(2.2.1) $|h_n| = O(1+|n|^{2+\epsilon})^{-1}$ *for some* $\epsilon > 0$,
(2.2.2) $m_0(0) = 1$,
(2.2.3) $M(\omega) + M(\omega + 0.1) \equiv 1$, *where* $M(\omega) = |m_0(\omega)|^2$,
(2.2.4) $m_0(\omega) \neq 0$ *on* Δ_{-1}.

Define $\hat{\phi}(\omega) = \prod_{k=1}^{\infty} m_0(\sigma^k(\omega))$. Then the function $\hat{\phi}(\omega)$ is an $L^2(\hat{G})$-function. The inverse Fourier transform $\phi(x)$ of $\hat{\phi}(\omega)$ has the property that $\{\tau_n\phi : n \in \Lambda\}$ is an orthonormal basis for a closed subspace V_0 of $L^2(G)$. If ϕ is regular, then a regular MRA $(V_j)_{j\in Z}$ is produced.

The proof of this proceeds essentially as in [8], suitably translated, via analogues of Lemma 1 and Lemma 2 of [8].

3. Wavelets on LCA groups. We return now to the general situation of §1. Suppose we have an MRA $(V_j)_{j\in Z}$ of $L^2(G)$. We shall show how orthogonal wavelets may be constructed from this MRA. (This will be very similar to [8] or [9, Chap. III, §2]. We will also present examples of MRAs and orthogonal wavelets on various groups.

We know that $V_j \subseteq V_{j+1}$ for all j. So, let W_j be the orthogonal complement of V_j in V_{j+1}, so $V_j \oplus W_j = V_{j+1}$. Let ϕ be the "father wavelet," whose translates span V_0. We seek ψ, whose translates span W_0 and are orthogonal, such that the translates of ψ are orthogonal to the translates of ϕ. Now ϕ and ψ belong to V_1, so

$$\phi(x) = \sum_{n\in\Lambda} a_n \phi(\rho(x)+n) \quad \text{and} \quad \psi(x) = \sum_{n\in\Lambda} b_n \phi(\rho(x)+n).$$

Upon taking Fourier transforms, we find

$$\hat{\phi}(\omega) = m_0(\omega)\hat{\phi}(\sigma(\omega)) \quad \text{and} \quad \hat{\psi}(\omega) = m_1(\omega)\hat{\phi}(\sigma(\omega)),$$

where

$$m_0(\omega) = \sum_{n\in\Lambda} \tfrac{1}{2} a_n \omega(\sigma(n)) \quad \text{and} \quad m_1(\omega) = \sum_{n\in\Lambda} \tfrac{1}{2} b_n \omega(\sigma(n)).$$

We may show that ψ and ϕ obey the desired orthogonality conditions when

$$\begin{bmatrix} m_0(\omega) & m_1(\omega) \\ m_0(\omega+\theta) & m_1(\omega+\theta) \end{bmatrix}$$

is a unitary matrix, where θ is an odd Λ-periodic character, i.e., θ is not Λ_{-1}-periodic.

To make this matrix unitary, it is enough to let $m_1(\omega) = W(\omega)\overline{m_0(\omega+\theta)}$, where $W(\omega) = \omega(k)$ for some $k \in \Lambda_{-1} \setminus \Lambda_0$ (k is a "half integer").

Taking inverse Fourier transforms of

$$\hat{\psi}(\omega) = W(\omega)\overline{m_0(\omega+\theta)\hat{\phi}(\sigma(\omega))},$$

we obtain

$$\psi(x) = \sum_{n\in\Lambda} (-1)^{\operatorname{sgn}(n)} \overline{a_{\overline{n}}} \phi(\rho(x) - n - \rho(k)), \tag{3.1}$$

where $\operatorname{sgn}(n) = 1$ if n is even (i.e., $n \in \Lambda_1$) or -1 if n is odd ($n \in \Lambda \setminus \Lambda_1$).

It now becomes clear (as in the usual case) that

$$\{2^{j/2}\psi(\rho^j(x) - n) : n \in \Lambda, j \in Z\}$$

is an orthogonal system of wavelets, for the MRA $(V_j)_{j\in Z}$.

Example 1. If we let $G = R$, $\Lambda = Z$, and $\rho(x) = 2x$, we find that (3.1) becomes equation (3.8) [9, Chap. III, §2].

Example 2. For G the locally compact Cantor group, (3.1) may be written

$$\psi(x) = \sum_{n\in\Lambda} \overline{a_n}(-1)^{|n|}\phi(\rho(x) + n + 1.0). \tag{3.2}$$

If we set $\phi(x) = 1_I(x)$, where I is the "unit" interval in G given by $I = \{x \in G : x_j = 0 \text{ for all } j \geq 0\}$, we find that $\phi(\sigma(x)) = \phi(x) + \phi(x + 1.0)$. Then our wavelet is

$$\psi(x) = \phi(\rho(x) + 1.0) - \phi(\rho(x)),$$

which is, of course, the Haar wavelet (when G is identified with the half-line under the mapping $x \to |x|$). Note that the multiresolution analysis here can be described without reference to the dual group as

$$V_j = \{f \in L^2(G) : f \text{ is constant on cosets of } D_j\ \},$$

where D_j is the subgroup of G consisting of all $x \in G$ such that $x_i = 0$ for $i \geq j$.

Example 3. We construct a different example of wavelets on the Cantor dyadic group by defining the scaling filter

$$m_0(\omega) = \begin{cases} 1 & \text{if } \omega_{-1} = \omega_{-2} = 0, \\ a & \text{if } \omega_{-1} = 0 \text{ and } \omega_{-2} = 1, \\ 0 & \text{if } \omega_{-1} = 1 \text{ and } \omega_{-2} = 0, \\ b & \text{if } \omega_{-1} = \omega_{-2} = 1, \end{cases}$$

where $|a|^2 + |b|^2 = 1$.

This filter is actually the Walsh-trigonometric polynomial

$$m_0 = c_0W_{00.0} + c_1W_{01.0} + c_2W_{10.0} + c_3W_{11.0},$$

where $c_0 = (1+a+b)/4$, $c_1 = (1+a-b)/4$, $c_2 = (1-a-b)/4$, and $c_3 = (1-a+b)/4$.

It is easy to see that the hypotheses of Theorem 2.2 are met. Wavelets corresponding to the compactly supported orthogonal wavelets of I. Daubechies result; in fact, we find that

$$\begin{aligned}\hat{\phi}(\omega) &= m_0(\sigma(\omega))m_0(\sigma^2(\omega))m_0(\sigma^3(\omega))\cdots \\ &= f(\omega) + af(0.1 + \omega) + abf(1.1 + \omega) + ab^2f(11.1 + \omega) + ab^3f(111.1 + \omega) + \cdots\end{aligned}$$

where $f = \chi_{\Delta_{-1}}$.

So the scaling function becomes a lacunary Walsh series

$$\phi(x) = \tfrac{1}{2}F(x)(1 + aW_{0.1}(x) + abW_{1.1}(x) + ab^2W_{11.1}(x) + ab^3W_{111.1}(x) + \cdots),$$

where $F = \chi_{\Delta_1}$. If we identify G with the real half-line, we get the literal Walsh function series

$$\phi(x) = \tfrac{1}{2}\chi_{[0,1)}(x/2)\left(1 + aW_1(x/2) + abW_3(x/2) + ab^2W_7(x/2) + ab^3W_{15}(x/2) + \cdots\right)$$

for $x \geq 0$.

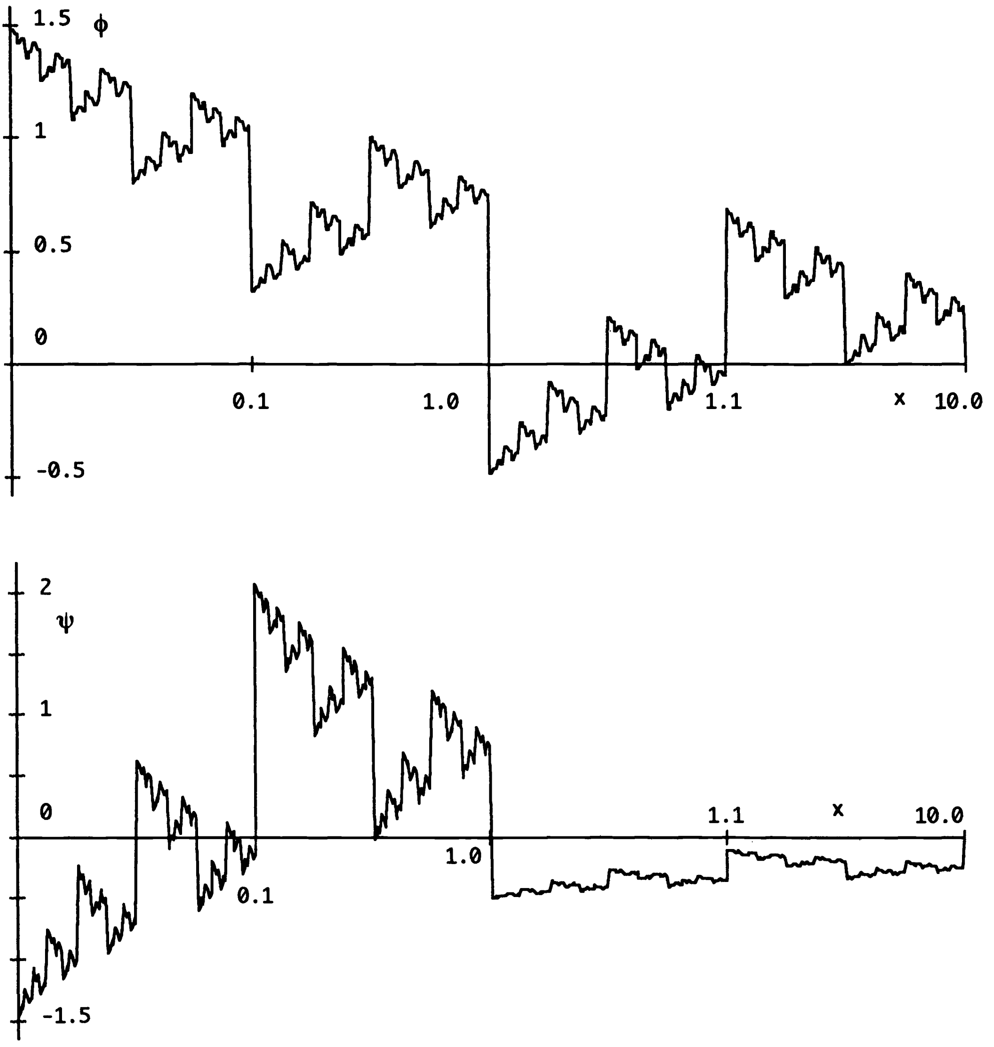


FIG. 1. *The dyadic scaling function ϕ of example* 3 (*top*) *and the corresponding dyadic wavelet ψ* (*bottom*), *with a set equal to* 0.8. *The horizontal axis is the group G identified as the half real line*; *as functions on the real line, ϕ and ψ have discontinuities at dyadic rational points, indicated here in binary notation.* (*These plottings were generated using Maple* V.)

Note that if $a = 1$ and $b = 0$, we recover the Haar scaling function $\phi(x) = \chi_{[0,1]}$. Plottings of these scaling functions for certain values of a are given in Figs. 1 and 2.

The corresponding wavelet ψ is given by (3.2) as in Example 2. Viewed as a function on the real half-line, we may write

$$\psi(x) = 2\bar{a}_0\phi(T_1(2x)) - 2\bar{a}_1\phi(2x) + 2\bar{a}_2\phi(T_3(2x)) - 2\bar{a}_3\phi(T_2(2x)),$$

where T_n is the function representing translation by n in the sense of the group G. (So the jth digit of the binary expansion of $T_n(x)$ is the sum modulo 2 of the jth binary digits of n and x. Thus, for example, $T_1(x)$ equals $x - 1$ if $j \leq x < j + 1$ for odd j and $T_1(x)$ equals $x + 1$ if $j \leq x < j + 1$ for even j.)

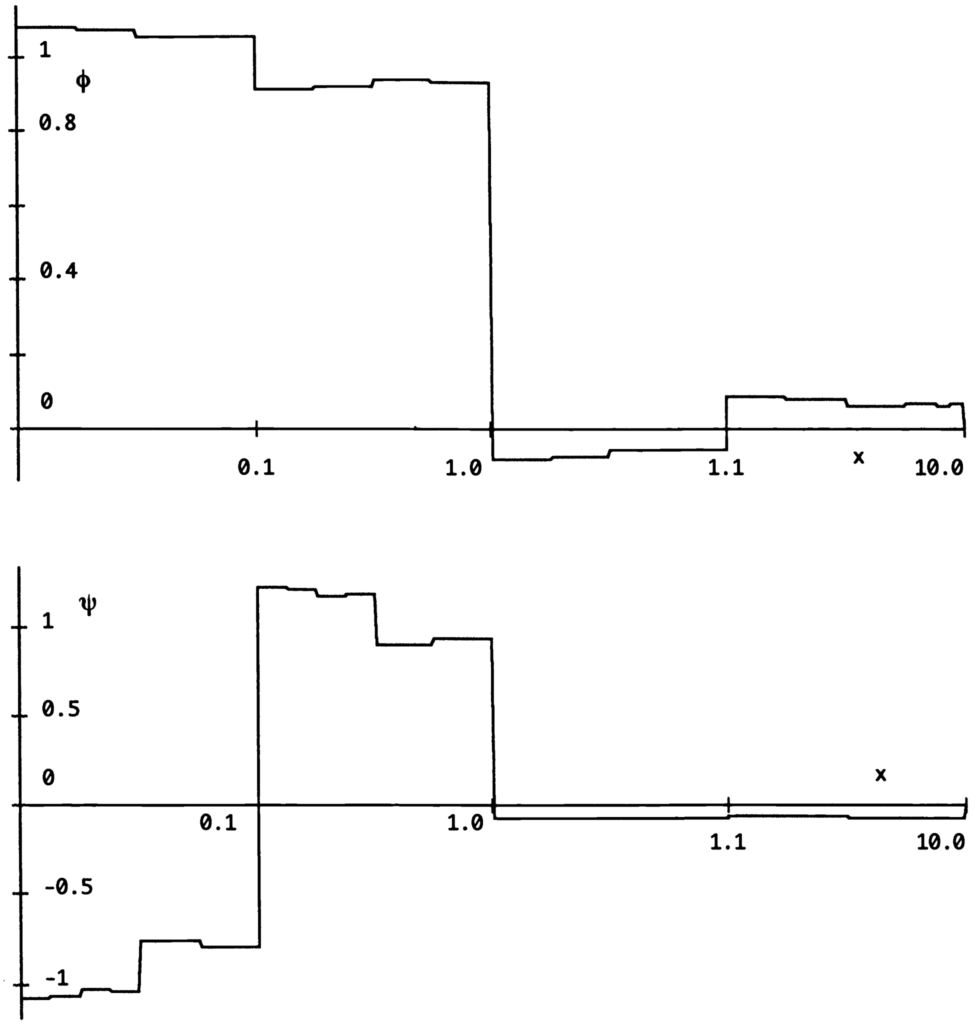


FIG. 2. *The dyadic scaling function ϕ of example* 3 (*top*) *and the corresponding dyadic wavelet* ψ (*bottom*), *with a set equal to* 0.99. *Note that for a close to* 1, *these should approach the Haar scaling function and wavelet.* (*These plottings were generated using Maple* V).

Plottings of these wavelets are given below for certain values of a; for a close to 1, the wavelet approaches the Haar wavelet.

Example 4. Our axioms (1.1.1)–(1.1.5) for dilations and translation appear to be (in some sense) one-dimensional. However, wavelets based on these axioms can be constructed on the plane or on R^n. See [4] for remarkable examples of wavelets on the plane or on R^2 based upon "matrix dilations". Their examples include the following:

Let $G = R^2$, $\Lambda = Z^2$ and $\rho(x, y) = (y, 2x)$; this choice of G, Λ, and ρ meet all the axioms. Let ϕ to be the indicator function of the unit square $0 \leq x \leq 1$, $0 \leq y \leq 1$, so

$$\phi((x, y)) = \phi(\rho(x, y) - (0, 1)) + \phi(\rho(x, y))$$

and ϕ generates an MRA of $L^2(R^2)$ as above. The "mother wavelet" ψ is then

$$\psi(x,y) = \phi(\rho(x,y) - (1,0)) - \phi(\rho(x,y))$$
$$= \begin{cases} 1 & \text{if } \frac{1}{2} < x \leq 1, 0 \leq y \leq 1, \\ -1 & \text{if } 0 \leq x \leq \frac{1}{2}, 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Of course, the axioms for dilation need not be dyadic on R^n. In fact, there may be a system of more than one mother wavelet in more general constructions; see [1].

4. Conclusions. We note that the results of this paper bear generalization in a number of ways.

We may seek to construct wavelets on other groups of this sort, such as Vilenkin groups (compact zero-dimensional Abelian metric groups), or on local fields. This presumably would require different frameworks for orthogonal wavelets; the structure of the group as well as the structure of its group of automorphisms would determine the possibilities, and we need not limit ourselves to the dyadic dilation of our axioms here.

We also anticipate that our results for the Cantor dyadic group may be strengthened to provide wavelet bases for spaces beyond $L^2(G)$. For example, [3] already provides a Littlewood–Paley theorem for the Cantor dyadic group (as well as a variety of other groups), suggesting that our wavelets might be a basis for $L^p(G)$ for $p > 1$.

Acknowledgments. I wish to express my thanks for the helpful suggestions and comments of the referee, in particular for the observation at the end of Example 2 of §3.

REFERENCES

[1] L. A. Baggett, A. Carey, W. Moran, and P. Ohring, *General existence theorem for orthonormal wavelets: An abstract approach*, Publ. Res. Inst. Math. Sci., 31 (1995), pp. 95–112.

[2] I. Daubechies, *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[3] R. E. Edwards and G. I. Gaudry, *Littlewood–Paley and Multiplier Theory*, Springer-Verlag, New York, 1985.

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

[5] E. Hewitt and K. A. Ross, *Abstract Harmonic Analysis* I, Springer-Verlag, Berlin, 1963.

[6] I. Kluvanek, *Sampling theorem in abstract harmonic analysis*, Matematicko-Fyzikálny Časopis Slovenská Akadémia Vied (Bratislava), 15 (1965), pp. 43–48.

[7] P. G. Lemarie, *Bases d'ondelettes sur les groupes de Lie stratfies*, Bull. Soc. Math. France, 117 (1989), pp. 211–233.

[8] S. G. Mallat, *Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.

[9] Y. Meyer, *Ondelettes et Opérateurs* I, Hermann, Paris, 1990.

[10] F. Schipp, W. R. Wade, and P. Simon, *Walsh Series: An Introduction to Dyadic Harmonic Analysis*, Adam Hilger, Bristol and New York, 1990.

[11] M. H. Taibleson, *Fourier analysis on local fields,* Mathematical Notes, no. 15, Princeton University Press, Princeton, 1975.

SIAM J. MATH. ANAL.
Vol. 27, No. 2, pp. 313–340, March 1996

001

ON L_p-THEORY OF STOCHASTIC PARTIAL DIFFERENTIAL EQUATIONS IN THE WHOLE SPACE*

N. V. KRYLOV†

Abstract. It is shown that equations like

$$du = (a^{ij}u_{x^ix^j} + b^iu_{x^i} + cu + f)\,dt + (\sigma^{ik}u_{x^i} + \nu^k u + g^k)\,dw_t^k, \quad t > 0,$$

with variable random coefficients and with zero initial condition have unique solutions in the Sobolev spaces W_p^2, $p \in [2, \infty)$, under natural ellipticity condition and under conditions that (i) a is uniformly continuous with respect to x, (ii) σ, ν have bounded first derivatives in x and all other coefficients are bounded, (iii) $f \in L_p$, $g \in W_p^1$. A corresponding result in the spaces of Bessel potentials H_p^n is proved, which implies that better differentiability properties of the coefficients and free terms of the equations lead to the better regularity of solutions. Applications to equations with space–time white noise are given.

Key words. stochastic equations, Bessel potentials, cylindrical white noise

AMS subject classifications. 60H15, 35R60

1. Introduction.

Evolutional stochastic partial differential equations (SPDEs) arise in many applications of probability theory and have been treated since long ago (see [14]). An example of a linear second-order SPDE is given by the following equation:

$$(1.1)\qquad du = (a^{ij}u_{x^ix^j} + b^iu_{x^i} + cu + f)\,dt + (\sigma^{ik}u_{x^i} + \nu^k u + g^k)\,dw_t^k, \quad t > 0.$$

The main purpose of this article is to develop a theory of solvability of the Cauchy problem for linear and some quasi-linear equations like (1.1) in spaces of summable functions with exponent of summability $p \geq 2$. If $p = 2$, so that we are concerned with solutions belonging to the Sobolev spaces $W_2^n(\mathbb{R}^d)$, such a theory does exist and is rather complete and satisfactory (see, for instance, [14]). Some results concerning the solvability of the first boundary-value problem in spaces like $W_2^n(D)$, where D is a smooth domain, can be found in [10] and [13]. Roughly speaking, the main tool in W_2^n-theory is integration by parts. There are also approaches based on semigroup methods [2], which work well for the equations with nonrandom leading coefficients and again in the Hilbert-space framework.

One of inconveniences of W_2^n-theory is that $W_2^n(\mathbb{R}^d) \subset C^{n-d/2}(\mathbb{R}^d)$ only if $2n > d$, and one can prove that the solutions belong to $W_2^n(\mathbb{R}^d)$ only if the coefficients are $n-2$ times continuously differentiable. Therefore, if we want to get the solutions m times continuously differentiable with respect to $x \in \mathbb{R}^d$, we have to suppose that the coefficients of the equation are more than $m + d/2 - 2$ times continuously differentiable even if the free terms are of class $C_0^\infty(\mathbb{R}^d)$. At the same time, $W_p^n(\mathbb{R}^d) \subset C^{n-d/p}(\mathbb{R}^d)$ if $pn > d$, and by taking p sufficiently large, one sees that the solutions have almost as many usual derivatives as generalized ones. Actually, exactly for this purpose the spaces $W_p^n(\mathbb{R}^d)$ with $p \geq 2$ have already been used in SPDE theory (see, for instance [14]), but the corresponding results obtained again by integration by parts were not sharp.

Another advantage of the W_p^n setting with $p \geq 2$ can be seen in the case of very popular equations with so-called cylindrical white noise (see, for instance, [15], [16],

* Received by the editors February 16, 1994; accepted for publication (in revised form) August 8, 1994. This research was partially supported by National Science Foundation grant DMS-9302516.

† School of Mathematics, University of Minnesota, Minneapolis, MN 55455.

[19], and references therein). Although these equations can be included in the general W_p^n-theory as particular examples for any $p \geq 2$ (see §4), for $p = 2$ we get only the solutions summable to any degree, and the solutions become continuous only for $p > 2$. By the way, as in [14] we consider n positive and negative, but in contrast with [14] we allow noninteger values for n. For general n we are working in the spaces of Bessel potentials $H_p^n(\mathbb{R}^d)$, and in the case of equations with cylindrical white noise, we take n slightly less than $(-3/2)$.

Our main tool is the theory of spaces $H_p^n(\mathbb{R}^d)$, borrowed from [17], together with a result from [9] which is an analog of the so-called maximal regularity property of stochastic convolutions in Hilbert spaces obtained by Da Prato (cf. [1]).

The article is organized as follows: In §2 we investigate equations with constant coefficients, which not only leads us to a class of equations which can be treated but also allows us to find a Banach space $\mathcal{H}_p^n(T)$ in which equations with variable coefficients make perfect sense. These spaces play the same role as spaces $W_p^{1,2}$ in the theory of parabolic second-order PDEs, and their properties, stated in §3, present one part of our main results. Section 3 also contains other main results, two of which are proved in §5. In §4 we consider an application of our main results to equations with cylindrical white noise. In particular, we obtain an existence theorem which extends corresponding known results (see [19]) to equations with random and variable coefficients. We also give a short proof of a generalization of a result from [15] concerning the nonexplosion of solutions of a nonlinear SPDE.

We finish the section by introducing several notations to be used throughout the paper.

Let $(\Omega, \mathcal{F}, P)$ be a complete probability space, $(\mathcal{F}_t, t \geq 0)$ be an increasing filtration of σ-algebras $\mathcal{F}_t \subset \mathcal{F}$ containing all P-null subsets of Ω, and $\mathcal{P}$ be the predictable σ-field related to $(\mathcal{F}_t, t \geq 0)$.

We fix a separable Hilbert space H. By using a well-known procedure, we identify the space of all bounded linear functionals on H with H. We also fix numbers $p \geq 2$, $K, \delta > 0$, and an integer $d \geq 1$.

Let w be an H-valued $\mathcal{F}_t$-adapted Wiener process with covariance operator $\mathbb{Q}$. Define E as the set of all real-valued linear functions e on $\mathbb{Q}^{1/2}H$ such that $|e|_E := |e\mathbb{Q}^{1/2}|_H < \infty$. Obviously, E is a Hilbert space.

Denote $L_p = L_p(\mathbb{R}^d)$, $||\cdot||_p = ||\cdot||_{L_p}$. In the case of functions $g = g(x)$ taking values in E, we write

$$||g||_p := ||\,|g|_E||_p.$$

We will need the spaces of Bessel potentials (also called the Sobolev spaces with fractional derivatives) $H_p^n = H_p^n(\mathbb{R}^d)$ for all values of $n \in (-\infty, \infty)$. We recall (see, for instance [17]) that for integers $n \geq 0$ the space H_p^n coincides with the Sobolev space $W_p^n = W_p^n(\mathbb{R}^d)$, and in general H_p^n is the closure of $C_0^\infty = C_0^\infty(\mathbb{R}^d)$ with respect to the norm

$$||u||_{n,p} := ||(I - \Delta)^{n/2}u||_p.$$

Observe that by definition the set C_0^∞ is dense in H_p^n and by Theorem 2.3 (ii) of [17] the latter is a subset of the space $\mathcal{D}$ of real-valued Schwartz distributions on $\mathbb{R}^d$ defined on C_0^∞. It is also useful to note that $||\cdot||_{n,p} \leq ||\cdot||_{m,p}$ for $m \geq n$. We apply the same definitions to E-valued functions g, specifically,

$$||g||_{n,p} := ||\,|(I - \Delta)^{n/2}g|_E||_p.$$

Finally, for a stopping time τ, we denote $(\!(0,\tau]\!] = \{(\omega,t) : 0 < t \leq \tau(\omega)\}$,

$$\mathbb{H}^n_p(\tau) = L_p((\!(0,\tau]\!], \mathcal{P}, H^n_p), \ \ \mathbb{H}^n_p = \mathbb{H}^n_p(\infty),$$

$$\mathbb{H}^n_p(\tau, E) = L_p((\!(0,\tau]\!], \mathcal{P}, H^n_p(\mathbb{R}^d, E)), \ \ \mathbb{H}^n_p(E) = \mathbb{H}^n_p(\infty, E), \ \ \mathbb{L}_{\dots}\dots = \mathbb{H}^0_{\dots}\dots .$$

The norms in these spaces are defined in an obvious way. As is conventional, elements of spaces like $\mathbb{H}^n_p$ are treated as functions rather than distributions or classes of equivalent functions, and if we know that a function of this class has a modification with better properties, then we always consider this modification. For instance, if we take $u \in H^n_p$ and $n - d/p > 0$, then u has a bounded continuous modification, but we talk about $\sup_x u(x)$ instead of sup of this modification.

2. Equations with coefficients independent of x. In this section, we consider equation (1.1) when $b^i = c = \nu^k = 0$ and the coefficients a and σ do not depend on x. Throughout the section, we fix real-valued functions $a^{ij}(t)$ and E-valued functions $\sigma^i(t)$ defined for $i, j = 1, \dots, d$ on $\Omega \times (0, \infty)$. Define

$$\alpha^{ij}(t) = \frac{1}{2}(\sigma^i(t), \sigma^j(t))_E$$

and assume that a and σ are $\mathcal{P}$-measurable functions, and in the matrix sense

$$(a^{ij}) = (a^{ij})^*, \ \ K(\delta^{ij}) \geq (a^{ij}) \geq (a^{ij} - \alpha^{ij}) \geq \delta(\delta^{ij}).$$

More precisely, our goal in this section is to investigate the equation

(2.1)
$$du(t,x) = (a^{ij}(t)u_{x^ix^j}(t,x) + f(t,x))\,dt + (\sigma^i(t)u_{x^i}(t,x) + g(t,x))\,dw_t, \ \ t > 0.$$

DEFINITION 2.1. *Denote by $\mathfrak{D}$ the set of all $\mathcal{D}$-valued functions u (written as $u(t,x)$ in a common abuse of notation) on $\Omega \times [0,\infty)$ such that for any $\phi \in C_0^\infty$,*

(i) *the function (u,ϕ) is $\mathcal{P}$-measurable,*

(ii) *for any $\omega \in \Omega$ and $T \in (0,\infty)$ we have*

$$\int_0^T \sup_{x\in\mathbb{R}^d} |(u(t,x+\cdot),\phi)|^2\,dt < \infty. \tag{2.2}$$

In the same way, we define $\mathcal{D}(E)$ and $\mathfrak{D}(E)$ by considering E-valued linear functionals on C_0^∞ (or bilinear forms $(u,\phi)(h)$, $\phi \in C_0^\infty$, $h \in \mathbb{Q}^{1/2}H$) and replacing $|\cdot|$ in (2.2) by $|\cdot|_E$.

DEFINITION 2.2. *Let $f, u \in \mathfrak{D}$, $g \in \mathfrak{D}(E)$. We say that the equality*

$$du(t,x) = f(t,x)\,dt + g(t,x)\,dw_t, \ \ t > 0 \tag{2.3}$$

holds in the sense of distributions if for any $\phi \in C_0^\infty$, with probability 1 we have

$$(u(t,\cdot),\phi) = (u(0,\cdot),\phi) + \int_0^t (f(s,\cdot),\phi)\,ds + \int_0^t (g(s,\cdot),\phi)\,dw_s \tag{2.4}$$

for all $t \geq 0$, where the last stochastic integral is understood as an Itô stochastic integral of the E-valued function $(g(s,\cdot),\phi)$ with respect to w_s.

We will understand equation (2.1) in the sense of distributions. We start with the following simple statement related to Definition 2.2.

LEMMA 2.1. *Let $f, u \in \mathfrak{D}$, $g \in \mathfrak{D}(E)$. Define*

$$x_t = \int_0^t \sigma(s)\, dw_s.$$

Then equality (2.3) *holds (in the sense of distributions) if and only if (in the sense of distributions) for the function $v(t,x) = u(t, x - x_t)$, we have*

$$dv(t,x) = [f(t, x - x_t) + a^{ij}(t) v_{x^i x^j}(t,x) - (g_{x^i}(t, x - x_t), \sigma^i(t))_E]\, dt$$

$$+ [g(t, x - x_t) - v_{x^i}(t,x)\sigma^i(t)]\, dw_t, \quad t > 0.$$

Proof. First notice that obviously $v(t,x)$, $f(t, x - x_t)$, and $(g_{x^i}(t, x - x_t), \sigma^i(t))_E$ belong to $\mathfrak{D}$ and $g(t, x - x_t)$ $v_{x^i}(t,x)\sigma^i(t)$ belong to $\mathfrak{D}(E)$. Furthermore, for any $\phi \in C_0^\infty$ the function $(u(t, x + \cdot), \phi)$ has a stochastic differential in t for any x and is infinitely differentiable with respect to x. Now our assertion immediately follows from the Itô–Wentzell formula. The lemma is proven. □

LEMMA 2.2. *Let $f \in \mathfrak{D}$, $g \in \mathfrak{D}(E)$, u_0 be a $\mathcal{D}$-valued function on Ω. Then the following assertions hold.*

(i) *In $\mathfrak{D}$ there can exist only one (up to evanescence) solution of equation* (2.1) *with the initial condition $u(0, \cdot) = u_0$.*

(ii) *Let $u \in \mathfrak{D}$ satisfy equation* (2.1) *(in the sense of distributions). Let $\{\mathcal{G}_t, t \geq 0\}$ be an increasing family of σ-fields $\mathcal{G}_t \subset \mathcal{F}_t$ and assume that $w_t = w_t^{(1)} + w_t^{(2)}$, where $w^{(i)}$ are Wiener processes such that $w^{(1)}$ is $(\mathcal{G}_t)$-adapted and the process $w^{(2)}$ is independent of $(\mathcal{G}_t)$. Assume that the processes a, f, σ, and g are $\mathcal{G}_t$-adapted, and that there exists an $n \in (-\infty, \infty)$ such that $f \in \mathbb{H}_2^n(T)$, $g \in \mathbb{H}_2^n(T, E)$ for any $T \in (0, \infty)$ and*

$$E\|u(0, \cdot)\|_{n,2}^2 < \infty.$$

Then the function $\tilde{u}(t,x) := E\{u(t,x)|\mathcal{G}_t\}$ is a solution of the equation

$$d\tilde{u} = (a^{ij}\tilde{u}_{x^i x^j} + f)\, dt + (\sigma^i \tilde{u}_{x^i} + g)\, dw_t^{(1)}, \quad t > 0. \tag{2.5}$$

Proof. (i) As always we can take $f \equiv 0, g \equiv 0$, and $u_0 \equiv 0$, and by Lemma 2.1, it suffices to consider only the case $\sigma \equiv 0$. But then we are left with a parabolic equation with coefficients independent of x, and the uniqueness of its solution in our class of functions is a standard fact.

(ii) Actually, the statement means that there exists a solution $\tilde{u}$ of equation (2.5) such that for any $\phi \in C_0^\infty$ and $t \geq 0$,

$$(\tilde{u}(t, \cdot), \phi) = E\{(u(t, \cdot), \phi)|\mathcal{G}_t\} \quad \text{(a.s.)}.$$

To prove this version of our statement, we first notice that, according to [14], equation (2.1) has a unique solution v in the space $\mathbb{H}_2^{n+1}(T)$ for any T. Moreover, v is continuous (a.s.) as an H_2^n-valued process and

$$E \sup_{t \leq T} \|v(t, \cdot)\|_{n,2}^2 < \infty \quad \forall T < \infty, \tag{2.6}$$

so that v is a $\mathfrak{D}$-solution of (2.1). From (i), it follows that our function u belongs to $\mathbb{H}_2^{n+1}(T)$ for any T, and (2.6) holds for u. Furthermore, in the sense of the Hilbert space H_2^{n-1} with probability 1 for all t at once,

$$u(t) = u_0 + \int_0^t [a^{ij}(s)u_{x^ix^j}(s) + f(s)]\,ds + \int_0^t [\sigma^i(s)u_{x^i}(s) + g(s)]\,dw_s.$$

By Theorem 1.4.7 of [14], or rather by its Hilbert-space counterpart, there exists an H_2^n-valued, $\mathcal{G}_t$-predictable function $\bar{u}(t)$ such that for almost any t we have $\bar{u}(t) = E\{u(t)|\mathcal{G}_t\}$ (a.s.) and

$$\bar{u}(t) = u_0 + \int_0^t [a^{ij}(s)\bar{u}_{x^ix^j}(s) + f(s)]\,ds + \int_0^t [\sigma^i(s)\bar{u}_{x^i}(s) + g(s)]\,dw_s^{(1)} \tag{2.7}$$

for almost all t and ω. The right-hand side here is a continuous H_2^{n-1}-valued process; therefore, we can and will take $\bar{u}(t)$ to be a continuous process in H_2^{n-1}. We denote this version of $\bar{u}(t)$ by $\tilde{u}(t)$, and we show that $\tilde{u}$ is the function we need.

By construction, $\tilde{u} \in \mathfrak{D}$ and $\tilde{u}$ is a solution of (2.5). It remains only to observe that, again by [14, Thm. 1.4.7], for any t the expression $E\{u(t)|\mathcal{G}_t\}$ can be represented by the right-hand side of (2.7) almost surely. The lemma is proven. □

THEOREM 2.1. *Take $n \in \mathbb{R}$ and let $f \in \mathbb{H}_p^{n-1}$, $g \in \mathbb{H}_p^n(E)$. Then*

(i) *equation* (2.1) *with zero initial condition has a (unique) solution*

$$u \in \bigcap_T \mathbb{H}_p^{n+1}(T);$$

(ii) *for this solution, we have $u \in C((0,\infty), H_p^n)$ almost surely, and*

$$||u_{xx}||_{\mathbb{H}_p^{n-1}} \le N(d,n,p,\delta,K)(||f||_{\mathbb{H}_p^{n-1}} + ||g||_{\mathbb{H}_p^n(E)}); \tag{2.8}$$

(iii) *for any $\lambda, t > 0$,*

$$E\sup_{s\le t}(e^{-p\lambda s}||u(s,\cdot)||_{n,p}^p) \le N(||e^{-\lambda s}f||_{\mathbb{H}_p^{n-1}(t)}^p + ||e^{-\lambda s}g||_{\mathbb{H}_p^n(t,E)}^p), \tag{2.9}$$

where $N = N(d,p,\delta,K,\lambda)$.

To prove the theorem, we will invoke two lemmas. The first of them is the first main result of [9]. Take an $\mathbb{R}^d$-valued (standard) Wiener process B_t and define

$$T_t h(x) = Eh(x + B_t).$$

LEMMA 2.3. *Let $-\infty \le a < b \le \infty$, $g \in L_p((a,b) \times \mathbb{R}^d, E)$. Then*

$$\int_{\mathbb{R}^d}\int_a^b \left[\int_a^t |\nabla T_{t-s}g(s,\cdot)(x)|_E^2\,ds\right]^{p/2} dt\,dx \le N\int_{\mathbb{R}^d}\int_a^b |g(t,x)|_E^p\,dt\,dx,$$

where the constant N depends only on d and p.

LEMMA 2.4. *Let $\sigma^k \equiv 0, a^{ij} \equiv (1/2)\delta^{ij}$, $n = 0$. Then the assertions of Theorem* 2.1 *hold true. Moreover,*

(iii)$'$ *for any $\lambda, t > 0$, we have*

$$E\sup_{s\le t}(e^{-p\lambda s}||u(s,\cdot)||_p^p) + E\int_0^t e^{-p\lambda s}||\,|u|^{(p-2)/p}|u_x|^{2/p}(s,\cdot)||_p^p\,ds$$

$$\leq N(d,p,\delta,K,\lambda)(||e^{-\lambda s}f||^p_{\mathbb{H}^{-1}_p(t)} + ||e^{-\lambda s}g||^p_{\mathbb{L}_p(t,E)}). \tag{2.10}$$

Proof. It is well known that there exists a continuous linear operator

$$P : H^{-1}_p \to (L_p)^{d+1}$$

such that if $h \in H^{-1}_p$ and $Ph = (h_0, \tilde{h}^1, \dots, \tilde{h}^d)$, then $h = h_0 + \operatorname{div}\tilde{h}$ and

$$||\tilde{h}||_p + ||h_0||_p \leq N(d,p)||h||_{-1,p}. \tag{2.11}$$

Actually, one can take $\tilde{h} = -\operatorname{grad}(1-\Delta)^{-1}h$ and $h_0 = h - \operatorname{div}\tilde{h}$.

Define $(f_0, \tilde{f}) = Pf$. Then equation (2.1) takes the form

$$du = \left(\frac{1}{2}\Delta u + f_0 + \operatorname{div}\tilde{f}\right) dt + g\, dw_t \tag{2.12}$$

and we supply it with zero initial condition. We will prove that for arbitrary $f_0, \tilde{f}^i \in \mathbb{L}_p$, assertions (i), (ii) (with $n = 0$), and (iii)$'$ hold for (2.12) in place of (2.1). Of course, in (2.8) and (2.10) we take $f = f_0 + \operatorname{div}\tilde{f}$.

We can certainly confine ourselves to the case in which

$$f_0(t,x) = \sum_{i=1}^m I_{(\tau_{i-1},\tau_i]}(t) f_{0i}(x), \quad \tilde{f}(t,x) = \sum_{i=1}^m I_{(\tau_{i-1},\tau_i]}(t)\tilde{f}_i(x),$$

$$g(t,x) = \sum_{j=1}^m g^j(t,x)h_j, \quad g^j(t,x) = \sum_i^m I_{(\tau_{i-1},\tau_i]}(t) g_{ij}(x), \tag{2.13}$$

where (h_j) is an orthonormal basis in H of eigenvectors of $\mathbb{Q}$, $m < \infty$, τ_i are bounded stopping times, $\tau_{i-1} \leq \tau_i$, and $f_{0i}, \tilde{f}_i, g_{ij} \in C^\infty_0$.

Set $w^j_t = (h_j, w_t)_H$,

$$v(t,x) = \int_0^t g(s,x)\, dw_s = \sum_{i,j=1}^m g_{ij}(x)(w^j_{t\wedge\tau_i} - w^j_{t\wedge\tau_{i-1}}),$$

$$u(t,x) = v(t,x) + \int_0^t T_{t-s}\left[\frac{1}{2}\Delta v + f\right](s,\cdot)(x)\, ds, \quad \forall t \geq 0.$$

As easy to see the function $u - v$ is infinitely differentiable in (t,x) and satisfies the equation

$$\frac{\partial z}{\partial t} = \frac{1}{2}\Delta z + \frac{1}{2}\Delta v + f.$$

It follows that for any x the function $u(t,x)$ satisfies (2.12) almost surely. From our explicit formulas and from the finiteness of f and g, it also follows that $u \in C([0,\infty), H^n_p)$ for any n (and for any ω), that $u \in \mathfrak{D}$, and that u is a solution of (2.12).

Next, we want to obtain some bounds of norms of u. Let

$$u_1(t,x) = \int_0^t T_{t-s} f(s,x)\,ds.$$

According to [12] (for any ω),

$$\tag{2.14} ||u_{1xx}||_{L_p(\mathbb{R}_+,H_p^{-1})} \le N||f||_{L_p(\mathbb{R}_+,H_p^{-1})}.$$

Furthermore,

(2.15)
$$N^{-1}||u_{xx} - u_{1xx}||^p_{\mathbb{H}_p^{-1}} \le ||u_x - u_{1x}||^p_{\mathbb{L}_p} = \int_0^\infty \int_{\mathbb{R}^d} E|u_x - u_{1x}|^p(t,x)\,dx\,dt.$$

To make some further transformations of this formula, we note that if $z^k = z^k(x)$ are bounded $\mathcal{B}(\mathbb{R}^d)$-measurable functions, then by Itô's formula applied to the increment over $[0,t]$ of

$$\left(\int_r^t T_{t-s} z^k\,ds\right)(w^k_{r\wedge\tau_2} - w^k_{r\wedge\tau_1})$$

as a function of r, we obtain (a.s.)

$$0 = -\int_0^t (w^k_{r\wedge\tau_2} - w^k_{r\wedge\tau_1}) T_{t-r} z^k\,dr + \int_0^t I_{(\tau_1,\tau_2]}(r)\left(\int_r^t T_{t-s} z^k\,ds\right) dw^k_r.$$

By using this for our particular g or by using the stochastic version of the Fubini theorem, for any $t \ge 0$ and $x \in \mathbb{R}^d$ we get

$$u_x(t,x) - u_{1x}(t,x) = v_x(t,x) + \int_0^t T_{t-s} \sum_{k=1}^m \int_0^s \frac{1}{2}\Delta g^k_x(r,x)\,dw^k_r\,ds$$

$$= v_x(t,x) - \sum_{k=1}^m \int_0^t \int_r^t \frac{d}{ds} T_{t-s} g^k_x(r,x)\,ds\,dw^k_r = \sum_{k=1}^m \int_0^t T_{t-r} g^k_x(r,x)\,dw^k_r \quad \text{(a.s.)}.$$

Hence by the Burkholder–Davis–Gundy inequality,

$$\begin{aligned} E|u_x - u_{1x}|^p(t,x) &\le NE\left[\int_0^t \sum_{k=1}^m \lambda_k |T_{t-r} g^k_x(r,x)|^2\,dr\right]^{p/2} \\ &= NE\left[\int_0^t |T_{t-r} g_x(r,x)|^2_E\,dr\right]^{p/2}, \end{aligned}$$

λ_k being the eigenvalue of $\mathbb{Q}$ corresponding to h_k, so that $E|w^k_t|^2 = \lambda_k t$, $k = 1,2,3,\ldots$.

By plugging this into (2.15) and by applying Lemma 2.3, we obtain

$$||u_x - u_{1x}||^p_{\mathbb{L}_p} \le NE\int_0^\infty \int_{\mathbb{R}^d} \left[\int_0^t |\nabla T_{t-s} g(s,x)|^2_E\,ds\right]^{p/2} dx\,dt \le N||g||^p_{\mathbb{L}_p(E)}.$$

This along with (2.14) gives us (2.8) for $n = 0$.

To prove (2.10) with a sufficiently large λ, it suffices to repeat the corresponding arguments from [11] or [14] related to integration by parts in a formula for $|u(t,x)|^p \exp(-\lambda pt)$, which one gets from equation (2.12) satisfied pointwise. One might only observe that by denoting $q = p/(p-1)$ and using the Hölder inequality and (2.11), one gets

$$\int_{\mathbb{R}^d} |u(s,x)|^{p-2} u_x(s,x) \cdot \tilde{f}(s,x)\,dx \le \int_{\mathbb{R}^d} (|u|^{(p-2)/2}|u_x|)^q |u|^{q(p-2)/2}\,dx$$

$$+ \|\tilde{f}(s,\cdot)\|_p^p \le N\|f(s,\cdot)\|_{-1,p}^p + N_1\|u(s,\cdot)\|_p^p + \frac{1}{2}\|\,|u(s,\cdot)|^{(p-2)/p}|u_x(s,\cdot)|^{2/p}\|_p^p,$$

$$\int_{\mathbb{R}^d} |u(s,x)|^{p-2}u(s,x)f_0(s,x)\,dx \le \|f_0(s,\cdot)\|_p^p + \|u(s,\cdot)\|_p^p$$
$$\le N\|f(s,\cdot)\|_{-1,p}^p + \|u(s,\cdot)\|_p^p.$$

From estimates (2.10) and (2.8), we conclude that $u \in \bigcap_T \mathbb{H}_p^1(T)$. Finally, the assertion about the arbitrariness of λ in (2.10) can be easily justified by rescaling arguments when instead of f, g, and w one takes $(c^2 f, cg)(c^2 t, cx)$ and $c^{-1}w_{c^2 t}$ and gets $u(c^2 t, cx)$ instead of $u(t,x)$. The lemma is proven. □

Remark 2.1. Although (2.14) holds for all $p \in (1,\infty)$, it follows from [9] that (2.8) is not true if $p < 2$.

Proof of Theorem 2.1. Since one can apply the operator $(I-\Delta)^{n/2}$ to both sides of (2.1), it suffices to prove assertions (i) and (ii) only for $n = 0$. Furthermore, our norms are translation invariant and $\bigcap_T \mathbb{H}_p^{n+1}(T) \subset \mathfrak{D}$; hence by Lemma 2.1, we need to consider only the case $\sigma \equiv 0$. As in Lemma 2.4, we can assume that f and g are as in (2.13). In this case, as we know from [11] and [14], equation (2.1) has a unique solution u which belongs to $C_b([0,T] \times \mathbb{R}^d)$ and $C([0,T], L_2)$ almost surely for any $T < \infty$. It follows that $u \in C([0,T], L_p)$ almost surely for any $T < \infty$.

Since the matrix a is uniformly nondegenerate, by making a nonrandom time change, we can reduce the general case to the case $2a \ge I$. In this case, define the matrix-valued function $\bar{\sigma}(t) = \bar{\sigma}^*(t) \ge 0$ as a solution of the equation $\bar{\sigma}^2(t) + I = 2a(t)$. Furthermore, without loss of generality, we assume that on our probability space we are also given a d-dimensional Wiener process B_t independent of $\mathcal{F}_t$.

Now consider the equation

$$dv(t,x) = \left[\frac{1}{2}\Delta v(t,x) + f\left(t, x - \int_0^t \bar{\sigma}(s)\,dB_s\right)\right]dt + g\left(t, x - \int_0^t \bar{\sigma}(s)\,dB_s\right)dw_t$$

with zero initial condition. By Lemmas 2.4, 2.1, and 2.2, there is a function v possessing properties (i) through (iii) listed in Theorem 2.1 such that

$$u(t,x) = E\left\{v\left(t, x + \int_0^t \bar{\sigma}(s)\,dB_s\right)|\mathcal{F}_t\right\}$$

in the sense explained in the proof of Lemma 2.2. We use this representation of u and the already mentioned relation $u \in C([0,T], L_p)$. Bearing also in mind the properties of v and using the Hölder inequality, we immediately get all the needed properties of u, thus finishing the proof of our theorem. □

Remark 2.2. By using the self-similarity of equation (2.1), one can obtain further estimates from estimates like (2.9). For instance, bearing in mind that $H^1_p = W^1_p$, one sees that for $n = 1$ and $\lambda = 1/p$, estimate (2.9) implies that

$$E\sup_{s\le t}\{||u_x(s,\cdot)||^p_p + ||u(s,\cdot)||^p_p\} \le N(d,p,\delta,K)e^t(||f||^p_{\mathbb{L}_p(t)} + ||g_x||^p_{\mathbb{L}_p(t,E)} + ||g||^p_{\mathbb{L}_p(t,E)}).$$

Let us take a constant $c > 0$ and consider $(c^2f, cg)(c^2t, cx), c^{-1}w_{c^2t}$, and $u(c^2t, cx)$ instead of f, g, w, and u. Then from the last estimate, we get

$$E\sup_{s\le t}\{c^{p-d}||u_x(c^2s,\cdot)||^p_p + c^{-d}||u(c^2s,\cdot)||^p_p\}$$

$$\le Ne^t(c^{2p-(d+2)}||f||^p_{\mathbb{L}_p(c^2t)} + c^{2p-(d+2)}||g_x||^p_{\mathbb{L}_p(c^2t,E)} + c^{p-(d+2)}||g||^p_{\mathbb{L}_p(c^2t,E)}),$$

the constant N taken the same as above. It follows that for $c, t \ge 0$,

$$E\sup_{s\le t}\{||u_x(s,\cdot)||^p_p + c^{-p}||u(s,\cdot)||^p_p\}$$

$$\le Ne^{t/c^2}c^{p-2}(||f||^p_{\mathbb{L}_p(t)} + ||g_x||^p_{\mathbb{L}_p(t,E)} + c^{-p}||g||^p_{\mathbb{L}_p(t,E)}).$$

Upon setting $c^2 = t$ and considering $(I - \Delta)^{(n-1)/2}u$ instead of u, we conclude that under the conditions of Theorem 2.1 for any $t > 0$,

$$E\{\sup_{s\le t}||u_x(s,\cdot)||^p_{n-1,p} + t^{-p/2}\sup_{s\le t}||u(s,\cdot)||^p_{n-1,p}\}$$

$$\le N(d,p,\delta,K)t^{(p-2)/2}(||f||^p_{\mathbb{H}^{n-1}_p(t)} + ||g_x||^p_{\mathbb{H}^{n-1}_p(t,E)} + t^{-p/2}||g||^p_{\mathbb{H}^{n-1}_p(t,E)}).$$

We will later prove a much deeper estimate than (2.9).

3. A Banach-space setting and main results. Here we state our main results, the first two of which are proved in §5. Fix a number $T \in (0,\infty)$ and a stopping time $\tau \le T$. For $n \in \mathbb{R}$ and

$$(f,g) \in \mathcal{F}^n_p(\tau) := \mathbb{H}^n_p(\tau) \times \mathbb{H}^{n+1}_p(\tau, E),$$

set

$$||(f,g)||_{\mathcal{F}^n_p(\tau)} := ||f||_{\mathbb{H}^n_p(\tau)} + ||g||_{\mathbb{H}^{n+1}_p(\tau,E)}.$$

DEFINITION 3.1. *For a function $u \in \mathbb{H}^n_p(\tau)$, we write $u \in \mathcal{H}^n_p(\tau)$ if there exists $(f,g) \in \mathcal{F}^{n-2}_p(\tau)$ such that for any $\phi \in C^\infty_0$, equality* (2.4) *holds almost surely for all $t \le \tau$ and $u(0,\cdot) \in L_p(\Omega, H^n_p)$. In this case, we define $\mathcal{H}^n_{p,0}(\tau) = \mathcal{H}^n_p(\tau) \cap \{u : u(0,\cdot) = 0\}$,*

$$||u||_{\mathcal{H}^n_p(\tau)} = ||u_{xx}||_{\mathbb{H}^{n-2}_p(\tau)} + ||(f,g)||_{\mathcal{F}^{n-2}_p(\tau)} + (E||u(0,\cdot)||^p_{n,p})^{1/p}. \tag{3.1}$$

Remark 3.1. Since ordinary and stochastic integrals are "incomparable," the function u defines the couple f, g uniquely. Therefore, the norm in (3.1) is well defined. In (3.1) we use such a restrictive norm of the initial value as $||u(0,\cdot)||_{n,p}$

only for simplicity. Actually, we will always get rid of the initial condition by considering $u(t,x) - u(0,x)$ instead of u, and in (3.1) we could replace $||u(0,\cdot)||_{n,p}$ by $||u(0,\cdot)||_{n-2/p,p}$ if instead of $u(t,x)-u(0,x)$ we take $u(t,x) - T_t u(0,x)$.

The following theorem contains some basic properties of the spaces $\mathcal{H}_p^n(\tau)$. As we will see in §5, assertion (ii) in this theorem is a simple consequence of Theorem 2.1. By using Theorem 2.1, it can obviously be obtained not only for $p=2$ but for any $p \geq 2$. However, for $p > 2$ the much stronger statement (iii) is available.

THEOREM 3.1. (i) *The spaces* $\mathcal{H}_p^n(\tau)$ *and* $\mathcal{H}_{p,0}^n(\tau)$ *are Banach spaces with the norm given by* (3.1).

(ii) *For any function* $u \in \mathcal{H}_2^n(\tau)$, *we have* $u \in C([0,\tau], H_2^{n-1})$ *(a.s.) and*

$$E \sup_{t \leq \tau} ||u(t,\cdot)||_{n-1,2}^2 \leq N(d,n,T)||u||_{\mathcal{H}_2^n(\tau)}^2.$$

(iii) *If* $p > 2$, $1/2 > \beta > \alpha > 1/p$, *then for any function* $u \in \mathcal{H}_p^n(\tau)$, *we have* $u \in C^{\alpha-1/p}([0,\tau], H_p^{n-2\beta})$ *(a.s.) and for any stopping time* $\eta \leq \tau$,

(3.2)
$$E||u(t\wedge\eta,\cdot) - u(s\wedge\eta,\cdot)||_{n-2\beta,p}^p \leq N(d,\beta,p,T)|t-s|^{\beta-1/p}||u||_{\mathcal{H}_p^n(\tau)}^p \quad \forall t,s \leq T;$$

$$E||u(t,\cdot)||_{C^{\alpha-1/p}([0,\tau],H_p^{n-2\beta})}^p \leq N(d,\beta,\alpha,p,T)||u||_{\mathcal{H}_p^n(\tau)}^p. \tag{3.3}$$

(iv) *If* $\alpha := n - d/p > 0$ *and* $u \in \mathcal{H}_p^n(\tau)$, *then* $u \in L_p(\Omega \times (0,\tau), \mathcal{C}^\alpha(\mathbb{R}^d))$, *where* $\mathcal{C}^\alpha(\mathbb{R}^d)$ *is the Zygmund space (which differs from the ordinary Hölder spaces* $C^\alpha(\mathbb{R}^d)$ *only if* α *is an integer). In addition,*

$$E\int_0^\tau ||u(t,\cdot)||_{\mathcal{C}^\alpha(\mathbb{R}^d)}^p\,dt \leq N(d,n,p,T)||u||_{\mathcal{H}_p^n(\tau)}^p.$$

(v) *If* $m < n$, $n - d/p = m - d/q$, *and* $u \in \mathcal{H}_p^n(\tau)$, *then*

$$E\int_0^\tau ||u(t,\cdot)||_{m,q}^p\,dt \leq N(d,n,m,p,T)||u||_{\mathcal{H}_p^n(\tau)}^p.$$

(vi) *If* $q \geq p > 2$ *and* $\theta \in (0,1)$, *then for*

$$m < n + \frac{d}{q} - \frac{d+2(1-\theta)}{p}, \quad u \in \mathcal{H}_p^n(\tau),$$

we have $u \in L_{p/\theta}((0,\tau), H_q^m)$ *(a.s.) and*

$$E\left(\int_0^\tau ||u(t,\cdot)||_{m,q}^{p/\theta}\,dt\right)^\theta \leq N(d,p,q,n,m,\theta,T)||u||_{\mathcal{H}_p^n(\tau)}^p.$$

In particular (take $\theta = p/q$*),*

$$E\left(\int_0^\tau ||u(t,\cdot)||_{m,q}^q\,dt\right)^{p/q} \leq N(d,p,q,n,m,T)||u||_{\mathcal{H}_p^n(\tau)}^p$$

if

$$q > p > 2, \quad m < n - (d+2)\left(\frac{1}{p} - \frac{1}{q}\right).$$

Remark 3.2. Assertions (ii) and (iii) obviously imply that $\mathcal{H}_p^n(\tau) \subset \mathbb{H}_p^n(\tau)$ and

$$||u||_{\mathbb{H}_p^n(\tau)} \leq N(d,n,T)||u||_{\mathcal{H}_p^n(\tau)}.$$

We are now going to state our main results concerning SPDEs. Fix $n \in (-\infty,\infty)$ and fix a number $\gamma \in [0,1)$ such that $\gamma = 0$ if $n = 0, \pm 1, \pm 2, \dots$; otherwise $\gamma > 0$ and is such that $|n| + \gamma$ is not an integer. Consider the general equation

(3.4)

$$du(t,x) = [a^{ij}(t,x)u_{x^ix^j}(t,x) + f(u,t,x)]\,dt + [\sigma^i(t,x)u_{x^i}(t,x) + g(u,t,x)]\,dw_t,$$

where a^{ij} and f are real-valued and σ^i and g are E-valued functions defined for $\omega \in \Omega$, $t \geq 0$, $x \in \mathbb{R}^d$, $u \in H_p^{n+2}$, $i,j = 1,\dots,d$.

We make the following assumptions, where as in §2 we define

$$\alpha^{ij}(t,x) = \frac{1}{2}(\sigma^i(t,x), \sigma^j(t,x))_E.$$

Assumption 3.1 (coercivity). For any $\omega \in \Omega$, $t \geq 0$, $x, \lambda \in \mathbb{R}^d$, we have

$$K|\lambda|^2 \geq [a^{ij}(t,x) - \alpha^{ij}(t,x)]\lambda^i\lambda^j \geq \delta|\lambda|^2.$$

Assumption 3.2 (uniform continuity of a and σ). For any $\varepsilon > 0, i, j$, there exists a $\kappa_\varepsilon > 0$ such that

$$|a^{ij}(t,x) - a^{ij}(t,y)| + ||\sigma^i(t,x) - \sigma^i(t,y)||_E^2 \leq \varepsilon \tag{3.5}$$

whenever $|x - y| \leq \kappa_\varepsilon, t \geq 0, \omega \in \Omega$.

Actually, below we impose much stronger conditions on σ.

Assumption 3.3. For any $x \in \mathbb{R}^d$ and $e \in \mathbb{Q}^{1/2}H$, the functions $a^{ij}(t,x)$ and $\sigma^i(t,x)e$ are real-valued predictable functions, and for any $\omega \in \Omega$ and $t \geq 0$, we have

$$a^{ij}(t,\cdot) \in C^{|n|+\gamma}, \quad \sigma^i(t,\cdot) \in C^{|n|+\gamma+1}(\mathbb{R}^d, E).$$

Assumption 3.4. For any $u \in H_p^{n+2}$, the functions $f(u,t,x)$ and $g(u,t,x)$ are predictable as functions taking values in H_p^n and $H_p^{n+1}(\mathbb{R}^d, E)$, respectively.

Assumption 3.5. For any $t \geq 0, \omega, i, j$,

$$||a^{ij}(t,\cdot)||_{C^{|n|+\gamma}} + ||\sigma^i(t,\cdot)||_{C^{|n|+\gamma+1}(\mathbb{R}^d,E)} \leq K, \quad (f(0,\cdot,\cdot), g(0,\cdot,\cdot)) \in \mathcal{F}_p^n(\tau).$$

Assumption 3.6. For any $\varepsilon > 0$, there exists a constant K_ε such that for any $u, v \in H_p^{n+2}$, t, ω, we have

$$\begin{aligned} ||f(u,t,\cdot) - f(v,t,\cdot)||_{n,p} &+ ||g(u,t,\cdot) - g(v,t,\cdot)||_{n+1,p} \\ &\leq \varepsilon||u - v||_{n+2,p} + K_\varepsilon||u - v||_{n,p}. \end{aligned} \tag{3.6}$$

THEOREM 3.2. *Let Assumptions* 3.1–3.6 *be satisfied and let*

$$u_0 \in L_p(\Omega, \mathcal{F}_0, H_p^{n+2}).$$

Then the Cauchy problem for equation (3.4) *on* $[0,\tau]$ *with the initial condition* $u(0,\cdot) = u_0$ *has a unique solution* $u \in \mathcal{H}_p^{n+2}(\tau)$. *For this solution, we have*

$$||u||_{\mathcal{H}_p^{n+2}(\tau)} \leq N(||f(0,\cdot,\cdot)||_{\mathbb{H}_p^n(\tau)} + ||g(0,\cdot,\cdot)||_{\mathbb{H}_p^{n+1}(\tau)} + (E||u_0||_{n+2,p}^p)^{1/p}),$$

where the constant N depends only on $d, n, \gamma, p, \delta, K, T$, and the functions κ_ε and K_ε.

To discuss the theorem we need the following lemma.

LEMMA 3.1. *For real-valued (measurable) functions a on $\mathbb{R}^d$, define $\bar{a} = ||a||_{C^{|n|+\gamma}}$ if $n \neq 0$, $\bar{a} = \sup_x |a(x)| =: ||a||_B$ if $n = 0$. Let $\zeta \in C_0^\infty$ be a nonnegative function such that $\int \zeta(x)\,dx = 1$ and define $\zeta_k(x) = k^{-d}\zeta(x/k)$, $k = 1, 2, 3, \dots$. We assert that for any $u \in H_p^n$, we have the following:*

(i) $||au||_{n,p} \leq N\bar{a}||u||_{n,p}$, *where the constant N depends only on d, p, n, and γ;*

(ii) $||u * \zeta_k||_{n,p} \leq ||u||_{n,p}$, $||u - u * \zeta_k||_{n,p} \to 0$.

Proof. If $n \neq 0, \pm1, \pm2, \dots$ (and $\gamma > 0$), then one gets (i) by Corollary 2.8.2 (ii) of [17]. If n is a nonnegative integer, then (i) follows from the Leibnitz rule (and the fact that $H_p^n = W_p^n$). For negative integers (and generally for negative n) (i) follows easily by duality.

As for (ii), the first inequality follows from the Minkowski inequality and the second one is derived as usual owing to denseness of C_0^∞ in H_p^n. The lemma is proven. □

Remark 3.3. Of course, by a solution to the Cauchy problem for equation (3.4) on $[0, \tau]$ with the given initial condition u_0, we understand a function $u \in \mathcal{H}_p^{n+2}(\tau)$ such that for any test function $\phi \in C_0^\infty$, almost surely one has

$$\begin{aligned}(u(t,\cdot), \phi) = (u_0, \phi) &+ \int_0^t (a^{ij}(s,\cdot)u_{x^ix^j}(s,\cdot) + f(u,s,\cdot), \phi)\,ds \\ &+ \int_0^t (\sigma^i(s,\cdot)u_{x^i}(s,\cdot) + g(u,s,\cdot), \phi)\,dw_s, \quad \forall t \in [0,\tau],\end{aligned}$$

where by Lemma 3.1 $a^{ij}u_{x^ix^j} \in H_p^n$, $\sigma^i u_{x^i} \in H_p^{n+1}(\mathbb{R}^d, E)$ whenever $u \in H_p^{n+2}$.

Remark 3.4. Two main ideas in the proof of this theorem are quite standard. The first one, reduction to equations with constant coefficients, will be seen very clearly. The second one, which is somewhat hidden, actually, consists of the introduction of the new unknown function $v = (I - \Delta)^{n/2}u$, which reduces the case of general n to the case $n = 0$. The equation for v is pseudodifferential, and we note that more general pseudodifferential equations can be considered too.

Remark 3.5. By Theorem 14.2 of [5], for any $u \in H_p^{n+2}$ and $m \in [n, n+2]$, we have

$$||u||_{m,p} \leq N||u||^{\theta}_{n+2,p}||u||^{1-\theta}_{n,p},$$

where $\theta = (m - n)/2$ and N depends only on d, n, m, and p. This shows that the last norm in (3.6) can be replaced by $||u - v||_{n+\varepsilon+1,p}$ once $|\varepsilon| < 1$.

Remark 3.6. A typical application of Theorem 3.2 occurs when $f(u,t,x) = b^i(t,x)u_{x^i} + c(t,x)u + f(t,x)$ and $g(u,t,x) = \nu(t,x)u + g(t,x)$, so that (3.4) becomes

$$du = (a^{ij}u_{x^ix^j} + b^iu_{x^i} + cu + f)\,dt + (\sigma^iu_{x^i} + \nu u + g)\,dw_t. \tag{3.7}$$

To describe the appropriate assumptions, we take $\varepsilon \in (0,1)$ and denote

$$\begin{array}{llll} n_b = n + \gamma & \text{if} \ \ n \geq 0, & n_b = 0 & \text{if} \ \ n \in (-1, 0], \\ n_\nu = n + 1 + \gamma & \text{if} \ \ n \geq -1, & n_\nu = 0 & \text{if} \ \ n \in (-2, -1], \\ n_c = n + \gamma & \text{if} \ \ n \geq 0, & n_c = 0 & \text{if} \ \ n \in (-2, 0], \end{array}$$

$$\begin{array}{ll} n_b = -n - 1 + \varepsilon & \text{if} \ \ n \leq -1, \\ n_\nu = -n - 2 + \varepsilon & \text{if} \ \ n \leq -2, \\ n_c = -n - 2 + \varepsilon & \text{if} \ \ n \leq -2. \end{array}$$

Assume that b, c,and ν are appropriately measurable and

$$b^i(t,\cdot) \in C^{n_b}, \;\; c(t,\cdot) \in C^{n_c}, \;\; \nu(t,\cdot) \in C^{n_\nu}(\mathbb{R}^d, E),$$

$$f(t,\cdot) \in H^n_p, \;\; g(t,\cdot) \in H^{n+1}_p(\mathbb{R}^d, E),$$

$$||b^i(t,\cdot)||_{C^{n_b}} + ||c(t,\cdot)||_{C^{n_c}} + ||\nu(t,\cdot)||_{C^{n_\nu}(\mathbb{R}^d,E)} \leq K, \quad (f(\cdot,\cdot), g(\cdot,\cdot)) \in \mathcal{F}^n_p(\tau),$$

where we understand the space C^m as $B(\mathbb{R}^d)$ if $m = 0$. It turns out then that the assumptions of Theorem 3.2 about $f(u,t,x)$ and $g(u,t,x)$ are satisfied. To show this, it suffices to apply Remark 3.5 and to notice that, for instance, if $n \geq -1$, then $||\nu u||_{n+1,p} \leq N||u||_{n+1,p}$ by Lemma 3.1; if $n \in (-2,-1]$, then obviously $||\nu u||_{n+1,p} \leq ||\nu u||_p \leq N||u||_p = N||u||_{n+1+(-n-1),p}$, and $-n-1 \in [0,1)$; if $n \leq -2$, then Lemma 3.1 yields $||\nu u||_{n+1,p} \leq ||\nu u||_{n+2-\varepsilon_1,p} \leq N||u||_{n+2-\varepsilon_1,p} \leq N||u||_{n+2-\varepsilon,p}$, where $\varepsilon_1 \in (0,\varepsilon)$. The terms $||b^i u_{x^i}||_{n,p}, ||cu||_{n,p}$ are considered similarly.

Actually, the above conditions on b, c, and ν can be considerably relaxed if in addition one applies deeper theorems about multiplyers from [17].

THEOREM 3.3. *Assume that for $m = 1,2,3,\ldots$, we are given a^{ij}_m, σ^i_m, f_m, g_m, and u_{0m} having the same sense as in Theorem 3.2 and verifying the same assumptions as a^{ij}, σ^i, f, g and u_0 with the same constants $\delta, K, \kappa_\varepsilon$, and K_ε. Let (h_k) be an orthonormal basis in H consisting of eigenvectors of $\mathbb{Q}$ and let P_m be the operator of orthogonal projection of H onto* Span $(h_1,\ldots,h_m)$. *Let $\zeta(x)$ be a real function of class C^∞_0 such that $\zeta(x) = 1$ if $|x| \leq 1$ and $\zeta(x) = 0$ if $|x| \geq 2$. Define $\zeta_k(x) = \zeta(x/k)$ and assume that for any $k = 1,2,3,\ldots, t \geq 0$, $\omega \in \Omega$,*

$$||\zeta_k\{a^{ij}(t,\cdot) - a^{ij}_m(t,\cdot)\}||_{n,p} + ||\zeta_k\{\sigma^i(t,\cdot) - \sigma^i_m(t,\cdot)\}||_{n+1,p} \to 0$$

as $m \to \infty$. Finally, let $E||u_{0m} - u_0||^p_{n+2,p} \to 0$ and

$$||(f(u,\cdot,\cdot), g(u,\cdot,\cdot)) - (f_m(u,\cdot,\cdot), g_m(u,\cdot,\cdot))||_{\mathcal{F}^n_p(\tau)} \to 0$$

whenever $u \in \mathcal{H}^{n+2}_p(\tau)$. Take the function u from Theorem 3.2 and for any m define $u_m \in \mathcal{H}^{n+2}_p(\tau)$ as the (unique) solution of the Cauchy problem for the equation

$$du_m(t,x) = [a^{ij}_m(t,x)u_{mx^ix^j}(t,x) + f_m(u_m,t,x)]\,dt$$

$$+ [\sigma^i_m(t,x)u_{mx^i}(t,x) + g_m(u_m,t,x)]\,dP_m w_t$$

on $[0,\tau]$ with the initial condition $u_m(0,\cdot) = u_{0m}$. Then $||u - u_m||_{\mathcal{H}^{n+2}_p(\tau)} \to 0$ as $m \to \infty$.

Proof. We have

$$d(u(t) - u_m(t)) = [a^{ij}_m(u - u_m)_{x^ix^j} + (a^{ij} - a^{ij}_m)u_{x^ix^j} + (f(u) - f_m(u_m))]\,dt$$

$$+ [\sigma^i_m P_m(u - u_m)_{x^i} + (\sigma^i - \sigma^i_m P_m)u_{x^i} + (g(u) - g_m(u_m)P_m)]\,dw_t,$$

$$f(u) - f_m(u_m) = [f_m(u) - f_m(u_m)] + [f(u) - f_m(u)],$$

$$g(u)-g_m(u_m)P_m=[g_m(u)-g_m(u_m)]P_m+g(u)[I-P_m]+[g(u)-g_m(u)]P_m,$$

$$\sigma^i-\sigma^i_mP_m=(\sigma^i-\sigma^i_m)P_m+\sigma^i(I-P_m).$$

Hence from our assumptions by Theorem 3.2 for any $t\geq 0$, we obtain

$$\|u-u_m\|_{\mathcal{H}^{n+2}_p(\tau\wedge t)}\leq\frac{1}{2}\|u-u_m\|_{\mathcal{H}^{n+2}_p(\tau\wedge t)}+N\|u-u_m\|_{\mathbb{H}^{n+1}_p(\tau\wedge t)}+NI_m,$$

where

$$I_m=(E\|u_0-u_{0m}\|^p_{n+2,p})^{1/p}+\|f(u)-f_m(u)\|_{\mathbb{H}^n_p(\tau)}+\|(a^{ij}-a^{ij}_m)u_{x^ix^j}\|_{\mathbb{H}^n_p(\tau)}$$

$$+\|(\sigma^i-\sigma^i_m)u_{x^i}\|_{\mathbb{H}^{n+1}_p(\tau,E)}+\|\sigma^i(I-P_m)u_{x^i}\|_{\mathbb{H}^{n+1}_p(\tau,E)}$$

$$+\|g(u)(I-P_m)\|_{\mathbb{H}^{n+1}_p(\tau,E)}+\|g(u)-g_m(u)\|_{\mathbb{H}^{n+1}_p(\tau,E)}.$$

We collect like terms in the last inequality and then apply Theorem 3.1. This yields

$$E\|u-u_m\|^p_{n+1,p}(\tau\wedge t)\leq N\|u-u_m\|^p_{\mathcal{H}^{n+2}_p(\tau\wedge t)}\leq NI^p_m+N\int_0^t E\|u-u_m\|^p_{n+1,p}(\tau\wedge s)\,ds,$$

where $t\leq T$ and N is independent of m. Gronwall's inequality allows us to drop the last term on the right. Next, we let m go to infinity and use our assumptions. Then we get

(3.8)
$$\limsup_{m\to\infty}\|u-u_m\|_{\mathcal{H}^{n+2}_p(\tau)}\leq N\limsup_{m\to\infty}I_m=N\limsup_{m\to\infty}\{\|(a^{ij}-a^{ij}_m)u_{x^ix^j}\|_{\mathbb{H}^n_p(\tau)}$$

$$+\|(\sigma^i-\sigma^i_m)u_{x^i}\|_{\mathbb{H}^{n+1}_p(\tau,E)}+\|\sigma^i(I-P_m)u_{x^i}\|_{\mathbb{H}^{n+1}_p(\tau,E)}+\|g(u)(I-P_m)\|_{\mathbb{H}^{n+1}_p(\tau,E)}\}.$$

Here, by the dominated convergence theorem,

$$\|g(u)(I-P_m)\|^p_{\mathbb{H}^{n+1}_p(\tau,E)}=E\int_0^T\int_{\mathbb{R}^d}\left[\sum_{k>m}((I-\Delta)^{(n+1)/2}g(u)\mathbb{Q}^{1/2},h_k)^2_H\right]^{p/2}dxdt\to 0,$$

and the same is true for $\|\sigma^i(I-P_m)u_{x^i}\|_{\mathbb{H}^{n+1}_p(\tau,E)}$ ($\sigma^iu_{x^i}\in\mathbb{H}^{n+1}_p(\tau,E)$ by Remark 3.3).

Next, by Lemma 3.1 for any $v\in C^\infty_0$ and k so large that $v\zeta_k=v$, we have

(3.9) $$\|(a^{ij}-a^{ij}_m)u_{x^ix^j}\|_{n,p}\leq N\|(u-v)_{x^ix^j}\|_{n,p}+\|(a^{ij}-a^{ij}_m)v_{x^ix^j}\|_{n,p},$$

$$\|(a^{ij}-a^{ij}_m)v_{x^ix^j}\|_{n,p}=\|\zeta_k(a^{ij}-a^{ij}_m)v_{x^ix^j}\|_{n,p}\leq N\|\zeta_k(a^{ij}-a^{ij}_m)\|_{n,p}\|v\|_{C^{|n|+2+\gamma}},$$

where the constants N do not depend on m and k. Thus,

$$\lim_{m\to\infty}\|(a^{ij}-a^{ij}_m)u_{x^ix^j}\|_{n,p}\leq N\|(u-v)_{x^ix^j}\|_{n,p},$$

and from the arbitrariness of v, we conclude that the left-hand side is zero for those ω and t for which $u \in H_p^{n+2}$. If we again apply Lemma 3.1, then we see that the pth power of the left-hand side of (3.9) is bounded by an integrable function. This and the dominated convergence theorem imply that

$$\lim_{m \to \infty} ||(a^{ij} - a^{ij}_m)u_{x^ix^j}||_{\mathbb{H}_p^n(\tau)} = 0.$$

Similar arguments take care of the remaining term in (3.8). The theorem is proven. □

COROLLARY 3.1. *Let Assumptions* 3.1–3.5 *be satisfied and take $f(u,t,x)$ and $g(u,t,x)$ satisfying the conditions from Remark* 3.6. *Also take the functions ζ_k from Lemma* 3.1 *and define*

$$(a_k, b_k, c_k, \sigma_k, \nu_k, f_k, g_k) = (a, b, c, \sigma, \nu, f, g) * \zeta_k.$$

Finally, define $u_k \in \mathcal{H}_p^{n+2}(T)$ as solutions of the Cauchy problem for the equations

$$du_k = (a_k^{ij} u_{kx^ix^j} + b_k^i u_{kx^i} + c_k u_k + f_k)\,dt + (\sigma_k^i u_{kx^i} + \nu_k u_k + g_k)\,dP_k w_t$$

with the initial conditions $u_k(0,\cdot) = u_0 * \zeta_k$. Then $||u - u_k||_{\mathcal{H}_p^{n+2}(\tau)} \to 0$.

To show this, it suffices to apply Lemma 3.1, repeat our argument about (3.9), and notice that

$$(\sigma_k^i, \sigma_k^j)_E \lambda^i \lambda^j = |(\sigma^i \lambda^i) * \zeta_k|_E^2 \le \zeta_k * [(\sigma^i, \sigma^j)_E \lambda^i \lambda^j].$$

COROLLARY 3.2. *Let Assumptions* 3.1–3.5 *be satisfied and also let them be satisfied for a $p = q$, where $q \ge 2$. Take $f(u,t,x)$ and $g(u,t,x)$ satisfying the conditions from Remark* 3.6. *Then the solution u from Theorem* 3.2 *belongs also to $\mathcal{H}_q^{n+2}(\tau)$.*

Without loss of generality, assume $p < q$ and let v be the solution of the same initial value problem but belonging to $\mathcal{H}_q^{n+2}(\tau)$ (such a unique v exists by Theorem 3.2). We have only to show that $v = u$. In light of Corollary 3.1, we can suppose that our assumptions are satisfied for any n. In this case, by Theorem 3.1,

$$E \sup_{t \le \tau} ||u||_{C^m}^p < \infty, \quad E \int_0^\tau ||u_{xx}(t,\cdot)||_{m,p}^p\,dt < \infty$$

for any $m = 1, 2, 3, \dots$. It follows that

$$\int_0^\tau ||u_{xx}(t,\cdot)||_{m,r}^r\,dt < \infty \quad \text{(a.s.)}$$

for any $r \ge p$. Take $m = 0$ and $r = q$ here and define

$$\tau_k = \tau \wedge \inf\left\{t : \int_0^t ||u_{xx}(s,\cdot)||_q^q\,ds \ge k\right\}.$$

Then obviously $u \in \mathcal{H}_q^2(\tau_k)$. Since v lies in the same class, by uniqueness $u(t,\cdot) = v(t,\cdot)$ for $t \le \tau_k$ (a.s.). It remains only to observe that $\tau_k \uparrow \tau$ when $k \to \infty$.

Our last result concerns the maximum principle. Its proof is absolutely standard. One needs only to apply Corollary 3.1 (along with statements (ii) and (iii) of Theorem 3.1) and the maximum principle from [14] or [11].

THEOREM 3.4 (maximum principle). *Let Assumptions* 3.1–3.5 *be satisfied and take $f(u,t,x)$ and $g(u,t,x)$ satisfying the conditions from Remark* 3.6. *Suppose that for any ω and t we have $u_0 \ge 0$, $f(t,\cdot) \ge 0$ (in the sense of distributions), and $g(t,\cdot) \equiv 0$. Then the solution u of the Cauchy problem for the* linear *equation* (3.7) *with the initial condition $u(0,\cdot) = u_0$ verifies $u(t,\cdot) \ge 0$ for all $t \in [0,\tau]$ almost surely.*

4. An application. In this section, we consider the one-dimensional equation with space–time white noise. Thus $d = 1$. Fix a number $\kappa \in (0, 1/2]$ and define $n = -\kappa - 3/2$.

4.1. Let $a(t, x) = a(\omega, t, x)$ and $b(t, x) = b(\omega, t, x)$ be real-valued functions defined on $\Omega \times \mathbb{R}_+ \times \mathbb{R}$. Assume that

(i) for any ω and t, the function a is twice continuously differentiable with respect to x and the function b is once continuously differentiable with respect to x and $||a||_{C^2} + ||b||_{C^1} \leq K$, $K \geq a \geq \delta$;

(ii) for any $x \in \mathbb{R}$, the processes a and b are predictable.

Suppose also that we are given measurable real functions $f(t, x, u)$ and $g(t, x, u)$ on $\Omega \times \mathbb{R}_+ \times \mathbb{R}^2$ such that

(iii) for any x and u, the processes $f(t, x, u)$ and $g(t, x, u)$ are predictable;

(iv) for any ω, t, and x, the functions $f(t, x, u)$ and $g(t, x, u)$ satisfy the Lipschitz condition with the constant K with respect to u;

(v) $E\int_0^T\{||f(t,\cdot,0)||_p^p + ||g(t,\cdot,0)||_p^p\}\,dt < \infty$ for all $T < \infty$.

We also take an $\mathcal{F}_t$-adapted Wiener process B_t in L_2 with unit covariance operator (the so-called cylindrical Wiener process) and consider the equation

(4.1)
$$du(t,x) = [a(t,x)u''(t,x) + b(t,x)u'(t,x) + f(t,x,u(t,x))]\,dt + g(t,x,u(t,x))\,dB_t,$$

which as always is understood in the sense of distributions. Specifically, by the solution of this equation we mean a real-valued function $u(\omega, t, x)$ of class $\mathfrak{D}$ such that for any $\phi \in C_0^\infty$, we have $u\phi \in L_2([0,T], L_2)$, $T > 0$ almost surely, the process $(u(t,\cdot), \phi)$ is continuous, and

$$(u(t,\cdot),\phi) = (u(0,\cdot),\phi) + \int_0^t ((u(s,\cdot),(a(s,\cdot)\phi(\cdot))'' - (b(s,\cdot)\phi(\cdot))' + f(s,\cdot,u(s,\cdot)))\,ds$$
$$+ \int_0^t (g(s,\cdot,u(s,\cdot))\phi(\cdot), dB_s),$$

for all t almost surely, where the last stochastic integral is a stochastic integral of the L_2-valued function $g(t,\cdot,u(t,\cdot))\phi(\cdot)$ with respect to B_t.

In order to apply Theorem 3.2, we have to find the corresponding objects H, E, and w_t and rewrite the equation in our terms. First of all, we make a standard imbedding of B_t into a Hilbert space. To this end, we take an orthonormal basis $(\psi_k, k \geq 1)$ in L_2 and define H to be the space of all formal series $h = \sum_k h^k\psi_k$, where h^k are real numbers such that $|h|_H^2 := \sum_k k^{-20}(h^k)^2 < \infty$. Observe that the functions $\tilde{\psi}_k := k^{10}\psi_k$ form an orthonormal basis in H. Next, let

$$B_t^k := (B_t, \psi_k), \quad w_t := \sum_{k=1}^{\infty} \psi_k B_t^k.$$

Since $E\sum_k k^{-20}(\psi_k, B_t)^2 = t\sum_k k^{-20} < \infty$, the process w_t takes values in H. Its covariance operator $\mathbb{Q}$ can be found from the formula

$$t(h_1, \mathbb{Q}h_2)_H = E(h_1, w_t)_H(h_2, w_t)_H$$
$$= E\left(\sum_{k=1}^{\infty} k^{-20}h_1^k B_t^k\right)\left(\sum_{k=1}^{\infty} k^{-20}h_2^k B_t^k\right) = t\sum_{k=1}^{\infty} k^{-40}h_1^k h_2^k.$$

Hence

$$\mathbb{Q}h = \sum_{k=1}^{\infty} \frac{1}{k^{20}} h^k \psi_k, \quad \mathbb{Q}^{1/2}h = \sum_{k=1}^{\infty} \frac{1}{k^{10}} h^k \psi_k, \quad \mathbb{Q}^{1/2}\tilde{\psi}_k = \psi_k.$$

If $h \in H$, then $\mathbb{Q}^{1/2}h$ (as an element of H) is represented by a formal series which converges in L_2. Therefore, we sometimes identify $\mathbb{Q}^{1/2}h$ with the corresponding element of L_2. In the following lemma, we also set

$$R(x) = \frac{1}{2\sqrt{\pi}} |x|^{-(1-2\kappa)/2} \int_0^\infty t^{-(5-2\kappa)/4} e^{-tx^2 - 1/(4t)}\, dt.$$

Observe that R is infinitely differentiable outside the origin, decreases exponentially fast as $|x| \to \infty$, and behaves near the origin like $|x|^{-(1-2\kappa)/2}$ if $\kappa < (1/2)$ and $-\log|x|$ if $\kappa = 1/2$.

LEMMA 4.1. (i) *Let $f \in L_{2,\mathrm{loc}}$. For $\phi \in C_0^\infty$ define (Gf, ϕ) as a function on L_2 by the formula*

$$(Gf, \phi)(h) = (f\phi, h)_{L_2} = \int_{\mathbb{R}} fh\phi\, dx.$$

Then $G : L_{2,\mathrm{loc}} \to \mathcal{D}(E)$ and $|(Gf, \phi)|_E = ||f\phi||_2$.

(ii) *If $f \in L_p$, then $Gf \in H_p^{n+1}(E)$ and*

(4.2)

$$||Gf||_{n+1,p}^2 = \left(\int_{\mathbb{R}} \left\{ \int_{\mathbb{R}} R^2(y) f^2(x-y)\, dy \right\}^{p/2} dx \right)^{2/p} \le (N||f||_p^2) \wedge (||f||_2^2 \cdot ||R||_p^2),$$

where $N = ||R||_2^2 < \infty$, and if $p(1-2\kappa) > 2$, then

$$||Gf||_{n+1,p} \le N(\kappa) ||f||_2^{2\kappa p/(p-2)} ||f||_p^{1-2\kappa p/(p-2)}. \tag{4.3}$$

(iii) *If $f(t)$ is an $L_{2,\mathrm{loc}}$-valued predictable function such that*

$$\int_0^T \sup_y ||f(t) I_{|\cdot - y| \le R}||_2^2\, dt < \infty \quad \forall R, T \in (0, \infty),$$

then $Gf(\cdot) \in \mathfrak{D}(E)$ and for any $T > 0, \phi \in C_0^\infty$, we have

$$\int_0^T (\phi f(t), dB_t) = \int_0^T (Gf(t), \phi)\, dw_t \quad (a.s.). \tag{4.4}$$

Proof. (i) As we have explained above, functions on L_2 can be considered as functions on $\mathbb{Q}^{1/2}H$; therefore, linear functions on L_2 are linear functions on $\mathbb{Q}^{1/2}H$. Next,

$$|(Gf, \phi)|_E = |(Gf, \phi) \circ \mathbb{Q}^{1/2}|_H = \sup_{|h|_H = 1} (Gf, \phi)(\mathbb{Q}^{1/2}h) = \sup_{|h|_H=1} \sum_k \frac{1}{k^{10}} h^k \int_{\mathbb{R}} f\phi\psi_k\, dx,$$

and the last expression equals $||f\phi||_2$. In particular, this implies that $Gf \in \mathcal{D}(E)$.

(iii) The relation $Gf(\cdot) \in \mathfrak{D}(E)$ follows from (i). By one of definitions, in the mean-square sense, we have

(4.5)

$$\int_0^T (Gf(t),\phi)\,dw_t = \sum_k \int_0^T (Gf,\phi)(\tilde\psi_k)\,d(w_t,\tilde\psi_k)_H = \sum_k \int_0^T (Gf,\phi)(\tilde\psi_k)k^{-10}\,dB_t^k$$
$$= \sum_k \int_0^T (Gf,\phi)(\psi_k)\,dB_t^k = \sum_k \int_0^T (f(t)\phi,\psi_k)_{L_2}\,dB_t^k.$$

It remains only to observe that the last expression coincides with the left-hand side of (4.4) (a.s.).

(ii) We know that for any $\phi \in C_0^\infty$, we have

$$(I-\Delta)^{(n+1)/2}\phi(x) = (I-\Delta)^{-(1+2\kappa)/4}\phi(x) = \int_{\mathbb{R}} R(x-y)\phi(y)\,dy.$$

This implies that for any $h \in L_2$, $\phi \in C_0^\infty$,

$$((I-\Delta)^{(n+1)/2}Gf,\phi)(h) = (R*Gf,\phi)(h) := (Gf,R*\phi)(h)$$
$$= \int_{\mathbb{R}} f(R*\phi)h\,dx = \int_{\mathbb{R}} \phi(R*(fh))\,dx,$$

and $(I-\Delta)^{(n+1)/2}Gf$ is a usual function on $\mathbb{R}$ with values in E defined by

$$(I-\Delta)^{(n+1)/2}Gf(x)(h) = R*(fh)(x) = (Gf,R(x-\cdot))(h).$$

By (i) (or as in (i)), it follows that

$$|(I-\Delta)^{(n+1)/2}Gf(x)|_E = ||fR(x-\cdot)||_2.$$

We thus get the equality in (4.2). The inequality in (4.2) follows from the Minkowski inequality. To prove (4.3), we use that $R^2(y) \le N|y|^{2\kappa-1}$ and we minimize with respect to $\varepsilon > 0$ after the following computations:

$$\left(\int_{\mathbb{R}}\left\{\int_{\mathbb{R}} R^2(y)f^2(x-y)\,dy\right\}^{p/2} dx\right)^{2/p} \le \left(\int_{\mathbb{R}}\left\{\int_{\mathbb{R}} I_{|y|\le\varepsilon}R^2(y)f^2(x-y)\,dy\right\}^{p/2} dx\right)^{2/p}$$
$$+\left(\int_{\mathbb{R}}\left\{\int_{\mathbb{R}} I_{|y|\ge\varepsilon}R^2(y)f^2(x-y)\,dy\right\}^{p/2} dx\right)^{2/p}$$
$$\le ||I_{|y|\le\varepsilon}R^2||_1||f||_p^2 + ||I_{|y|\ge\varepsilon}R^p||_1^{2/p}||f||_2^2$$
$$\le N\varepsilon^{2\kappa}||f||_p^2 + N\varepsilon^{-(1-2\kappa-2/p)}||f||_2^2.$$

The lemma is proven. □

Remark 4.1. The method of converting integrals with respect to B_t into integrals with respect to w_t can be generalized. In [6] a similar construction is used in order to reduce general (not necessarily continuous) stochastic integrals with respect to *martingale measures* to integrals with respect to Hilbert-space-valued processes.

THEOREM 4.1. *Take* $\kappa \in (0,1/2)$ *and* $u_0 \in L_p(\Omega,\mathcal{F}_0,H_p^{(1/2)-\kappa})$. *Then for any* T, *in the space* $\mathcal{H}_p^{(1/2)-\kappa}(T)$, *equation* (4.1) *with the initial condition* u_0 *has a unique solution* u. *Moreover,*

$$||u||_{\mathcal{H}_p^{1/2-\kappa}(T)} \le N(||f(\cdot,\cdot,0)||_{\mathbb{H}_p^{-3/2-\kappa}(T)} + ||g(\cdot,\cdot,0)||_{\mathbb{L}_p(T)} + (E||u_0||^p_{1/2-\kappa,p})^{1/p}),$$

where the constant N depends only on κ, p, δ, K, and T.

Proof. By Lemma 4.1 (iii) for functions from $\mathcal{H}_p^{1/2-\kappa}(T)$, equation (4.1) takes the form

$$du = [au'' + bu' + f(u)]\,dt + \bar{g}(u)\,dw_t,$$

where $\bar{g}(u)(t,x) := G(g(t,\cdot,u(t,\cdot)))(x)$. We will apply Theorem 3.2 to this equation. Its assumptions concerning a and σ are obviously satisfied. Next, if $u \in H_p^{n+2}$, then $bu' \in H_p^{n+1} \subset H_p^n$, $u \in L_p$, $f(u) \in L_p \subset H_p^n$. Furthermore, by Lemma 3.1 (cf. Remark 3.6),

$$||bu'||_{n,p} \le ||bu'||_{-1,p} \le N||u'||_{-1,p} \le N||u||_p = N||u||_{n+2-((1/2)-\kappa),p},$$

$$||f(u) - f(v)||_{n,p} \le ||f(u) - f(v)||_p \le K||u - v||_p.$$

Consequently (see Remark 3.5), the assumptions of Theorem 3.2 concerning $bu' + f(u)$ are satisfied. To check the remaining assumptions about $\bar{g}(u)$, it suffices to notice that by Lemma 4.1 (ii) we have

$$||\bar{g}(0)(t,\cdot)||_{n+1,p} \le N||g(t,\cdot,0)||_p,$$

$$||\bar{g}(u)(t,\cdot) - \bar{g}(v)(t,\cdot)||_{n+1,p} \le ||\bar{g}(u)(t,\cdot) - \bar{g}(v)(t,\cdot)||_p \le K||u(t,\cdot) - v(t,\cdot)||_p.$$

The theorem is proven. □

Remark 4.2. In order to apply the approximation theorem (Theorem 3.3) or Theorem 3.4 on the maximum principle, it is useful to notice that if instead of w_t we take its approximations by $P_m w_t$ as in Theorem 3.3, then the corresponding equation becomes

$$du(t,x) = [a(t,x)u''(t,x) + b(t,x)u'(t,x) + f(t,x,u(t,x))]\,dt + \sum_{k=1}^{m} g(t,x,u(t,x))\,dB_t^k.$$

This follows immediately from (4.5).

Remark 4.3. Additional information about Hölder continuity properties of the solution is readily obtained from the properties of elements of $\mathcal{H}$ listed in Theorem 3.1.

Remark 4.4. We could take a noise even "whiter" than B_t. Indeed it is not hard to see that B_t above can be cylindrical Wiener process in $H_2^{-\varepsilon}$ with unit covariance operator, where $\varepsilon \in (0, 1/2)$.

4.2. Take a, b, and g as in §4.1, and take a bounded real-valued $\mathcal{P} \times \mathcal{B}(\mathbb{R})$-measurable function $c(t,x) = c(\omega,t,x)$. Assume that $g(t,x,u) = 0$ for $u \le 0$, fix a number $\lambda \in [0, 1/2)$, and define $h(t,x,u) = g(t,x,u)u^\lambda$.

Theorem 4.1 can be easily applied to prove that the equation

$$du = (au'' + bu' + cu)\,dt + h(u)\,dB_t \tag{4.6}$$

has a unique solution defined for all t if the initial condition u_0 is nonnegative and, say, is nonrandom and belongs to C_0^∞. Furthermore, $\sup_{t \le T, x} |u(t,x)|$ is finite (a.s.) for any $T < \infty$. These facts for equation (4.6) considered on a finite space interval with $a \equiv 1$, $b \equiv c \equiv 0$, and $h(u) = u_+^{1+\lambda}$ and with zero boundary data were discovered in [15] with the help of a quite different approach. By using the maximum principle, it is not hard to show that our assertion implies the result of [15].

To begin our investigation of (4.6), let us fix $T \in (0, \infty)$ and take $\kappa \in (0, 1/2)$ and $p > 6$ such that

$$\frac{1}{2} - \kappa - \frac{3}{p} > 0, \quad \eta := 1 + \lambda - \kappa - \frac{1}{p} < 1.$$

By Theorem 4.1 for any $m = 1, 2, 3 \dots$, the equation

$$du_m = (au_m'' + bu_m' + cu_m)\,dt + h(u_m \wedge m)\,dB_t \tag{4.7}$$

with the initial condition $u_m(0, \cdot) = u_0$ has a unique solution $u_m \in \mathcal{H}_p^{1/2-\kappa}(T)$. By Theorem 3.1 (iii) and since $H_p^r \subset C^{r-d/p}$ whenever $r - d/p > 0$, we have that $u_m \in C([0,T] \times \mathbb{R})$ (a.s.) and $E||u_m||^p_{C([0,T]\times\mathbb{R})}$ is finite. We need only to show that for a constant $r > 0$ the expression $E\sup_{t\leq T,x}|u_m(t,x)|^r$ is bounded by a constant independent of m. Indeed, then, for large m with probability as close to 1 as we like, the function u_m satisfies (4.6) on $[0, T]$.

From Theorem 3.4 on the maximum principle (also use that $h(u_m \wedge m) = u_m\nu$, where ν is a bounded function), we get $u_m \geq 0$. Next take $\zeta_k(x)$ from Theorem 3.3, multiply (4.7) by $\zeta_k e^{-Kt}$, where $K = \sup(|a''| + |b'| + |c|)$, integrate by parts (that is, use the definition of solutions), and take expectations. Then for any stopping time $\tau \leq T$, we obtain

$$\begin{aligned}
e^{-KT}E(\zeta_k, u_m(\tau,\cdot)) &\leq E(\zeta_k, u_m(\tau,\cdot))e^{-K\tau} \\
&= (\zeta_k, u_0) + E\int_0^\tau (a\zeta_k'' + (2a' - b)\zeta_k' + (a'' - b' + c - K)\zeta_k, u_m)e^{-Kt}\,dt \\
&\leq N + NE\int_0^T (|\zeta_k''| + |\zeta_k'|, u_m)\,dt \leq N + \frac{N}{k}k^{1-1/p}E\int_0^T ||u_m(t,\cdot)||_p\,dt,
\end{aligned}$$

$$E(\zeta_k, u_m(\tau,\cdot)) \leq N + Nk^{-1/p}\left(E\int_0^T ||u_m(t,\cdot)||_p^p\,dt\right)^{1/p} \leq N + Mk^{-1/p},$$

where the last constant N is independent of m, k, and τ and M is independent of k. Since this inequality is true for any stopping time $\tau \leq T$, with the same N and M for any number $r \in (0, 1)$,

$$E\sup_{t\leq T}\left(\int_{\mathbb{R}} \zeta_k(x)u_m(t,x)\,dx\right)^r \leq \frac{N + Mk^{-1/p}}{1-r}, \quad E\sup_{t\leq T}\left(\int_{\mathbb{R}} u_m(t,x)\,dx\right)^r \leq \frac{N}{1-r},$$

where the latter relation is obtained from the former one by the monotone convergence theorem.

After this, for $\alpha := 2\kappa p/(p-2)$ and for any stopping time $\tau \leq T$, by Theorems 3.1 (iii) and 3.2 and Lemma 4.1 (ii), we obtain

$$\begin{aligned}
E||u_m||^p_{C([0,\tau]\times\mathbb{R})} &\leq N\left(1 + E\sup_{t\leq\tau}||Gh(u_m \wedge m)(t,\cdot)||^p_{n+1,p}\right) \\
&\leq N + NE\sup_{t\leq\tau}\{||h(u_m)(t,\cdot)||_2^{p\alpha}||h(u_m)(t,\cdot)||_p^{p(1-\alpha)}\} \\
&\leq N + NE\sup_{t\leq\tau}\{||u_m^{1+\lambda}(t,\cdot)||_2^{p\alpha}||u_m^{1+\lambda}(t,\cdot)||_p^{p(1-\alpha)}\},
\end{aligned}$$

where the constants N are independent of τ and m. Since this is true for any $\tau \leq T$, by the well-known martingale inequality (see, for instance, Lemma 1.1 in [8]), for any stopping time $\tau \leq T$ and any number $\beta \in (0,1)$, it also holds that

$$E||u_m||^{p\beta}_{C([0,\tau]\times\mathbb{R})} \leq N + NE\sup_{t\leq\tau}\{||u^{1+\lambda}(t,\cdot)||_2^{p\alpha\beta}||u^{1+\lambda}(t,\cdot)||_p^{p(1-\alpha)\beta}\}.$$

Take $\beta = (1-\eta)(\kappa p+1)^{-1}$ and use the simple relations

$$||u^{\mu}||_p \leq \sup|u|^{\mu-1/p}||u||_1^{1/p}, \quad \left(\lambda+\frac{1}{2}\right)\alpha+\left(\lambda+1-\frac{1}{p}\right)(1-\alpha) = \eta, \quad \frac{p\alpha}{2}+1-\alpha = \kappa p+1.$$

Then it is seen that

$$E||u_m||^{p\beta}_{C([0,\tau]\times\mathbb{R})} \leq N + NE||u||^{p\beta\eta}_{C([0,\tau]\times\mathbb{R})}\sup_{t\leq\tau}||u(t,\cdot)||_1^{1-\eta}$$

$$\leq N + N(E||u_m||^{p\beta}_{C([0,\tau]\times\mathbb{R})})^{\eta}(E\sup_{t\leq T}||u(t,\cdot)||_1)^{1-\eta} \leq N + N(E||u_m||^{p\beta}_{C([0,\tau]\times\mathbb{R})})^{\eta}.$$

We emphasize that all constants N are independent of m and τ. Therefore, the last inequality implies the desired absolute estimate of $E||u_m||^{p\beta}_{C([0,T]\times\mathbb{R})}$.

Remark 4.5. If we first let $p \to \infty$ and then let $\kappa \uparrow 1/2$, then we see that $p\beta$ can be made as close to $1-2\lambda$ as we wish.

5. Proofs of Theorems 3.1 and 3.2.

In this section, we prove our main results.

Proof of Theorem 3.1. We will prove assertions (ii)–(vi) only when $u \in \mathcal{H}^n_{p,0}(T)$. The reader can easily obtain these assertions in the general case by considering $u(t,x)-u(0,x)$ instead of u and making obvious modifications in our arguments.

Take $u \in \mathcal{H}^n_{p,0}(\tau)$, $\bar{f}$, and g such that (2.3) is satisfied with $\bar{f}$ and g instead of f and g. Notice that on $[0,\tau]$, the function u satisfies the equation

$$du = (\Delta u + f)\,dt + g\,dw_t, \tag{5.1}$$

where $f = \bar{f} - \Delta u \in \mathbb{H}^{n-2}_p(\tau)$. By Theorem 2.1, the equation

$$dv = (\Delta v + fI_{t\leq\tau})\,dt + gI_{t\leq\tau}\,dw_t$$

on $[0,T]$ with zero initial condition has a unique solution $v \in \mathcal{H}^n_{p,0}(T)$. The difference $u-v$ satisfies the heat equation on $[0,\tau]$ with zero initial condition. It follows that $u(t,\cdot) = v(t,\cdot)$ on $[0,\tau]$, and by Theorem 2.1 (ii),

$$||v||_{\mathcal{H}^n_p(T)} \leq N(d,n,p)(||f||_{\mathbb{H}^{n-2}_p(\tau)} + ||g||_{\mathbb{H}^{n-1}_p(\tau,E)}) \leq N(d,n,p)||u||_{\mathcal{H}^n_p(\tau)},$$

$$||f||_{\mathbb{H}^{n-2}_p(\tau)} + ||g||_{\mathbb{H}^{n-1}_p(\tau,E)} \leq N(d,n,p)||v||_{\mathcal{H}^n_p(\tau)}$$

The former inequality shows that we can confine ourselves to the case when $\tau \equiv T$. The latter one allows us to consider solutions of equations (5.1) only and to prove our assertions with $||(f,g)||_{\mathcal{F}^{n-2}_p(T)}$ in place of $||u||_{\mathcal{H}^n_p(T)}$. Therefore, below we take $\tau = T$ and take the function $u \in \mathcal{H}^n_{p,0}(T)$ as a solution of (5.1), and we notice at once that assertion (ii) immediately follows from Theorem 2.1.

(iii) We can and will suppose that $n = 2\beta$. As in Lemma 2.4, it suffices to consider the case of f and g as in (2.13). This will justify our later computations.

By one of imbedding theorems for the Slobodetsky spaces (see, for instance, [17] or [3]), for any continuous L_p-valued function $h(t)$ and $s \le t$, we have

$$||h(t) - h(s)||_p^p \le N(\alpha,p)(t-s)^{\alpha p-1} \int_s^t \int_s^t I_{r_2 > r_1} \frac{||h(r_2) - h(r_1)||_p^p}{|r_2 - r_1|^{1+\alpha p}}\, dr_1 dr_2$$

$$= N(\alpha,p)(t-s)^{\alpha p-1} \int_0^{t-s} \frac{dy}{y^{1+\alpha p}} \int_s^{t-y} ||h(r+y) - h(r)||_p^p\, dr. \tag{5.2}$$

Actually, the space L_p here can be replaced by any Banach space. We will also need Theorem 14.11 from [5], which implies that for any $h \in L_p$, $\theta \in (0,1]$, and $y \in [0,T]$,

$$\begin{aligned} ||e^{-y/2}T_y h||_p &= ||(1-\Delta)^{1-\theta} e^{-y/2} T_y[(1-\Delta)^{-(1-\theta)}h]||_p \\ &\le N(d,p,\theta)\frac{1}{y^{1-\theta}}||(I-\Delta)^{-(1-\theta)}h||_p, \\ ||(T_y - I)h||_p &\le \int_0^y |||[\Delta(I-\Delta)^{-1}](I-\Delta)^{1-\theta}T_v[(1-\Delta)^{\theta}h]|||_p\, dv \\ &\le N\int_0^y ||(I-\Delta)^{1-\theta}T_v[(1-\Delta)^{\theta}h]||_p\, dv \\ &\le N(d,p,\theta,T)\int_0^y v^{\theta-1}\, dv||h||_{2\theta,p} = Ny^{\theta}||h||_{2\theta,p}. \end{aligned} \tag{5.3}$$

We apply (5.2) to u_1 and $u_2 := u - u_1$, where u_1 is introduced in Lemma 2.4. In order to avoid repetitions of some arguments, we denote $f^{(1)} = f$, $f^{(2)} = g$, $dw_t^{(1)} = dt$, and $dw_t^{(2)} = dw_t$. It is easy to see that

$$u_i(r+y) - u_i(r) = (T_y - I)u_i(r) + \int_0^y T_{y-v}f^{(i)}(r+v)\, dw_{r+v}^{(i)}.$$

Therefore,

$$E||u_i(r+y) - u_i(r)||_p^p \le N(A_i(r,y) + B_i(r,y)),$$

$$E||u_i(t) - u_i(s)||_p^p \le N(t-s)^{\alpha p-1}(I_i(t,s) + J_i(t,s)),$$

where

$$A_i(r,y) = E||(T_y - I)u_i(r)||_p^p, \quad B_i(r,y) = E\left\|\int_0^y T_{y-v}f^{(i)}(r+v)\, dw_{r+v}^{(i)}\right\|_p^p,$$

$$I_i(t,s) = \int_0^{t-s} \frac{dy}{y^{1+\alpha p}} \int_s^{t-y} A_i(r,y)\, dr, \quad J_i(t,s) = \int_0^{t-s} \frac{dy}{y^{1+\alpha p}} \int_s^{t-y} B_i(r,y)\, dr.$$

By using the Burkholder–Davis–Gundy inequality, the Hölder inequality, and (5.3), we get

$$\begin{aligned} B_2(r,y) &\le NE\int_{\mathbb{R}^d}\left[\int_0^y v^{2\beta-1}v^{1-2\beta}|T_v g(r+y-v)|_E^2\, dv\right]^{p/2} dx \\ &\le Ny^{\beta p-1}E\int_0^y ||(I-\Delta)^{\beta-1/2}g(r+y-v)||_p^p\, dv, \end{aligned}$$

$$
\begin{aligned}
J_2 &\le NE\int_0^{t-s}\frac{dy}{y^{2-(\beta-\alpha)p}}\int_s^{t-y}dr\int_0^y ||g(r+y-v)||^p_{n-1,p}\,dv\\
&\le NE\int_0^{t-s}\frac{dy}{y^{2-(\beta-\alpha)p}}\int_0^y dv\int_s^t ||g(r)||^p_{n-1,p}\,dr\\
&= N(t-s)^{(\beta-\alpha)p}E\int_s^t ||g(r)||^p_{n-1,p}\,dr,
\end{aligned}
$$

$$
\begin{aligned}
B_1(r,y) &= E\int_{\mathbb{R}^d}\left|\int_0^y v^{\beta-1}v^{1-\beta}T_v f(r+y-v)\,dv\right|^p\,dx\\
&\le Ny^{\beta p-1}E\int_0^y ||f(r+y-v)||^p_{n-2,p}\,dv,\\
J_1(t,s) &\le N(t-s)^{(\beta-\alpha)p}E\int_s^t ||f(r)||^p_{n-2,p}\,dr.
\end{aligned}
$$

Finally, by again using (5.3) and results from [9] and [12], we conclude

$$
\begin{aligned}
I_2(t,s) &\le NE\int_0^{t-s}\frac{dy}{y^{1-(\beta-\alpha)p}}\int_s^{t-y} ||u_2(r)||^p_{2\beta,p}\,dr\\
&\le N(t-s)^{(\beta-\alpha)p}E\int_0^t ||g(r)||^p_{2\beta-1,p}\,dr,\\
I_1(t,s) &\le N(t-s)^{(\beta-\alpha)p}E\int_0^t ||f(r)||^p_{2\beta-2,p}\,dr.
\end{aligned}
$$

Collecting all these estimates, we get (3.2) at least for $\eta = T\,(=\tau)$. In the general case, it suffices to notice that instead of points t and s on the left in estimate (5.2), one can obviously take any two points between them.

The proof of (3.3) goes exactly the same way, the only difference being that this time we use the following consequence of (5.2):

$$
\sup_{0\le s<t\le T}\frac{||h(t)-h(s)||^p_p}{(t-s)^{\alpha p-1}} \le N(\alpha,p)\int_0^T\int_0^T I_{r_2>r_1}\frac{||h(r_2)-h(r_1)||^p_p}{|r_2-r_1|^{1+\alpha p}}\,dr_1dr_2.
$$

Assertions (ii) and (iii) imply (i) almost automatically. They also imply (iv) and (v) in view of well-known imbedding theorems saying that under conditions in (iv) and (v), we have $H^n_p \subset C^\alpha$ and $H^n_p \subset H^m_q$, respectively (see, for instance, Remarks 2.7.1/2 and 2.7.1/3 in [17]). From the same imbedding theorems and from the interpolation theorem (Theorem 2.4.2) in [17], we have

$$
||u||_{m(\theta)-d/p+d/q,q} \le N(d,p,q,m(0),m(1),\theta)||u||^{1-\theta}_{m(0),p}||u||^{\theta}_{m(1),p}
$$

whenever $1 < p < q < \infty$, $\theta \in (0,1)$, $m(\theta) := (1-\theta)m(0) + \theta m(1) \ne m(0)$, and $u \in H^n_p$. Theorem 14.2 from [5] shows that the case $p = q$ actually needs not to be excluded. Note also that under the conditions in (vi), there is a β such that

$1/2 > \beta > 1/p$ and $m \le k := n - 2\beta(1-\theta) - d/p + d/q$. Therefore,

$$E\left(\int_0^T ||u(t,\cdot)||_{m,q}^{p/\theta}\,dt\right)^\theta \le E\left(\int_0^T ||u(t,\cdot)||_{k,q}^{p/\theta}\,dt\right)^\theta$$
$$\le NE\left(\int_0^T ||u(t,\cdot)||_{n-2\beta,p}^{(1-\theta)p/\theta}||u(t,\cdot)||_{n,p}^{p}\,dt\right)^\theta$$
$$\le NE\sup_{t\le T}||u(t,\cdot)||_{n-2\beta,p}^{(1-\theta)p}\left(\int_0^T ||u(t,\cdot)||_{n,p}^{p}\,dt\right)^\theta .$$

To prove (vi) it remains only to apply the Hölder inequality and assertion (iii). The theorem is proven. □

To prove Theorem 3.2, we need some auxiliary constructions.

DEFINITION 5.1. *Assume that for $\omega \in \Omega$ and $t \ge 0$, we are given operators*

$$L(t,\cdot) : H_p^{n+2} \to H_p^n, \quad \Lambda(t,\cdot) : H_p^{n+2} \to H_p^{n+1}(\mathbb{R}^d, E).$$

Assume that

(i) *for any ω and t, the operators $L(t,u)$ and $\Lambda(t,u)$ are continuous (with respect to u);*

(ii) *for any $u \in H_p^{n+2}$, the processes $L(t,u)$ and $\Lambda(t,u)$ are predictable;*

(iii) *for any $\omega \in \Omega$, $t \ge 0$, and $u \in H_p^{n+2}$, we have*

$$||L(t,u)||_{n,p} + ||\Lambda(t,u)||_{n+1,p} \le N_{L,\Lambda}(1 + ||u||_{n+2,p}),$$

where $N_{L,\Lambda}$ is a constant.

Then for a function $u \in \mathcal{H}_p^{n+2}(T)$, we write

$$(L,\Lambda)u = -(f,g)$$

if $(f,g) \in \mathcal{F}_p^n(T)$, and in the sense of distributions for $t \in [0,T]$, we have that

$$u(t) = u(0) + \int_0^t (L(s,u(s)) + f(s))\,ds + \int_0^t (\Lambda(s,u(s)) + g(s))\,dw_s \quad \text{(a.s.)}.$$

Remark 5.1. By virtue of our conditions on L and Λ, for any $u \in \mathcal{H}_p^{n+2}(T)$, we have $(L(u),\Lambda(u)) \in \mathcal{F}_p^n(T)$. It immediately follows that the operator (L,Λ) is well defined on $\mathcal{H}_p^{n+2}(T)$, and

$$||(L,\Lambda)u||_{\mathcal{F}_p^n(T)} \le (1 + N_{L,\Lambda})||u||_{\mathcal{H}_p^{n+2}(T)} + N_{L,\Lambda}T^{1/p}.$$

Observe that $||u_x||_{n,p} \le N||(I-\Delta)^{1/2}u||_{n,p}$ for any $u \in H_p^{n+1}$. This shows that, in terms of Definition 5.1, Theorem 2.1 has the following version.

THEOREM 5.1. *Let a and σ satisfy the assumptions from §2. Define*

$$Lu = a^{ij}u_{x^ix^j}, \quad \Lambda u = \sigma^i u_{x^i}.$$

Then the operator (L,Λ) is a one-to-one operator from $\mathcal{H}_{p,0}^{n+2}(T)$ onto $\mathcal{F}_p^n(T)$ and the norm of its inverse is less than a constant depending only on d, p, δ, and K (thus independent of T).

Next we prove a perturbation result. It needs a proof because we do not allow ε to depend on T.

THEOREM 5.2. *Take the operators L and Λ from Theorem 5.1, and let some operators L_1 and Λ_1 satisfy the requirements from Definition 5.1. We assert that there exists a constant $\varepsilon \in (0,1)$ depending only on d, p, δ, and K such that if for a constant K_1 and any $u, v \in H_p^{n+2}$, $t \geq 0$, $\omega \in \Omega$, we have*

$$||L_1(t,u) - L_1(t,v)||_{n,p} + ||\Lambda_1(t,u) - \Lambda_1(t,v)||_{n+1,p}$$

$$\leq \varepsilon||u_{xx} - v_{xx}||_{n,p} + K_1||u - v||_{n+1,p}, \tag{5.4}$$

then for any $(f,g) \in \mathcal{F}_p^n(T)$, there exists a unique solution $u \in \mathcal{H}_{p,0}^{n+2}(T)$ of the equation

$$(L + L_1, \Lambda + \Lambda_1)u = -(f,g).$$

Furthermore, for this solution u, we have

$$||u||_{\mathcal{H}_p^{n+2}(T)} \leq N||(L_1(\cdot,0) + f, \Lambda_1(\cdot,0) + g)||_{\mathcal{F}_p^n(T)},$$

where N depends only on d, p, δ, K, K_1, and T and N is independent of T if $K_1 = 0$.

Proof. Take $u \in \mathcal{H}_{p,0}^{n+2}(T)$, observe that $(L_1(u), \Lambda_1(u)) \in \mathcal{F}_p^n(T)$, and by using Theorem 5.1, define $v \in \mathcal{H}_{p,0}^{n+2}(T)$ as the unique solution of the equation $(L,\Lambda)v = -(f + L_1(u), g + \Lambda_1(u))$. By denoting $v = Ru$, we define an operator $R : \mathcal{H}_{p,0}^{n+2}(T) \to \mathcal{H}_{p,0}^{n+2}(T)$. The only thing we have to establish is that for an integer $m > 0$ (under control), the operator R^m is a contraction in $\mathcal{H}_{p,0}^{n+2}(T)$.

By Theorem 5.1, for $t \leq T$,

$$\begin{aligned}||Ru - Rv||^p_{\mathcal{H}_p^{n+2}(t)} &\leq N||(L_1(u) - L_1(v), \Lambda_1(u) - \Lambda_1(v))||^p_{\mathcal{F}_p^n(t)} \\ &\leq N_0\varepsilon^p||u - v||^p_{\mathcal{H}_p^{n+2}(t)} + N_0K_1^p\int_0^t E||u(s) - v(s)||^p_{n+1,p}\,ds,\end{aligned}$$

with the constant N_0 depending only on d, p, δ, and K. This gives the desired result if $K_1 = 0$. In the general case, by Theorem 3.1 (or by Theorem 2.1),

$$E||u(s) - v(s)||^p_{n+1,p} \leq N_1||u - v||^p_{\mathcal{H}_p^{n+2}(s)},$$

where $s \leq T$ and N_1 depends only on d, p, and T. It follows that for $t \leq T$ and $\theta := N_0\varepsilon^p$, we have

$$||Ru - Rv||^p_{\mathcal{H}_p^{n+2}(t)} \leq \theta||u - v||^p_{\mathcal{H}_p^{n+2}(t)} + N_2\int_0^t ||u - v||^p_{\mathcal{H}_p^{n+2}(s)}\,ds,$$

where N_2 depends only on d, p, δ, K, K_1, and T. Hence, by induction,

$$\begin{aligned}||R^m u - R^m v||^p_{\mathcal{H}_p^{n+2}(t)} &\leq \theta^m||u - v||^p_{\mathcal{H}_p^{n+2}(t)} \\ &+\sum_{k=1}^m \binom{m}{k}\theta^{m-k}N_2^k\int_0^t \frac{(t-s)^{k-1}}{(k-1)!}||u - v||^p_{\mathcal{H}_p^{n+2}(s)}\,ds,\end{aligned}$$

$$||R^m u - R^m v||^p_{\mathcal{H}^{n+2}_p(T)} \leq \sum_{k=0}^{m} \binom{m}{k} \theta^{m-k} \frac{1}{k!} (TN_2)^k ||u - v||^p_{\mathcal{H}^{n+2}_p(T)}$$
$$\leq 2^m \theta^m \max_k \frac{1}{k!} (TN_2/\theta)^k ||u - v||^p_{\mathcal{H}^{n+2}_p(T)}.$$

This allows us to find a needed m, and the theorem is proven. □

We finish our preparations by showing how Lemma 3.1 will be used.

Remark 5.2. To some extent, in what follows, the most important consequence of assertion (i) in Lemma 3.1 is that for any a from the lemma there exists a new norm $]|\cdot|[_{n,p}$ in H^n_p such that

$$]|au|[_{n,p} \leq 2N||a||_B]|u|[_{n,p},$$

where N is the same constant as in Lemma 3.1. To show this, it suffices to observe that for $a_m(x) = a(x/m)$ and $u_m(x) = u(x/m)$, we have

$$||(m^2 I - \Delta)^{n/2}(au)||_p = m^{n-d/p}||(I - \Delta)^{n/2}(a_m u_m)||_p$$
$$\leq N\bar{a}_m m^{n-d/p}||(I - \Delta)^{n/2} u_m||_p = N\bar{a}_m||(m^2 I - \Delta)^{n/2} u||_p$$
$$\leq N\left(||a||_B + \frac{1}{m^{(|n|+\gamma)\wedge 1}}\bar{a}\right)||(m^2 I - \Delta)^{n/2} u||_p.$$

Alternatively, it would be sufficient for our needs to know that

$$||au||_{n,p} \leq N(||a||_B||u||_{n,p} + ||a||_{C^{|n|+\gamma}}||u||_{n-1,p}).$$

Unfortunately, the author could not find the last inequality in the literature, though some very interesting information related to the subject can be found in [7].

Now we perform the main step in proving Theorem 3.2. Below we suppose that its assumptions are satisfied.

LEMMA 5.1. *There exists an $\varepsilon = \varepsilon(d, p, n, \gamma, \delta, K) > 0$ such that if $\tau = T$ and*

(i) *condition* (3.5) *is satisfied with this ε for all x, y, t, and ω and*

(ii) *f and g are independent of u,*

then there exists a unique solution $u \in \mathcal{H}^{n+2}_{p,0}(T)$ of equation (3.4). *Furthermore, for this solution u, we have*

$$||u||_{\mathcal{H}^{n+2}_p(T)} \leq N||(f, g)||_{\mathcal{F}^n_p(T)},$$

where N depends only on d, p, δ, K, and T.

Proof. Define $a(t) = a(t, 0)$ and $\sigma(t) = \sigma(t, 0)$, take operators L and Λ from Theorem 5.1 corresponding to $a(t)$ and $\sigma(t)$, and let

$$L_1(t, u)(x) = [a^{ij}(t, x) - a^{ij}(t)]u_{x^i x^j}(x), \quad \Lambda_1(t, u)(x) = [\sigma^i(t, x) - \sigma^i(t)]u_{x^i}(x).$$

In view of Theorem 5.2, to prove existence and uniqueness, we have only to check that if ε in (3.5) is sufficiently small, then the operators L_1 and Λ_1 satisfy condition (5.4) with as small ε as we like and with K_1 under control. Observe that by Lemma 3.1,

$$||[a^{ij}(t, \cdot) - a^{ij}(t)]u_{x^i x^j}||_{n,p} \leq N||a(t, \cdot) - a(t, 0)||_{C^{|n|+\gamma}}||u_{xx}||_{n,p},$$

$$||[\sigma^i(t, \cdot) - \sigma^i(t)]u_{x^i}||_{n+1,p} \leq N||\sigma(t, \cdot) - \sigma(t, 0)||_{C^{|n|+\gamma+1}}||u_x||_{n+1,p}.$$

Since $||u_x||_{n+1,p} \leq N(||u_{xx}||_{n,p} + ||u||_{n+1,p})$, the lemma holds true if

$$||a(t,\cdot)-a(t,0)||_{C^{|n|+\gamma}} + ||\sigma(t,\cdot)-\sigma(t,0)||_{C^{|n|+\gamma+1}(\mathbb{R}^d,E)} \leq \varepsilon_0 = \varepsilon_0(d,p,n,\gamma,\delta,K) \quad \forall\, t.$$

Next, observe that for $a_m(t,x) = a(t/m^2, x/m)$, we have

$$||a_m(t,\cdot) - a_m(t,0)||_{C^{|n|+\gamma}} \leq \varepsilon + 2m^{-[((|n|+\gamma)\wedge 1]}K,$$

and an analogous inequality holds for σ. It follows that for m sufficiently large, the statements of the lemma are true if we replace a, σ, w_t, f, g, and T in equation (3.4) by $a_m, \sigma_m, mw_{t/m^2}$, $m^{-2}f(t/m^2, x/m)$, $m^{-1}g(t/m^2, x/m)$, and m^2T, respectively. After this, it remains only to fix an appropriate m and make an obvious change of variables in the above-mentioned modification of equation (3.4) (and use that $I-\Delta \sim m^2I-\Delta$). The lemma is proven. □

Finally, we need the following result from [9], which in a sense is essentially covered by Theorem 2.4.7 from [18].

LEMMA 5.2. *Let $\delta > 0$ and let $\zeta_k \in C^\infty$, $k = 1,2,3,\ldots$. Assume that for any multiindex α and $x \in \mathbb{R}^d$,*

$$\sup_{x\in\mathbb{R}^d} \sum_k |D^\alpha \zeta_k(x)| \leq M(\alpha),$$

where $M(\alpha)$ are some constants. Then there exists a constant $N = N(d,n,M,\delta)$ such that for any $f \in H_p^n$,

$$\sum_k ||\zeta_k f||^p_{H_p^n} \leq N||f||^p_{H_p^n}.$$

If in addition

$$\sum_k |\zeta_k(x)|^p \geq \delta,$$

then for any $f \in H_p^n$,

$$||f||^p_{n,p} \leq N \sum_k ||\zeta_k f||^p_{n,p}.$$

Proof of Theorem 3.2. First observe that by considering $u(t,x) - u_0(x)$ and using Lemma 3.1, one easily reduces the general situation to the case $u_0 = 0$. For an obvious reason, we can assume that $\tau \equiv T$. In this case, define

$$Lu = a^{ij}(t,x)u_{x^ix^j}(t,x), \quad \Lambda u = \sigma^i(t,x)u_{x^i}(t,x),$$

and let $\{\zeta_k : k = 1,2,3,\ldots\}$ be a standard partition of unity such that for any k, condition (3.5) is satisfied for x and y in the support of ζ_k with ε from Lemma 5.1. Then by this lemma, for any $u \in \mathcal{H}^{n+2}_{p,0}(T)$ and for any k,

$$\begin{aligned} ||u\zeta_k||_{\mathcal{H}_p^{n+2}(T)} &\leq N||(L,\Lambda)(u\zeta_k)||_{\mathcal{F}_p^n(T)} \\ &\leq N||\zeta_k(L,\Lambda)u||_{\mathcal{F}_p^n(T)} + N||(uL\zeta_k + 2(a\zeta_{kx}, u_x), u\Lambda\zeta_k)||_{\mathcal{F}_p^n(T)}. \end{aligned}$$

By summing up the pth powers of the extreme terms and applying Lemma 5.2, we conclude that for any $u \in \mathcal{H}^{n+2}_{p,0}(T)$,

$$||u||_{\mathcal{H}_p^{n+2}(T)} \leq N(||(L,\Lambda)u||_{\mathcal{F}_p^n(T)} + ||u||_{\mathcal{H}_p^{n+1}(T)}).$$

Actually, the last term in parentheses on the right can be dropped, which follows from Gronwall's inequality and assertions (ii) and (iii) in Theorem 3.1 (cf. the proof of Theorem 3.3). Hence, the standard method of continuity (cf., for instance, [4]) (when one considers $(1-\lambda)\delta^{ij}+\lambda a^{ij}$ and λw instead of a^{ij} and w and makes λ vary in $[0,1]$) and the above a priori estimate as usual yield the existence and uniqueness of solution for equation (3.4) when f and g are independent of u. To consider general f and g, it remains only to repeat the proof of Theorem 5.2 taking $f(u,t,x)$ and $g(u,t,x)$ instead of $f+L_1(u)$ and $g+\Lambda_1(u)$ there. The theorem is proven. □

Acknowledgment. The author thanks S. Lapic for pointing out several mistakes and misprints in the original version of the article.

REFERENCES

[1] G. Da Prato, *Some results on linear stochastic evolution equations in Hilbert spaces by the semi-group method*, Stochastic Anal. Appl., 1 (1983), pp. 57–88.

[2] G. Da Prato and J. Zabczyk, *A note on stochastic convolution*, Stochastic Anal. Appl., 10 (1992), pp. 143–153.

[3] A. M. Garsia, E. Rodemich, and H. Rumsey, Jr., *A real variable lemma and the continuity of paths of some Gaussian processes*, Indiana Univ. Math. J., 20 (1970), pp. 565–578.

[4] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[5] M. A. Krasnoselskii, E. I. Pustylnik, P. E. Sobolevskii, and P. P. Zabrejko, *Integral operators in spaces of summable functions*, Nauka, Moscow, 1966 (in Russian); Noordhoff International Publishing, Leyden, 1976 (English translation).

[6] I. Gyöngy and N. V. Krylov, *On stochastic equations with respect to semimartingales* I, Stochastics, 4 (1980), pp. 1–21.

[7] C. E. Kenig, G. Ponce, and L. Vega, *Well-posedness and scattering results for the generalized Korteweg–de Vries equation via the contraction principle*, Comm. Pure Appl. Math., 46 (1993), pp. 527–620.

[8] N. V. Krylov, *On moment estimates for quasiderivatives of solutions of stochastic equations with respect to initial data, and their application*, Mat. Sb., 136 (1988), pp. 510–529 (in Russian); Math. USSR Sb., 64 (1989), pp. 505–526 (English translation).

[9] ———, *A generalization of the Littlewood–Paley inequality and some other results related to stochastic partial differential equations*, Ulam Quart., 2 (4) (1994), pp. 16–26.

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

[11] N. V. Krylov and B. L. Rozovskii, *On the characteristics of degenerate second order parabolic Itô equations*, Trudy Sem. Petrovsk., 8 (1982), pp. 153–168 (in Russian); J. Soviet Math., 32 (1986), pp. 336–348 (English translation).

[12] O. A. Ladyzhenskaia, V. A. Solonnikov, and N. N. Ural'tceva, *Linear and quasi-linear equations of parabolic type*, Nauka, Moscow, 1967 (in Russian); American Mathematical Society, Providence, 1968 (English translation).

[13] S. Lapic, *On the first initial-boundary problem for SPDEs on domains with limited smoothness at the boundary*, Potential Anal., submitted.

[14] B. L. Rozovskii, *Stochastic evolution systems*, Kluwer, Dordrecht, 1990.

[15] C. Mueller, *Long time existence for the heat equation with a noise term*, Probab. Theory Related Fields, 90 (1991), pp. 505–517.

[16] D. Nualart and E. Pardoux, *White noise driven quasilinear SPDEs with reflection*, Probab. Theory Related Fields, 93 (1992), pp. 77–89.

[17] H. Triebel, *Theory of Function Spaces*, Birkhäuser, Basel, Boston, Stuttgart, 1983.

[18] ———, *Theory of Function Spaces* II, Birkhäuser, Basel, Boston, Berlin, 1992.

[19] J. B. Walsh, *An introduction to stochastic partial differential equations*, Lecture Notes in Math. 1180, Springer-Verlag, Berlin, New York, Heidelberg, 1986.

BEHAVIOUR IN THE LIMIT, AS $p \to \infty$, OF MINIMIZERS OF FUNCTIONALS INVOLVING p-DIRICHLET INTEGRALS[*]

ULF JANFALK[†]

Abstract. The purpose of this paper is to study the behaviour, as $p \to \infty$, of minimizers of functionals involving p-Dirichlet integrals in a bounded Lipschitz domain, $\Omega \subset \mathbf{R}^n$. In the case where Ω is a convex ring it is proved that the minimizers converge monotonically and uniformly. In the paper by T. Bhattacharya, E. DiBenedetto, and J. Manfredi [Limits as $p \to \infty$ of $\Delta_p u_p = f$ and related extremal problems, *Rend. Sem. Mat. Univ. Politec. Torino*, (1989), pp. 15–68], the problem of torsional creep is studied. Here the situation is generalized by introducing a more general functional and relaxing the boundary conditions. Various aspects of the Green function of the p-laplacian are considered and it is proved that the Green function is not symmetric if p is sufficiently large. Finally, it is proved that the extremals to the dual problem tend to zero in the mean as $p \to \infty$, outside a well-specified subset of Ω.

Key words. p-harmonic, convex duality, variational solution, capacitary function, p-Dirichlet integral, ridge, Green function

AMS subject classifications. Primary, 35J20; Secondary, 35B40

1. Introduction.

In this paper we will investigate the behaviour and convergence properties as $p \to \infty$ of sequences of minimizers of certain convex minimization problems in Ω involving p-Dirichlet integrals. Throughout the paper Ω will be a bounded domain in $\mathbf{R}^n$ with Lipschitz boundary $\partial\Omega$.

The paper by Bhattacharya et al. [BDM] constitutes an important background for this work. There the limits as $p \to \infty$ of solutions to the equation $\Delta_p u = f$, where Δ_p is the p-laplace or p-harmonic operator, are studied in connection with the problem of torsional creep. The limit process leads to the study of the operator Δ_∞ which is obtained as a formal limit of the operator Δ_p as $p \to \infty$ (see §2 for a more detailed description of these operators). Note that Δ_∞ is not of a variational nature. This leads to some complications in finding a suitable solution concept. In [BDM], four different solution concepts are considered. These are classical solutions, absolute minimals, viscosity solutions, and variational solutions. Their mutual relations are also investigated. Classical solutions and absolute minimals have been studied in some detail by Aronsson in two papers, [Ar1] and [Ar2], and among other things, it was shown that classical solutions to the Dirichlet problem for $\Delta_\infty u = 0$ are in some sense rare. Recently, Jensen has shown that absolute minimals are viscosity solutions of $\Delta_\infty u = 0$ [J, Thm. 1.15] and that the Dirichlet problem for $\Delta_\infty u = 0$ has a unique viscosity solution [J, Thm. 2.22, p. 70] provided the boundary values are in the space

$$\mathrm{Lip}(\partial\Omega) = \left\{ g \in C(\partial\Omega) \colon \sup_{x,y \in \partial\Omega} \frac{|g(x) - g(y)|}{d_\Omega(x,y)} < \infty \right\},$$

where d_Ω denotes the distance within Ω. This condition on the boundary values is also necessary for an absolute minimal and a variational solution to exist. By [BDM, Prop. 2.2, p. 27] variational solutions are viscosity solutions so the results of Jensen thereby show that the solution concept's absolute minimal, variational solution and viscosity solution are equivalent. Here we will use the variational solution approach since it is better suited for our purposes.

[*] Received by the editors July 23, 1993; accepted for publication (in revised form) June 7, 1994.

[†] Department of Mathematics, Linköping University, S-581 83 Linköping, Sweden.

The paper is organized as follows. In §2 we consider variational solutions of the limit equation, derived from capacitary functions for a particular type of condensers, called ring domains. We prove monotone and uniform convergence of the capacitary functions when Ω is a convex ring. This also implies uniqueness of the variational solution.

In §3 we generalize the situation in [BDM], e.g., we consider a slightly more general functional and relax the boundary conditions somewhat. We also introduce a weight function. We study the problem using methods from convex analysis. We derive a dual problem and investigate the relation between the primal and dual problems. The majority of the results we prove are related to those of [BDM] and so are the proofs. We end the section by proving that, if μ is a finite, positive Borel measure then the family $\{u_p\}_{p>n}$ of solutions of $-\Delta_p u = \mu$, $u = 0$ on $\partial\Omega$, has a unique limit as $p \to \infty$.

Section 4 deals with various aspects of the Green function for the p-laplacian. An interesting detail regarding the Green function is that it fits into the setting of both §§2 and 3. It is well known that the Green function is symmetric if $p = 2$. We show here that if p is sufficiently large then the Green function is not symmetric.

We finish this paper in §5 with a theorem concerning the asymptotic behaviour of the extremals of the dual problems. We prove, roughly speaking, that the "mass" of the solution to the dual problem concentrates on the set of uniqueness as $p \to \infty$ $(q \to 1+)$.

2. On the monotone convergence of capacitary functions in a convex ring and a comparison principle. The variational equation

$$\int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla\varphi\, dx = 0, \quad \text{for all } \varphi \in C_0^\infty(\Omega), \tag{2.1}$$

also written $\operatorname{div}(|\nabla u|^{p-2}\nabla u) = 0$, is the Euler equation for the p-Dirichlet integral

$$I_p(u) = \int_\Omega |\nabla u|^p\, dx \quad (1 < p < \infty).$$

The nonlinear operator associated with (2.1) is called the p-laplacian or the p-harmonic operator and will be denoted by Δ_p. Any function $u \in W^{1,p}_{loc}(\Omega)$ satisfying (2.1) will be called p-harmonic. The equation $\Delta_\infty u = 0$ is obtained as a formal limit of $\Delta_p u = 0$ as $p \to \infty$. To be explicit, suppose u is p-harmonic, $u \in C^2(\Omega)$, and $\nabla u \neq 0$; then

$$\begin{aligned} \Delta_p u &= \operatorname{div}(|\nabla u|^{p-2}\nabla u) \\ &= |\nabla u|^{p-4}\left((p-2)\sum_{i,j=1}^n u_{x_i}u_{x_j}u_{x_ix_j} + |\nabla u|^2\Delta u\right) = 0. \end{aligned} \tag{2.2}$$

Thus, dividing both sides by $(p-2)|\nabla u|^{p-4}$ and letting $p \to \infty$ we get formally

$$\Delta_\infty u = \sum_{i,j=1}^n u_{x_i}u_{x_j}u_{x_ix_j} = 0.$$

In [Ar2, Thm. 13, p. 425] it was shown that if $\Omega \subset \mathbf{R}^2$ then the class $C^2(\Omega)$ is too small to solve the Dirichlet problem

$$\begin{cases} \Delta_\infty u = 0 \text{ in } \Omega, \\ u = g \quad \text{on } \partial\Omega \end{cases} \tag{2.3}$$

if, e.g., Ω is the unit disk and g is any nonconstant function satisfying $g(x,y) = g(-x,-y)$ for x,y on the unit circle. Instead, an alternative solution concept, the so-called absolute minimals, was introduced. In [BDM] the concept of a variational solution was introduced. A variational solution of (2.3) is obtained by the following limit procedure. Let $g \in W^{1,\infty}(\Omega)$ be given. Denote by u_p the unique solution in $W^{1,p}(\Omega)$ of the boundary value problem

$$\begin{cases} \Delta_p u = 0 & \text{in } \Omega, \\ u - g \in W_0^{1,p}(\Omega). \end{cases} \tag{2.4}$$

Let $m > n$ and put $G_m = \{u_p : p > m\}$. In [BDM, pp. 25, 26] it is shown that there is an element $u_\infty \in W^{1,\infty}(\Omega)$ and a sequence $\{u_{p_k}\}_{k=1}^{\infty}$ from G_m such that

$$u_{p_k} \to u_\infty \quad \text{weakly in } W^{1,m}(\Omega), \tag{2.5}$$

$$u_{p_k} \to u_\infty \quad \text{in } C^{0,\lambda}(\overline{\Omega}), \text{ for any } \lambda \in \left(0, 1 - \frac{n}{m}\right) \tag{2.6}$$

as $p_k \to \infty$. Following the terminology of [BDM] we will call any element u_∞ obtained in this way a variational solution of $\Delta_\infty u = 0$. By [BDM, Prop. 2.2, p. 27] variational solutions are viscosity solutions. Thus by [J, Thm. 2.22, p. 70], we get that a variational solution is unique and it is easy to see that (2.5) and (2.6) hold with u_{p_k} replaced by u_p (see the proof of Theorem 2.5).

In this section we will study the variational solutions obtained from capacitary functions of a special type of condensers, called ring domains. By a ring domain Ω we mean a bounded doubly connected domain in $\mathbf{R}^n$. Denote by K the bounded component of $\mathbf{R}^n \setminus \Omega$ and put $D = \Omega \cup K$. We say that Ω is a convex ring if D and K are convex. Note that the definition requires that K is compact and contained in the open set D. Further, a convex ring is not a convex set.

DEFINITION 1. Let $1 < p < \infty$ and let K be a compact subset of the open set D. The quantity

$$\operatorname{cap}_p(K,D) = \inf\left\{ \int_D |\nabla v|^p \, dx : v \in W_0^{1,p}(D) \cap C(D),\ v \geq 1 \text{ on } K \right\}$$

is called the p-capacity of the condenser (K,D). A function $u \in W_0^{1,p}(D)$, $u \geq 1$ on K, such that

$$\operatorname{cap}_p(K,D) = \int_D |\nabla u|^p \, dx$$

is called a capacitary function with respect to p and (K,D). This is the definition of capacity according to [HKM]. If $p > n$ then the capacitary function with respect to p and (K,D) is equal to 1 on K. In [L, pp. 202–204], Lewis proves that the capacitary function of a ring domain is unique and that it is p-harmonic in Ω. The definition of capacity according to Lewis is somewhat different but yields the same result. In particular, the capacitary function of the convex ring $\Omega = D \setminus K$ is the same as the capacitary function of the condenser (K,D). The following theorem is also proved (see [L, Thm. 1, p. 204]).

THEOREM 2.1. *Given a convex ring $\Omega = D \setminus K$ and a constant p, $1 < p < \infty$, let u be the capacitary function corresponding to p and Ω. Then u has a continuous representative in $W_0^{1,p}(D)$ such that*

(1) *the set* $\{x : u(x) > t\}$ *is convex for* $0 \leq t < 1$,
(2) u *is real analytic in* Ω,
(3) *if* $u \not\equiv 0$ *and* $x \in \Omega$, *then all normal curvatures at* x *of the level surface* $\{y : u(y) = u(x)\}$ *are positive,*
(4) *if* $u \not\equiv 0$ *then* ∇u *is univalent in* Ω *and maps* $\{x : u(x) = t\}$, $0 < t < 1$, *onto a surface which is star shaped with respect to the origin.*

In particular, $\nabla u \neq 0$ in Ω (see [L, Lem. 2, p. 207]). In $\mathbf{R}^2$ the regularity part of Theorem 2.1 holds for any ring domain as follows from Theorem 2.2.

THEOREM 2.2. *Let* $\Omega = D \setminus K \subset \mathbf{R}^2$ *be a ring domain* (*not necessarily convex*) *and let* $1 < p < \infty$. *Then the corresponding capacitary function* u_p *is real analytic in* Ω.

Proof. We first observe that Ω being a ring domain implies that K is either a continuum of positive diameter or a single point. If K is a single point and $1 < p \leq 2$ then $\text{cap}_p(K, D) = 0$ so that $u_p \equiv 0$. We can thus assume that $\text{cap}_p(K, D) > 0$. If $p = 2$ the result is classical. In the case $p \neq 2$ we first claim that u_p can be extended continuously to K by $u_p = 1$ on K. For $p > 2$ this is obvious and in the case $1 < p < 2$ we do as follows. By [M1, 9.1.2, Prop. 1, p. 392] we have for any continuum e that

$$\text{cap}_p(e, \mathbf{R}^2) \sim d^{2-p}$$

where d is the diameter of e. Let $B_r(x)$ be the disc with radius r centered at x. Obviously, for every $x \in K$ and every $r > 0$ such that $B_r(x) \subset \Omega$ there is a continuum $K_r(x) \subset K \bigcap B_r(x)$ such that $r \leq \text{diam} K_r(x) \leq 2r$. Clearly, $\text{cap}_p(K \bigcap B_r(x), B_{2r}(x)) \geq \text{cap}_p(K \bigcap B_r(x), \mathbf{R}^2) \geq \text{cap}_p(K_r(x), \mathbf{R}^2)$ and by, e.g., [HKM, p. 35] we have that $\text{cap}_p(B_r(x), B_{2r}(x)) = c_p r^{2-p}$ where $c_p > 0$ is a constant. Hence, for every $x \in K$ the Wiener-type integral

$$\int_0^1 \left(\frac{\text{cap}_p(K \bigcap B_r(x), B_{2r}(x))}{\text{cap}_p(B_r(x), B_{2r}(x))} \right)^{\frac{1}{p-1}} \frac{dr}{r} = \infty.$$

By [M2, Thm., p. 236] the claim follows (see also [HKM, 6.17, p. 114; Cor. 6.28, p. 122]). For $0 < t < 1$, put $\Omega_t = \{x \in D \colon u_p(x) > t\}$. It is enough to show that $\nabla u_p(x) \neq 0$ for every $x \in \Omega$ (see [L, pp. 207–208]). Clearly, Ω_t is open. Further, Ω_t can have one component only. To see this, let V_1 be a component of Ω_t that intersects K and put $V_2 = \Omega_t \setminus V_1$. Since u_p is continuous on $\overline{\Omega}$ and $u_p = 1$ on K it is clear that $K \subset V_1$. This implies that $\overline{V}_2 \subset \Omega$. By the comparison principle (see [T, p. 312]) we then get that u_p is constant on V_2, which is impossible.

Suppose x_1 is such that $u_p(x_1) = t$ and $\nabla u_p(x_1) = 0$. Let N denote the order of the zero of ∇u_p at x_1, as defined in [Ar3, p. 76]. Using [Ar3, Thm. 4, pp. 83–84], we see that there is a $\delta > 0$ such that $\varphi(x) = u_p(x) - t$ has a representation in hodograph coordinates valid for $B_\delta = \{x : |x - x_1| < \delta\}$. From this representation it is evident that B_δ can be split into $2N$ parts such that φ is nonnegative in N parts, nonpositive in N parts, and the parts where φ has the same sign have x_1 as their only common point. Since Ω_t has only one component this situation is obviously impossible. Thus $\nabla u_p \neq 0$ on Ω. □

Before we proceed we need two more definitions.

DEFINITION 2. (1) A function $u \in W^{1,p}_{loc}(\Omega)$ is a p-supersolution of Δ_p if $-\Delta_p u \geq 0$ in the weak sense, i.e., if

$$\int_\Omega |\nabla u|^{p-2} \nabla u \cdot \nabla \varphi \, dx \geq 0$$

for all $\varphi \in C_0^\infty(\Omega)$ such that $\varphi \geq 0$.

(2) A function $u : \Omega \to \mathbf{R} \bigcup \{\infty\}$ is p-superharmonic if

(i) u is lower semicontinuous,

(ii) $u \not\equiv \infty$ in each component of Ω,

(iii) for each bounded open set D, such that $\overline{D} \subset \Omega$ and for each p-harmonic function $h \in C(\overline{D})$ the inequality $u \geq h$ on ∂D implies $u \geq h$ on D.

In [HKM, pp. 136–138] it is shown that a p-supersolution is always p-superharmonic, possibly after a change on a set of measure zero and that if the p-superharmonic function is in $W^{1,p}_{loc}(\Omega)$ then it is also a p-supersolution.

Lemma 2.4 relies heavily on Theorem 2.1(3). Therefore we digress a bit from the main subject of this section and study the normal curvature of the level sets of a smooth function. We will now derive the identity (2.7).

Let u be a C^2 function on an open subset of $\mathbf{R}^n$ and suppose $|\nabla u(x)| \neq 0$ for some x. Put $E_x = \{y : u(y) = u(x)\}$ and denote the tangent space at x by $T(x)$. The normal curvature, according to Lewis, of E_x at x in the direction of the unit vector $t \in T(x)$ is defined to be

$$K(t, x) = -\frac{u_{tt}(x)}{|\nabla u(x)|}$$

where the subscript t denotes differentiation in the direction given by $t \in T(x)$. Then

$$u_{tt} = t \cdot \nabla(t \cdot \nabla u) = \sum_{i,j=1}^{n} t_i t_j u_{x_i x_j}.$$

For $1 \leq r < k \leq n$ define the $n \times n$-matrix $A_{r,k} = (a_{i,j})_{i,j=1}^n$ by $a_{r,k} = 1$, $a_{k,r} = -1$, and $a_{i,j} = 0$ otherwise, i.e., $A_{r,k}\nabla u(x)$ is a vector in $\mathbf{R}^n$ having $-u_{x_k}$ and u_{x_r} as rth and kth components, respectively. Thus $\nabla u(x) \cdot A_{r,k}\nabla u(x) = 0$. Since $\nabla u(x) \neq 0$ it follows that $T(x)$ is spanned by $\{A_{r,k}\nabla u(x)\}_{1 \leq r < k \leq n}$. This is easily seen if one parametrizes the hyperplane $y \cdot \nabla u(x) = 0$. Put

$$t_{r,k} = \frac{A_{r,k}\nabla u(x)}{|A_{r,k}\nabla u(x)|}$$

if $A_{r,k}\nabla u(x) \neq 0$ and $t_{r,k} = 0$ otherwise. Here, as before, $1 \leq r < k \leq n$. With this choice of tangent directions we obtain

$$\begin{aligned}
&|\nabla u(x)| \sum_{r=1}^{n-1} \sum_{k=r+1}^{n} |A_{r,k}\nabla u(x)|^2 K(t_{r,k}, x) \\
&= -\sum_{r=1}^{n-1} \sum_{k=r+1}^{n} |A_{r,k}\nabla u(x)|^2 t_{r,k} \cdot \nabla(t_{r,k} \cdot \nabla u)(x) \\
&= -\sum_{r=1}^{n-1} \sum_{k=r+1}^{n} \left(u_{x_k}^2 u_{x_r x_r} + u_{x_r}^2 u_{x_k x_k} - 2u_{x_r} u_{x_k} u_{x_r x_k}\right) \\
&= -\sum_{i=1}^{n} \left(|\nabla u|^2 - u_{x_i}^2\right) u_{x_i x_i} + \sum_{\substack{i,j=1 \\ i \neq j}} u_{x_i} u_{x_j} u_{x_i x_j} \\
&= \Delta_\infty u - |\nabla u|^2 \Delta u.
\end{aligned}$$

Above, u and the derivatives of u that occur are to be evaluated at x. We have proved the following lemma.

LEMMA 2.3. *Suppose that u is a C^2 function in a neighbourhood of the point $x \in \mathbf{R}^n$ and that $\nabla u(x) \neq 0$. Then*

$$(2.7)\quad |\nabla u(x)| \sum_{r=1}^{n-1} \sum_{k=r+1}^{n} |A_{r,k}\nabla u(x)|^2 K(t_{r,k}, x) = (\Delta_\infty u)(x) - |\nabla u(x)|^2 (\Delta u)(x).$$

Note that the right-hand side of (2.7) is invariant under orthogonal coordinate transformations.

We now have sufficient background to prove the following fundamental lemma.

LEMMA 2.4. *Let $\Omega = D \setminus K$ be a convex ring and suppose $\operatorname{cap}_p(K, D) > 0$. For $1 < p < \infty$, denote by u_p the capacitary function corresponding to p and Ω. If $1 < p_1 < p_2$ then u_{p_2} is both p_1-superharmonic and a p_1-supersolution. Furthermore, u_{p_1} is both p_2-subharmonic and a p_2-subsolution. Hence $u_{p_1}(x) \leq u_{p_2}(x)$ in Ω.*

Proof. By Theorem 2.1, u_p is real analytic and $\nabla u_p(x) \neq 0$ for all $x \in \Omega$. Thus the p-laplacian can be interpreted in the classical sense. Hence, by (2.2) we have

$$(2.8)\qquad \Delta u_p = -\frac{p-2}{|\nabla u_p|^2} \Delta_\infty u_p$$

for all $p > 1$. When dealing with u_p we add the subscript p whenever needed. By Theorem 2.1, $K_p(t, x) > 0$. Since u_p is p-harmonic we get by inserting (2.8) into (2.7) that

$$(2.9)\qquad |\nabla u_p(x)| \sum_{r=1}^{n-1} \sum_{k=r+1}^{n} |A_{r,k}\nabla u_p(x)|^2 K_p(t_{r,k}, x) = (p-1)\Delta_\infty u_p > 0,$$

for all $p > 1$.

Now let $p_1 < p_2$. Then by (2.8) and (2.9) we get

$$\begin{aligned} -\Delta_{p_1} u_{p_2} &= -|\nabla u_{p_2}|^{p_1 - 4} \left((p_1 - 2)\Delta_\infty u_{p_2} + |\nabla u_{p_2}|^2 \Delta u_{p_2} \right) \\ &= -|\nabla u_{p_2}|^{p_1 - 4} \left((p_1 - 2)\Delta_\infty u_{p_2} - (p_2 - 2)\Delta_\infty u_{p_2} \right) \\ &= (p_2 - p_1)|\nabla u_{p_2}|^{p_1 - 4} \Delta_\infty u_{p_2} > 0. \end{aligned}$$

Consequently, u_{p_2} is a p_1-supersolution and therefore u_{p_2} is p_1-superharmonic. That u_{p_1} is a p_2-subsolution follows directly from above by replacing p_1 by p_2 and vice versa. □

By (2.5) and basic properties of p-superharmonic functions we get the following consequences.

THEOREM 2.5. *Let $\Omega = D \setminus K$ be a convex ring and let $f \in W^{1,\infty}(\Omega)$ be such that $f = 0$ on ∂D and $f = 1$ on K. Then the boundary value problem*

$$(2.10)\qquad \begin{cases} \Delta_\infty u = 0 & \text{in } \Omega, \\ u - f \in W_0^{1,\infty}(\Omega) \end{cases}$$

has a unique variational solution u_∞. Moreover, $u_p \to u_\infty$ in $C^{0,\lambda}(\overline{\Omega})$ for any $\lambda \in (0, 1)$ as $p \to \infty$ and $u_p(x) \nearrow u_\infty(x)$ as $p \to \infty$ for every $x \in \Omega$.

Proof. Let u_p be the solution to (2.4) and consider the set $G_n = \{u_p\}_{p>n}$. By Lemma 2.4 we have that u_p is p_1-superharmonic for any $1 < p_1 < p$. Let $\{u_{p_k}\}_{k=1}^{\infty} \subset$

G_n be a sequence with limit element u_∞ as in (2.5) and (2.6). From Definition 2(2(iii)) it then follows that $u_{p_i}(x) \leq u_p(x) \leq u_{p_j}(x)$ for all $x \in \Omega$ if $p_i \leq p \leq p_j$. Thus,

$$u_p \longrightarrow u_\infty \tag{2.11}$$

uniformly as $p \to \infty$ since $u_{p_k} \to u_\infty$ in $C^{0,\lambda}(\overline{\Omega})$, $\lambda \in (0, 1 - \frac{n}{p_1})$. To see that this holds for the entire net G_n, suppose on the contrary that there exists a sequence $\{u_{p_k^*}\}_{k=1}^\infty$ with $p_1^* > p_1$ and a $\delta > 0$ such that

$$\| u_{p_k^*} - u_\infty \|_{C^{0,\lambda}} \geq \delta \tag{2.12}$$

for all k and some $\lambda \in (0, 1 - \frac{n}{p_1})$. The sequence $\{u_{p_k^*}\}_{k=1}^\infty$ is bounded in $W_0^{1,p_1}(\Omega)$ (see [BDM, pp. 25–26]) and by the Rellich–Kondrachov theorem it has a subsequence converging to a function u_∞^* in $C^{0,\lambda}(\overline{\Omega})$. It follows immediately from (2.11) that

$$u_\infty^* = u_\infty$$

which contradicts (2.12).

To finish the proof, take $\lambda \in (0, 1)$ arbitrary and choose k large enough so that $\lambda < 1 - \frac{n}{p_k}$. Clearly the above reasoning can be repeated and we deduce

$$u_p \longrightarrow u_\infty$$

in $C^{0,\lambda}(\overline{\Omega})$ for any $\lambda \in (0, 1)$. □

Remark. By [BDM, p. 29], a variational solution is also a viscosity solution. Thus the uniqueness part of this theorem also follows from [J, Thm. 2.22, p. 70].

By applying Lemma 7.3 of [HKM, p. 132] we also obtain the following result.

COROLLARY 2.6. *The variational solution u_∞ of* (2.10) *is p-superharmonic for all $p > 1$.*

3. On the asymptotic behaviour of solutions to $-\Delta_p u = \mu$. Let $\Omega \subset \mathbf{R}^n$ be a bounded domain with Lipschitz boundary $\partial\Omega = \Gamma_0 \cup \Gamma_1$ where $\Gamma_0 \cap \Gamma_1 = \emptyset$ and $\Gamma_0 \neq \emptyset$. Let μ be a Borel measure on Ω such that $0 < |\mu|(\Omega) < \infty$ and let $f \in L^\infty(\Omega)$ be such that $0 < C \leq f \leq D < \infty$. For $n < p < \infty$ and $u \in W^{1,p}(\Omega)$, put

$$J_p(u) = \frac{1}{p} \int_\Omega f(x) |\nabla u|^p \, dx - \int_\Omega u \, d\mu$$

and define the class $\mathcal{K}_p = \{u \in W^{1,p}(\Omega) \ : \ u = 0 \text{ on } \Gamma_0\}$. Clearly $\mathcal{K}_p$ is a closed subspace of $W^{1,p}(\Omega)$ and hence it is itself a Banach space. Since the functions in $\mathcal{K}_p$ are continuous we can also assume that Γ_0 is closed. Consider the problem $\mathcal{P}_p$: Find $u_p \in \mathcal{K}_p$ such that

$$(\mathcal{P}_p) \qquad J_p(u_p) = \inf_{u \in \mathcal{K}_p} J_p(u).$$

When treating this problem we adopt the methods of convexity as described in [ET, Chap. 1–4] and we can, without loss of generality, assume that $D = 1$. By [Z, Cor. 4.5.2, p. 195; Rem., p. 75] there is a constant $C > 0$ depending on p, n, and the Bessel capacity of Γ_0 (which is bounded away from zero as long as $\Gamma_0 \neq \emptyset$, since $p > n$) such that

$$\| u \|_p \leq C \| \nabla u \|_p, \quad \text{for all } u \in \mathcal{K}_p \tag{3.1}$$

and by the Rellich–Kondrachov theorem (see, e.g., [Ad, Thm. 6.2, p. 144]) we have

$$\| u \|_\infty \leq C \| u \|_{W^{1,p}(\Omega)}, \text{ for all } u \in W^{1,p}(\Omega)$$

for $p > n$. Thus, for $p > n$ the functional J_p is coercive over $\mathcal{K}_p$ since

$$J_p(u) \geq \frac{1}{p} \| \nabla u \|_p^p - \| u \|_\infty |\mu|(\Omega) \to \infty \text{ as } \| u \|_{W^{1,p}(\Omega)} \to \infty$$

and strictly convex. Let $\frac{1}{p} + \frac{1}{q} = 1$. The problem $\mathcal{P}_p$ has a dual convex problem which we will denote $\mathcal{P}_q^*$. The dual problem is obtained using the method described in [ET, pp. 60–61]. We begin with another definition.

DEFINITION 3. Let V be a Banach space with dual space V^* and dual pairing $\langle \cdot, \cdot \rangle$. For any convex functional $J : V \to \mathbf{R}$ define a conjugate functional $J^* : V^* \to \mathbf{R}$ by

$$J^*(v^*) = \sup_{u \in V} [\langle u, v^* \rangle - J(u)].$$

Define functionals $F : \mathcal{K}_p \to \mathbf{R}$, $G : \mathrm{L}^p(\Omega; \mathbf{R}^n) \to \mathbf{R}$ by

$$F(u) = - \int_\Omega u \, d\mu, \quad G(v) = \frac{1}{p} \int_\Omega f |v|^p \, dx,$$

so that $J_p(u) = F(u) + G(\nabla u)$. Using the above definition one easily derives

$$F^*(u^*) = \begin{cases} 0 & \text{if } u^* = -\mu \\ +\infty & \text{otherwise} \end{cases}, \quad G^*(r) = \frac{1}{q} \int_\Omega f^{1-q} |r|^q \, dx$$

for $u^* \in (\mathcal{K}_p)^*$ and $r \in (\mathrm{L}^p(\Omega; \mathbf{R}^n))^* = \mathrm{L}^q(\Omega; \mathbf{R}^n)$, respectively (see [ET, pp. 19–20, p. 81]). The operator $\nabla : \mathcal{K}_p \to \mathrm{L}^p(\Omega; \mathbf{R}^n)$ has an adjoint operator

$$\nabla^* : (\mathrm{L}^p(\Omega; \mathbf{R}^n))^* = \mathrm{L}^q(\Omega; \mathbf{R}^n) \to (\mathcal{K}_p)^*$$

defined by

$$\langle \nabla u, v \rangle = \langle u, \nabla^* v \rangle \quad \text{for every } u \in \mathcal{K}_p \text{ and } v \in \mathrm{L}^q(\Omega; \mathbf{R}^n).$$

According to [ET, p. 61], the dual problem, $\mathcal{P}_q^*$, can be written as follows: find $r_q \in \mathrm{L}^q(\Omega; \mathbf{R}^n)$ such that

$$\sup_{r \in \mathrm{L}^q(\Omega; \mathbf{R}^n)} [-F^*(\nabla^* r) - G^*(-r)]$$

is attained for $r = r_q$. For the supremum to be finite we must have that $F^*(\nabla^* r)$ is finite, i.e.,

$$\int_\Omega r \cdot \nabla u \, dx = - \int_\Omega u \, d\mu \tag{3.2}$$

for all $u \in \mathcal{K}_p$. Put $\mathcal{K}_q^* = \{r \in \mathrm{L}^q(\Omega; \mathbf{R}^n) : \text{ (3.2) holds for } r\}$. Then the final version of the problem $\mathcal{P}_q^*$ can be written as follows: find $r_q \in \mathcal{K}_q^*$ such that

$$(\mathcal{P}_q^*) \qquad \sup_{r \in \mathcal{K}_q^*} \left[-\frac{1}{q} \int_\Omega f^{1-q} |r|^q \, dx \right] = -\frac{1}{q} \int_\Omega f^{1-q} |r_q|^q \, dx\,.$$

Denote the infimum in $\mathcal{P}_p$ by $\inf \mathcal{P}_p$ and the supremum in $\mathcal{P}_q^*$ by $\sup \mathcal{P}_q^*$.

THEOREM 3.1. *The problems $\mathcal{P}_p$ and $\mathcal{P}_q^*$ are uniquely solvable and*

$$\inf \mathcal{P}_p = \sup \mathcal{P}_q^* .$$

Let u_p be the solution to $\mathcal{P}_p$ and r_q be the solution to $\mathcal{P}_q^*$. Then

$$r_q = -f|\nabla u_p|^{p-2}\nabla u_p \ \textit{a.e.} \quad \textit{and} \quad \int_\Omega f|\nabla u_p|^p\,dx = \int_\Omega f^{1-q}|r_q|^q\,dx = \int_\Omega u_p\,d\mu .$$

Proof. The existence and uniqueness of the two solutions, u_p and r_q, follow from [ET, Prop. 1.2, p. 35], since J_p and G^* are coercive and strictly convex. The relation $\inf \mathcal{P}_p = \sup \mathcal{P}_q^*$ follows from [ET, Thm. 4.2, p. 60]. By that theorem we also have the following extremality relation: Let u_p be the solution to $\mathcal{P}_p$ and r_q be the solution to $\mathcal{P}_q^*$. Then

$$\begin{aligned} 0 &= \langle r_q, \nabla u_p\rangle + G(\nabla u_p) + G^*(-r_q) \\ &\geq \int_\Omega f\left(\frac{1}{p}|\nabla u_p|^p + \frac{1}{q}\frac{|r_q|^q}{f^q} - \frac{|r_q|}{f}|\nabla u_p|\right)\,dx \geq 0. \end{aligned}$$

Here, we have used the fact that $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$ for all $a, b \in \mathbf{R}^+$, i.e., Young's inequality and the Schwarz inequality. Thus we have equality if and only if $|r_q| = f|\nabla u_p|^{p-1}$ and r_q is antiparallel to ∇u_p. Hence, $r_q = -f|\nabla u_p|^{p-2}\nabla u_p$ a.e. Since $\inf \mathcal{P}_p = \sup \mathcal{P}_q^*$ we immediately get

$$\frac{1}{p}\int_\Omega f|\nabla u_p|^p\,dx - \int_\Omega u_p\,d\mu = -\frac{1}{q}\int_\Omega f|\nabla u_p|^p\,dx$$

which completes the proof. □

By (3.2) and Theorem 3.1 we conclude that u_p is a solution to the equation

$$\int_\Omega f|\nabla u|^{p-2}\nabla u \cdot \nabla\varphi\,dx = \int_\Omega \varphi\,d\mu \tag{3.3}$$

for all $\varphi \in \mathcal{K}_p$. Hence if we restrict φ to C_0^∞ it follows that u_p is a solution to the problem

$$\begin{cases} -\operatorname{div}(f|\nabla u|^{p-2}\nabla u) = \mu \text{ in } \mathcal{D}'(\Omega), \\ u|_{\Gamma_0} = 0. \end{cases}$$

Actually, (3.3) contains implicitly a natural boundary condition and under suitable smoothness assumptions one can show that $\nabla u_p \cdot \nu = 0$ on Γ_1, where ν is the outward normal.

In what follows we will study the limiting behaviour of solutions of (3.3) as $p \to \infty$. The results we will prove in this section, apart from Theorem 3.5, are slight generalizations of Proposition 2.1 and Theorems 4.1 and 4.2 in [BDM] (Theorem 3.3 and Corollary 3.4 here). The proofs are analogous but will be given for completeness. We start with a lemma on the convergence of sequences of solutions to problem $\mathcal{P}_p$.

LEMMA 3.2. *For $p > n$, let u_p be the solution to problem $\mathcal{P}_p$. Then there exists a sequence $\{p_k\}_{k=1}^\infty : p_1 > n$, $p_k \nearrow \infty$ as $k \to \infty$ and a function $u_\infty \in \mathcal{K}_\infty$ such that*

(a) *$u_{p_k} \to u_\infty$ weakly in $W^{1,m}(\Omega)$ for every $m \geq 1$,*

(b) *$u_{p_k} \to u_\infty$ in $C^{0,\lambda}(\overline{\Omega})$ for any $\lambda \in (0,1)$,*

as $k \to \infty$. Further, $\| \nabla u_\infty \|_\infty \leq 1$.

Conversely, if $\{p_k\}_{k=1}^\infty$ is any sequence tending to ∞ such that $\{u_{p_k}\}_{k=1}^\infty$ converges weakly to u_∞ in $W^{1,m}(\Omega)$ for some $m > n$, then the above statements hold.

Remarks. (1) In the case where $f \equiv 1$ and $\Gamma_0 = \partial\Omega$, the above lemma can be found in [BDM] with essentially the same proof (see, e.g., pp. 26, 33–34).

(2) Observe that different subsequences may a priori yield different limits.

(3) By applying Mazur's lemma (see, e.g., [ET, p. 6]) and Cantor's diagonalization process, it is possible to obtain a sequence $\{v_k\}_{k=1}^\infty$ of convex combinations of $\{u_{p_k}\}_{k=1}^\infty$ such that $v_k \to u_\infty$ strongly in $W^{1,m}(\Omega)$ for every $m \geq 1$. Further, v_k is almost optimal for $\mathcal{P}_p$ in the following sense: Given $\varepsilon > 0$ there exists a K such that $|J_p(u_p) - J_p(v_k)| < \varepsilon$ for all p and k with $k > p > K$.

Proof. Let u_p be a solution to (3.3) and put $\mathcal{E}_p = \int_\Omega f|\nabla u_p|^p \, dx$. From Theorem 3.1 we then deduce that

$$\mathcal{E}_p = \inf_{r \in \mathcal{K}_q^*} \int_\Omega f^{1-q}|r|^q \, dx. \tag{3.4}$$

Now, let $n < s < p$, $\frac{1}{s} + \frac{1}{t} = 1$, and let r_t be the extremal for $\mathcal{P}_t^*$. Then, by Hölders inequality and (3.4)

$$\left(\frac{1}{|\Omega|}\mathcal{E}_p\right)^{\frac{1}{q}} \leq \left(\frac{1}{|\Omega|}\int_\Omega f^{1-q}|r_t|^q \, dx\right)^{\frac{1}{q}} \leq \left(\frac{1}{|\Omega|}\int_\Omega f^{1-t}|r_t|^t \, dx\right)^{\frac{1}{t}} = \left(\frac{1}{|\Omega|}\mathcal{E}_s\right)^{\frac{1}{t}}.$$

Hence, the function

$$p \mapsto \left(\frac{1}{|\Omega|}\mathcal{E}_p\right)^{\frac{p-1}{p}}$$

is monotonically decreasing and therefore $\lim_{p\to\infty} \mathcal{E}_p$ exists and is finite. See [BDM, p. 34].

Put $\mathcal{E}_\infty = \lim_{p\to\infty} \mathcal{E}_p$. By Theorem 3.1 we have that $J_p(u_p) = -\frac{1}{q}\mathcal{E}_p \to -\mathcal{E}_\infty$ as $p \to \infty$. Clearly, $\mathcal{E}_\infty > 0$. Let $m > n$ and put $G_m = \{u_p : \ p \geq m\} \subset W^{1,m}(\Omega)$. Further, there exists a constant C, independent of m and p, such that

$$\left(\frac{1}{|\Omega|}\int_\Omega f|\nabla u_p|^m \, dx\right)^{\frac{1}{m}} \leq \left(\frac{1}{|\Omega|}\mathcal{E}_p\right)^{\frac{1}{p}} \leq C, \quad \forall p \geq m. \tag{3.5}$$

Hence, from (3.1) it follows that G_m is bounded in $W^{1,m}(\Omega)$. By the Rellich–Kondrachov theorem (see, e.g., [Ad, Thm. 6.2, p. 144]) the embedding

$$W^{1,m}(\Omega) \hookrightarrow C^{0,\lambda}(\overline{\Omega}), \quad 0 < \lambda < 1 - \frac{n}{m}$$

is compact. Thus, since $W^{1,m}(\Omega)$ is weakly sequentially compact, we obtain a sequence $\{u_{p_k}\}_{k=1}^\infty$, $p_k \nearrow \infty$ as $k \to \infty$, that converges to a function u_∞ weakly in $W^{1,m}(\Omega)$ and strongly in $C^{0,\lambda}(\overline{\Omega})$, $0 < \lambda < 1 - \frac{n}{m}$, as $k \to \infty$. Furthermore, we claim that $u_{p_k} \to u_\infty$ weakly in $W^{1,m_1}(\Omega)$ for any $m_1 > m$. Suppose this is false. Then there is a functional $\ell \in (W^{1,m_1}(\Omega))^*$ and a subsequence $\{u_{p_{k_j}}\}_{j=1}^{\infty}$, $p_{k_1} > m_1$, such that $|\langle \ell, u_{p_{k_j}} - u_\infty \rangle| \geq \delta > 0$. By (3.5) this subsequence is bounded in $W^{1,m_1}(\Omega)$ and therefore contains a subsequence that converges to an element u_∞^* weakly in $W^{1,m_1}(\Omega)$. Since $u_{p_k} \to u_\infty$ uniformly and since this last subsequence is also a subsequence of $\{u_{p_k}\}_{k=1}^{\infty}$, we immediately get that $u_\infty = u_\infty^*$, which is a contradiction. Thus $u_{p_k} \to u_\infty$ weakly in $W^{1,m_1}(\Omega)$ for all $m_1 > m$. Since the embedding of $W^{1,m}(\Omega)$ into $W^{1,m_1}(\Omega)$ is continuous if $1 \leq m_1 < m$ (by Hölder's inequality) it then follows that we in fact have weak convergence in $W^{1,m_1}(\Omega)$ for any $m_1 \geq 1$.

Finally, by the weak convergence we get

$$\| \nabla u_\infty \|_m \leq \liminf_{k\to\infty} \| \nabla u_{p_k} \|_m \leq \liminf_{k\to\infty} |\Omega|^{\frac{1}{m}} \left(\frac{1}{|\Omega|} \mathcal{E}_{p_k} \right)^{\frac{1}{p_k}} = |\Omega|^{\frac{1}{m}}$$

for every $m > n$. Thus, $\| \nabla u_\infty \|_\infty \leq 1$, i.e., $u_\infty \in \mathcal{K}_\infty$. □

Note that we cannot expect to have convergence in $C^{0,1}$, as shown by the following example. Let $\Gamma_0 = \partial\Omega$, $f \equiv 1$, and $\mu = \delta_{x_0}$ for some $x_0 \in \Omega$. Then u_p is the Green function for $-\Delta_p$ with pole at x_0 in Ω (see §4). In a neighbourhood of x_0 we have by [S, Thm. 1, p. 79] that

$$|u_p(x) - u_p(x_0)| \approx |x - x_0|^{\frac{p-n}{p-1}}$$

which is not Lipschitz continuous for any $p > n$.

Before we proceed we need to study some geometric properties of Ω.

DEFINITION 4. The *ridge*, $\mathcal{R}(\Omega)$, is the set of points $x \in \Omega$ such that there exist $y_1, y_2 \in \partial\Omega$, $y_1 \neq y_2$ with $|x - y_1| = |x - y_2| = \operatorname{dist}(x, \partial\Omega)$. This set can also be characterized as the set of $x \in \Omega$ such that $\operatorname{dist}(\cdot, \partial\Omega)$ is not differentiable at x (see [EH, Thm. 3.3, p. 149]).

Let $x, y \in \overline{\Omega}$. By [B, p. 25, 5.18] there exists a shortest curve in $\overline{\Omega}$ joining x and y. We denote by $d_{\overline{\Omega}}(x,y)$ the distance within $\overline{\Omega}$ from x to y and define it by

$$d_{\overline{\Omega}}(x,y) = \inf\{\text{length}(\gamma) : \gamma \subset \overline{\Omega} \text{ is a curve joining } x \text{ and } y\}.$$

For a general domain in $\mathbf{R}^n$ one usually uses a different definition for $d_{\overline{\Omega}}(x,y)$ to make $d_{\overline{\Omega}}(x,y)$ comparable with the distance within Ω (see [BDM, pp. 22–23]). However, with the above definition they are in our case comparable, since Ω is a Lipschitz domain. We then define the distance from $x \in \Omega$ to Γ_0 as

$$d_{\Gamma_0}(x) = \inf_{y\in\Gamma_0} d_{\overline{\Omega}}(x,y).$$

We can thus choose a minimizing sequence of curves, γ_n, with terminal points $y_n \in \Gamma_0$. By [B, Thm. 5.16, p. 24] there is a subsequence γ_{n_k} and a curve $\gamma_0 \subset \overline{\Omega}$ with terminal point $y_0 \in \Gamma_0$ (since Γ_0 is closed) such that γ_{n_k} converges uniformly to γ_0 and $\text{length}(\gamma_0) \leq \liminf \text{length}(\gamma_{n_k})$, i.e., $d_{\Gamma_0}(x) = \text{length}(\gamma_0)$.

Let L_z be the set of shortest curves in $\overline{\Omega}$ connecting $z \in \operatorname{supp}\mu$ and Γ_0, and let E be the subset of $\overline{\Omega}$ covered by the curves in L_z as z varies over $\operatorname{supp}\mu$. Clearly $\operatorname{supp}\mu \subset E$. Further, E is closed. Indeed, let $\{x_k\}_{k=1}^{\infty} \subset E$ and suppose $x_k \to x$ as $k \to \infty$. By the definition of E, each x_k lies on a shortest curve joining a point $z_k \in \operatorname{supp}\mu$ and a point $y_k \in \Gamma_0$. Since $\operatorname{supp}\mu$ and Γ_0 are closed we can also assume that there are $z \in \operatorname{supp}\mu$ and $y \in \Gamma_0$ such that $z_k \to z$ and $y_k \to y$ as $k \to \infty$, possibly after choosing an appropriate subsequence of $\{x_k\}_{k=1}^{\infty}$. By the continuity of

$d_{\overline{\Omega}}$ it follows that $d_{\overline{\Omega}}(y_k, x_k) \to d_{\overline{\Omega}}(y, x)$ and $d_{\overline{\Omega}}(x_k, z_k) \to d_{\overline{\Omega}}(x, z)$. Since x_k lies on a shortest curve joining z_k and y_k, we have $d_{\overline{\Omega}}(y_k, z_k) = d_{\overline{\Omega}}(y_k, x_k) + d_{\overline{\Omega}}(x_k, z_k) \to d_{\overline{\Omega}}(y, x) + d_{\overline{\Omega}}(x, z)$. From this it follows that $d_{\overline{\Omega}}(z, y) = d_{\overline{\Omega}}(y, x) + d_{\overline{\Omega}}(x, z)$, i.e., $x \in E$ since we would otherwise have a contradiction to the assumption $x_k \in E$.

For reasons explained below, E will be called the set of uniqueness. We now formulate the main theorem of this section.

THEOREM 3.3. *Let $p > n$ and let u_p be the corresponding solution to problem $\mathcal{P}_p$. Let u_∞ be the limit element of some subsequence $\{u_{p_k}\}_{k=1}^{\infty}$ as in Lemma* 3.2. *Then the following hold:*

(1) $\| \nabla u_\infty \|_\infty = 1$ *and* $|u_\infty(x)| \le d_{\Gamma_0}(x)$ *for every* $x \in \Omega$.

(2) *Suppose μ is a positive measure. Then $u_\infty(x) = d_{\Gamma_0}(x)$ for every $x \in E$.*

(3) *If in addition to* (2), $\Gamma_0 = \partial\Omega$ *and* $\mathcal{R}(\Omega) \subset \operatorname{supp}\mu$, *then* $u_\infty(x) \equiv \operatorname{dist}(x, \partial\Omega)$. *In this case, the entire net $\{u_p\}_{p>n}$ converges to* $\operatorname{dist}(\cdot, \partial\Omega)$ *in* $W^{1,m}(\Omega)$ *for any $m \ge 1$ and in $C^{0,\lambda}(\overline{\Omega})$ for any $\lambda \in (0,1)$ as $p \to \infty$, i.e., without choosing subsequences.*

Remark. In the case $\Gamma_0 = \partial\Omega$ and $f \equiv 1$ on Ω the above theorem can be found in [BDM, Thms. 1.1, p. 33, 4.1, and 4.2, p. 42]. See also [K, Thm. 1, p. 5].

Proof. Let u_∞ be the limit of some subsequence, $\{u_{p_k}\}_{k=1}^{\infty}$. Using Theorem 3.1 and the fact that $u_{p_k} \to u_\infty$ in $C^{0,\lambda}(\overline{\Omega})$ as $k \to \infty$ we can conclude

$$\mathcal{E}_\infty = \lim_{k\to\infty} \mathcal{E}_{p_k}^{\frac{p_k-1}{p_k}} = \lim_{k\to\infty} \mathcal{E}_{p_k}^{-\frac{1}{p_k}} \int\limits_\Omega u_{p_k}\, d\mu = \int\limits_\Omega u_\infty\, d\mu. \tag{3.6}$$

Using (3.3), Theorem 3.1, and Hölder's inequality we get

$$\int\limits_\Omega \varphi\, d\mu \le \mathcal{E}_p^{\frac{1}{q}} \| \nabla\varphi \|_p, \quad \forall\varphi \in \mathcal{K}_\infty$$

and hence

$$\frac{1}{\| \nabla\varphi \|_\infty} \int\limits_\Omega \varphi\, d\mu \le \mathcal{E}_\infty, \quad \forall\varphi \in \mathcal{K}_\infty. \tag{3.7}$$

Since $u_\infty \in \mathcal{K}_\infty$ and $\| \nabla u_\infty \|_\infty \le 1$ by Lemma 3.2, it follows from (3.6) and (3.7) that

$$\| \nabla u_\infty \|_\infty = 1,$$

which in turn implies

$$|u_\infty(x)| \le d_{\Gamma_0}(x) \tag{3.8}$$

for every $x \in \Omega$. This proves (1).

Suppose now that μ is a positive measure on Ω. Since $d_{\Gamma_0} \in \mathcal{K}_\infty$ and $|\nabla d_{\Gamma_0}| = 1$ whenever it exists, we have that d_{Γ_0} is an admissible test function for $\mathcal{P}_p$ for any $p > n$. Hence, by Theorem 3.1 (recall $\frac{1}{p_k} + \frac{1}{q_k} = 1$)

$$J_{p_k}(u_{p_k}) = -\frac{1}{q_k} \int\limits_\Omega u_{p_k}\, d\mu \le J_{p_k}(d_{\Gamma_0}) = \frac{1}{p_k}|\Omega| - \int\limits_\Omega d_{\Gamma_0}\, d\mu,$$

since u_p is the solution to $\mathcal{P}_p$. Thus, using (3.6), we obtain as $p_k \to \infty$

$$\int_\Omega d_{\Gamma_0}\, d\mu \leq \int_\Omega u_\infty\, d\mu,$$

since $u_{p_k} \to u_\infty$ uniformly as $k \to \infty$, which by (3.8) yields

$$u_\infty(x) = d_{\Gamma_0}(x), \quad \text{for every } x \in \operatorname{supp}\mu. \tag{3.9}$$

This proves (2).

Suppose that $\Gamma_0 = \partial\Omega$ and that suppμ is a proper subset of $\overline{\Omega}$ and consider the Lipschitz extension problem as formulated in [Ar1]. Here, the given compact set $F = \operatorname{supp}\mu \cup \partial\Omega$ and the function to be extended, g, equals $\operatorname{dist}(x, \partial\Omega)$ on F. It is clear that both u_∞ and $\operatorname{dist}(\cdot, \partial\Omega)$ are solutions to this extension problem.

Suppose $\mathcal{R}(\Omega) \subset \operatorname{supp}\mu$. Then $\operatorname{dist}(\cdot, \partial\Omega) \in C^1(\Omega \setminus \operatorname{supp}\mu)$ and by [Ar1, Thm. 4, p. 555], it is the unique solution. Thus $u_\infty(x) = \operatorname{dist}(x, \partial\Omega)$ for every $x \in \Omega$.

If $\operatorname{supp}\mu = \overline{\Omega}$ then by (3.9) we have uniqueness of the limit function. That the net $\{u_p\}_{p>n}$ converges to u_∞ in $W^{1,m}(\Omega)$ for any $m \geq 1$ as $p \to \infty$ then follows by the same arguments as in [BDM, p. 36]. The proof of the Hölder convergence is analogous to the proof of Lemma 3.2(b). □

We can now state a sensible limiting problem $\mathcal{P}_\infty$ of the problems $\mathcal{P}_p$: find $u_\infty \in \mathcal{K}_\infty$ such that $\| \nabla u_\infty \|_\infty \leq 1$ and

$$(\mathcal{P}_\infty) \qquad \int_\Omega u_\infty\, d\mu = \sup_{\substack{u \in \mathcal{K}_\infty \\ \|\nabla u\|_\infty \leq 1}} \int_\Omega u\, d\mu.$$

Observe that the weight function f has "disappeared" in the limit process and does not occur in $\mathcal{P}_\infty$. We immediately get the following corollary.

COROLLARY 3.4. *Any function u_∞, as described above, is an extremal to the problem $\mathcal{P}_\infty$. If μ is positive then d_{Γ_0} is clearly an extremal.*

The uniqueness result of R. Jensen [J, Thm. 2.22, p. 70] and the fact that variational solutions are viscosity solutions of $\Delta_\infty u = 0$ [BDM, Prop. 2.2, p. 27] enables us to prove the following theorem.

THEOREM 3.5. *Let $\Gamma_0 = \partial\Omega$, $f \equiv 1$, and suppose μ is positive. Then the entire net $\{u_p\}_{p>n}$ converges to a unique limit element u_∞ in $C^{0,\lambda}(\overline{\Omega})$ for any $\lambda \in (0,1)$ as $p \to \infty$, i.e., without choosing subsequences.*

Proof. Let u_∞ be the limit element of some subsequence $\{u_{p_k}\}_{k=1}^\infty$ ($p_k \nearrow \infty$ as $k \to \infty$). Then $u_\infty(x)$=$\operatorname{dist}(x, \partial\Omega)$ for every $x \in E$ by Theorem 3.3. Let U be a component of $\Omega \setminus E$ and let v_p be the solution of the problem

$$\begin{cases} \Delta_p v = 0 & \text{in } U, \\ v = \operatorname{dist}(\cdot, \partial\Omega) & \text{on } \partial U. \end{cases}$$

Denote by v_∞ the unique variational solution corresponding to $\{v_p\}_{p>n}$. By the weak comparison principle for p-harmonic functions we then get

$$\| u_{p_k} - v_{p_k} \|_{\infty,U} = \| u_{p_k} - v_{p_k} \|_{\infty,\partial U} \to \| u_\infty - \operatorname{dist}(\cdot, \partial\Omega) \|_{\infty,\partial U} = 0,$$

i.e., $u_\infty = v_\infty$, and it follows that u_∞ is independent of the actual subsequence chosen. That $\{u_p\}_{p>n}$ has the stated convergence properties then follows exactly as in the proof of Theorem 2.5. □

Remark. Note that $\Delta_\infty u_\infty = 0$ on $\Omega \setminus E$ in both the variational and viscosity senses.

4. Green functions. Let $x_0 \in \Omega$, $\Gamma_0 = \partial\Omega$, $f \equiv 1$, and let $\mu = c_0\delta_{x_0}$. For any extremal $u_p \in W_0^{1,p}(\Omega)$ of problem $\mathcal{P}_p$, $p > n$, we then have that $-\Delta_p u_p = c_0\delta_{x_0}$. If $c_0 = 1$ then u_p is a so-called Green function of Δ_p with pole at x_0 in Ω. The Green function will be denoted $G_p(x; x_0, \Omega)$. For $p = 2$ we get the classical Green function. Thus, by Theorem 3.5, the family $\{G_p(\cdot\,; x_0, \Omega)\}_{p>n}$ has a unique limit as $p \to \infty$.

We now turn to the question of symmetry for the Green function. For $p = 2$ it is well known that the Green function is symmetric in any domain Ω, i.e., $G_2(x; y, \Omega) = G_2(y; x, \Omega)$. However, it is not as well known that if $p = n \geq 3$ and $\Omega = B_r(x_0)$ (ball with radius r and center at x_0) then the Green function is still symmetric! Symmetry holds because Möbius transformations preserve n-harmonic functions, i.e., if we perform a conformal change of coordinates, then the transformed function is still n-harmonic (see [HKM, p. 286]). (The author would like to thank Professor Peter Lindqvist, Trondheim, for pointing out this fact.) Indeed, let $r = 1$ and $x_0 = 0$. Then

$$G_n(x; 0, B_1(0)) = \omega_n^{\frac{1}{n-1}} \log \frac{1}{|x|},$$

where ω_n is the area of the unit sphere in $\mathbf{R}^n$. A mapping is said to be a Möbius transformation if it can be written as a finite composition of translations, homotheties, orthogonal linear transformations, and inversions in spheres. Verification of the statements below requires some very elementary but tedious calculations. These calculations will be omitted. To show the invariance of the n-harmonic functions under Möbius transformations we need only show that it holds when the transformation is an inversion in a sphere. This is done by changing coordinates in the integral on the left-hand side of (3.3). Now, the mapping

$$M_\xi(x) = \frac{(1 - |\xi|^2)(x - \xi) - |x - \xi|^2\xi}{1 - 2x \cdot \xi + |x|^2|\xi|^2}$$

is a Möbius transformation since it can be written as

$$M_\xi = S \circ H \circ T_2 \circ S \circ T_1,$$

where $S(x) = \frac{x}{|x|^2}$, $H(x) = (1 - |\xi|^2)x$, $T_1(x) = x - \xi$, and $T_2(x) = x - \frac{\xi}{1-|\xi|^2}$. Further,

$$|M_\xi(x)| = \frac{|(1 - |\xi|^2)(x - \xi) - |x - \xi|^2\xi|}{1 - 2x \cdot \xi + |x|^2|\xi|^2} = \frac{|x - \xi|}{|x|\left|\dfrac{x}{|x|^2} - \xi\right|}$$

if $x \neq 0$ and $|M_\xi(0)| = |\xi|$. Note that the square of the denominator of the right-hand side equals the denominator of the fraction in the middle. Hence $|M_\xi(x)| = |M_x(\xi)|$, M_ξ takes $B_1(0)$ onto $B_1(0)$, and $M_\xi(\xi) = 0$. Thus if $|\xi| < 1$ then

$$G_n(x; \xi, B_1(0)) = \omega_n^{\frac{1}{n-1}} \log \left(\frac{|x|}{|x - \xi|} \left| \frac{x}{|x|^2} - \xi \right| \right)$$

is the Green function for $B_1(0)$ with pole at ξ. Thus the Green function for the n-laplacian is symmetric. Below we will show that the Green function is not symmetric if p is sufficiently large.

THEOREM 4.1. *Let $B_1(0) = \{x \in \mathbf{R}^n : |x| < 1\}$. For every $x \in B_1(0)$, $x \neq 0$, there is a $p_x > n$ such that*

$$G_p(x; 0, B_1(0)) > G_p(0; x, B_1(0))$$

if $p > p_x$.

Remark. We chose $B_1(0)$ for simplicity. A similar result holds for a more general class of domains, e.g., convex domains. Indeed, let $x \in \Omega$ and suppose Ω is convex. Take a point $y \in \Omega$ such that y lies on a shortest ray from x to $\partial\Omega$. Then $G_p(y; x, \Omega) > G_p(x; y, \Omega)$ if p is sufficiently large.

Proof. Let $a \in B_1(0)$, put $y_0=(1,0,\dots,0)$, and let $\mathbf{R}^n \ni x=(x_1,\dots,x_n)$. The p-laplacian is invariant under rotations and therefore it is no restriction to assume that a lies on the positive x_1-axis. Further, $B_1(0)$ is convex and thus there is a family of affine functions, $\{f_p\}_{p>n}$, such that $f_p(-y_0)=0$, $f_p(0)=G_p(a; a, B_1(0))$ and since f_p is p-harmonic we get that $f_p(x) \geq G_p(x; a, B_1(0))$ by the weak comparison principle. By Theorem 3.3, $G_p(x; 0, B_1(0)) \to \operatorname{dist}(x, \partial B_1(0))$ in $W_0^{1,p}(\Omega)$ and $G_p(x; a, B_1(0)) \to \operatorname{dist}(x, \partial B_1(0))$ for x on the segment from a to y_0, as $p \to \infty$. Clearly,

$$G_p(0; a, B_1(0)) \leq f_p(0) = \frac{G_p(a; a, B_1(0))}{|a|+1} \to \frac{\operatorname{dist}(a, \partial B_1(0))}{|a|+1}, \quad \text{as } p \to \infty,$$

by construction and since

$$G_p(a; 0, B_1(0)) \to \operatorname{dist}(a, \partial B_1(0)) > \frac{\operatorname{dist}(a, \partial B_1(0))}{|a|+1}$$

the statement follows. □

We end this section with some observations that relate Green functions and the p-capacity of a condenser.

PROPOSITION 4.2. *Let $x_0 \in \Omega \subset \mathbf{R}^n$ and $p > n$. Then*

$$\operatorname{cap}_p(\{x_0\}, \Omega) = G_p(x_0; x_0, \Omega)^{1-p}.$$

Proof. Since p-capacitary functions are p-harmonic it follows by the comparison principle that the function

$$u(x) = \frac{G_p(x; x_0, \Omega)}{G_p(x_0; x_0, \Omega)}$$

is the capacitary function corresponding to p and $(\{x_0\}, \Omega)$. Thus

$$\begin{aligned}\operatorname{cap}_p(\{x_0\}, \Omega) &= \int_\Omega |\nabla u|^p \, dx = \frac{1}{G_p(x_0; x_0, \Omega)^p} \int_\Omega |\nabla G_p(x; x_0, \Omega)|^p \, dx \\ &= G_p(x_0; x_0, \Omega)^{1-p},\end{aligned}$$

since $-\Delta_p G_p = \delta_{x_0}$ and $G_p \in W_0^{1,p}(\Omega)$. □

LEMMA 4.3. *Let $\Omega \subset \mathbf{R}^n$ be a bounded domain with Lipschitz boundary $\partial\Omega$ and let $p > n$. There are constants $C_1, C_2 > 0$ such that*

$$C_1 \operatorname{dist}(x, \partial\Omega)^{n-p} \leq \operatorname{cap}_p(\{x\}, \Omega) \leq C_2 \operatorname{dist}(x, \partial\Omega)^{n-p}$$

for every $x \in \Omega$.

Proof. Put $B_r(y) = \{x \in \mathbf{R}^n : |x-y| < r\}$. By [HKM, Thm. 2.2, p. 28; Ex. 2.12, p. 35], we have $\operatorname{cap}_p(\{x\}, \Omega) \leq \operatorname{cap}_p(\{x\}, B_{\operatorname{dist}(x,\partial\Omega)}(x)) = C_2 \operatorname{dist}(x, \partial\Omega)^{n-p}$.

Take $x_0 \in \Omega$, put $d = \operatorname{dist}(x_0, \partial\Omega)$, and denote by u_0 the capacitary function corresponding to $(\{x_0\}, \Omega)$. Extend u_0 to all of $\mathbf{R}^n$ by $u_0(x) = 0$ if $x \notin \Omega$ and let

$v(x) = u_0(x_0 + \delta x)$. Then

$$\text{(4.1)} \quad \operatorname{cap}_p(\{x_0\},\Omega) = \int_{\Omega} |\nabla u_0|^p \, dx \geq \int_{\Omega \cap B_\delta(x_0)} |\nabla u_0|^p \, dx = \delta^{n-p} \int_{B_1(0)} |\nabla v|^p \, dx.$$

It only remains to estimate the last integral from below. Since $p > n$ we have by the Rellich–Kondrachov theorem (see, e.g., [Ad, Thm. 6.2, p. 144]) that the embedding $W^{1,p}(B_1(0)) \hookrightarrow C(B_1(0))$ is compact, i.e., there is a constant $A > 0$ such that

$$\text{(4.2)} \qquad 1 = \sup_{x \in B_1(0)} |v(x)| \leq A \parallel v \parallel_{W^{1,p}(B_1(0))} .$$

Now, choose $\delta = 2d$ and put $E = \{x \in B_1(0) : v(x) = 0\}$. Since $\partial\Omega$ is Lipschitz we claim that there is a constant $C > 0$ depending only on Ω such that

$$\text{(4.3)} \qquad C \leq |E| \leq |B_1(0)|$$

for all $x_0 \in \Omega$. To see this, put

$$h(x) = \frac{|B_{2d}(x) \cap \Omega^c|}{|B_{2d}(x)|}.$$

Since Ω is a bounded Lipschitz domain it has both the interior and the exterior cone property (see [Ad, pp. 66–67]). Here we shall make use of the exterior cone property, i.e., there is a finite cone S such that every $x \in \partial\Omega$ is the vertex of a cone $S_x \subset \Omega$ which is congruent to S. Now, for every $x \in \Omega$ we have that

$$1 \geq h(x) \geq \frac{|B_{2d}(x) \cap S_{\tilde{x}}|}{|B_{2d}(x)|} \geq \frac{|B_d(\tilde{x}) \cap S_{\tilde{x}}|}{|B_{2d}(x)|},$$

which is clearly bounded away from zero. Here $\tilde{x} \in \partial\Omega$ is such that $\operatorname{dist}(x, \partial\Omega) = |x - \tilde{x}|$.

From [Z, Thm. 4.4.2, p. 188] we have that for every $u \in W^{1,p}(B_1(0))$ that is equal to 0 on E

$$\parallel u \parallel_{L^p(B_1(0))} \leq C \parallel \nabla u \parallel_{L^p(B_1(0))},$$

where $C > 0$ depends only on p and $|E|$. Using (4.3) we then get that there is a $C' > 0$ independent of the particular E such that

$$\text{(4.4)} \qquad \parallel u \parallel_{W^{1,p}(B_1(0))} \leq C' \parallel \nabla u \parallel_{L^p(B_1(0))}$$

for all $u \in W^{1,p}(B_1(0))$ such that $u = 0$ on E.

Inserting (4.2) and (4.4) into (4.1) yields

$$\begin{aligned} \operatorname{cap}_p(\{x_0\},\Omega) &\geq (2\operatorname{dist}(x_0,\partial\Omega))^{n-p} \parallel \nabla v \parallel^p_{L^p(B_1(0))} \\ &\geq C''\operatorname{dist}(x_0,\partial\Omega)^{n-p} \parallel v \parallel^p_{W^{1,p}(B_1(0))} \\ &\geq C\operatorname{dist}(x_0,\partial\Omega)^{n-p} \end{aligned}$$

and we conclude $C_1\operatorname{dist}(x,\partial\Omega)^{n-p} \leq \operatorname{cap}_p(\{x\},\Omega) \leq C_2\operatorname{dist}(x,\partial\Omega)^{n-p}$. □

Combining Proposition 4.2 and Lemma 4.3 we immediately get the following result.

PROPOSITION 4.4. *Let $\Omega \subset \mathbf{R}^n$. Then there are constants $C_1, C_2 > 0$ such that*

$$C_1 \mathrm{dist}(x, \partial\Omega)^{\frac{p-n}{p-1}} \leq G_p(x; x, \Omega) \leq C_2 \mathrm{dist}(x, \partial\Omega)^{\frac{p-n}{p-1}}$$

for every $x \in \Omega$, which in turn implies

$$G_p(x; x, \Omega) \to 0 \text{ as } \mathrm{dist}(x, \partial\Omega) \to 0.$$

For further results on Green functions for a more general class of operators we refer to [Ho]. There, the case where $1 < p \leq n$ is considered but most of the results carry over to the case where $p > n$, with essentially the same proofs.

5. On the "thinning" behaviour of the extremals to the $\mathcal{P}_q^*$ problem. Below we will show that if the set of uniqueness $E \neq \overline{\Omega}$, then $r_q \to 0$ in the mean as $q \to 1+$ on sets that "stay away" from the set of uniqueness. As a corollary of this result (or rather, its proof) we will find that in this case the $\mathcal{P}_\infty$ problem has several different extremals.

A proof of the theorem below, in the case of two dimensions and a single Dirac measure, was suggested to the author by Professor Gunnar Aronsson, Linköping.

THEOREM 5.1. *Let μ be a positive Borel measure with $\mu(\Omega) < \infty$. Let $p > n$, $\frac{1}{p} + \frac{1}{q} = 1$, and let u_p and r_q denote the extremals of problems $\mathcal{P}_p$ and $\mathcal{P}_q^*$, respectively. Suppose that the set of uniqueness E is a proper subset of $\overline{\Omega}$ and let $F_\varepsilon = \{x \in \Omega : d_{\overline{\Omega}}(x, E) > \varepsilon\}$ for any $\varepsilon > 0$. Then*

$$\int_{F_\varepsilon} |r_q| \, dx = \int_{F_\varepsilon} f |\nabla u_p|^{p-1} \, dx \to 0, \quad \text{as } q \to 1+ \quad (p \to \infty)$$

for each $\varepsilon > 0$.

Remark. In some cases the sets F_ε can be chosen in a different way so that $\overline{F}_\varepsilon \cap E \neq \emptyset$, e.g., if $\mu = \sum_{k=1}^N c_k \delta_{x_k}$, $c_k > 0$ and $\Gamma_0 = \partial\Omega$. Then we can choose F_ε so that $\overline{F}_\varepsilon \cap E = \bigcup_{k=1}^N \{x_k\}$ and each component of $\Omega \setminus F_\varepsilon$ is the intersection of Ω and a cone with vertex at some x_k (see Fig. 1).

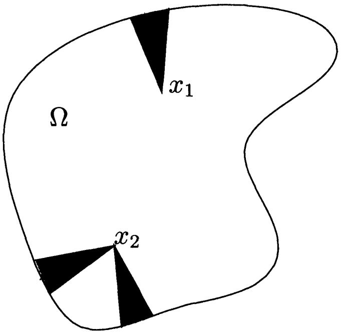


FIG. 1. *A typical example of the situation described in the remark above. The shaded areas indicate $\Omega \setminus F_\varepsilon$.*

Proof. The idea of this proof is to construct a function $\varphi_0 \in \mathcal{K}_\infty$ such that $|\nabla \varphi_0(x)| \leq 1 - \delta$ on F_ε for some $\delta > 0$, $\| \nabla \varphi_0 \|_{\infty,\Omega} = 1$, and $\varphi_0(x) = d_{\Gamma_0}$ for

$x \in \operatorname{supp}\mu$. Suppose that we have carried out this construction. From Lemma 3.1 and (3.2) it then follows that

$$-\int_\Omega r_q \cdot \nabla\varphi\, dx = \int_\Omega \varphi\, d\mu, \quad \forall\varphi \in \mathcal{K}_\infty.$$

Thus,

$$\int_\Omega \varphi_0\, d\mu = -\int_\Omega r_q \cdot \nabla\varphi_0\, dx \le \int_{\Omega\setminus F_\varepsilon} |r_q|\, dx + (1-\delta)\int_{F_\varepsilon} |r_q|\, dx \tag{5.1}$$

and since $\int_\Omega u_p\, d\mu \to \int_\Omega d_{\Gamma_0}\, d\mu$ by Theorem 3.3 as $p \to \infty$, it follows from (5.1) that (recall $0 < C \le f \le 1$)

$$\begin{aligned}\delta\int_{F_\varepsilon} |r_q|\, dx &\le \int_\Omega |r_q|\, dx - \int_\Omega \varphi_0\, d\mu \le |\Omega|^{\frac{1}{p}}\left(\int_\Omega f^{1-q}|r_q|^q\, dx\right)^{\frac{1}{q}} - \int_\Omega \varphi_0\, d\mu \\ &= |\Omega|^{\frac{1}{p}}\left(\int_\Omega u_p\, d\mu\right)^{\frac{p-1}{p}} - \int_\Omega d_{\Gamma_0}\, d\mu \to 0\end{aligned}$$

as $q \to 1+$ $(p \to \infty)$.

Put

$$\varphi_0(x) = d_{\Gamma_0}(x) \sup_{z\in\operatorname{supp}\mu} \frac{d_{\Gamma_0}(z)}{d_{\Gamma_0}(x) + d_{\overline{\Omega}}(x,z)} \tag{5.2}$$

(see Fig. 2). Now it only remains to prove that φ_0 has the properties stated. Clearly,

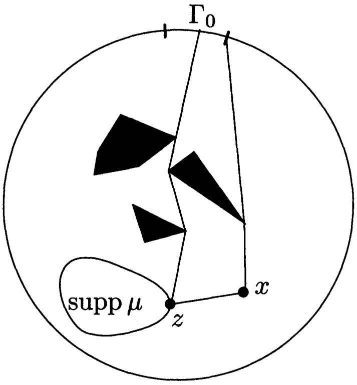


FIG. 2. *Illustration of the curves involved when calculating* $\varphi_0(x)$. *The shaded areas are parts of the complement of* Ω.

φ_0 is continuous on $\overline{\Omega}$ and $\varphi_0(x) \le d_{\Gamma_0}(x)$ with equality if and only if $x \in E \cup \Gamma_0$ since E is closed. Let $a, b \in \Omega$ and suppose $\varphi_0(a) > \varphi_0(b)$. Let $z_a \in \operatorname{supp}\mu$ be such that the supremum in the expression for φ_0 is attained. Then it follows from the triangle inequality that

$$0 < \varphi_0(a) - \varphi_0(b) \le d_{\Gamma_0}(z_a) \frac{d_{\overline{\Omega}}(a,b) d_{\overline{\Omega}}(b,z_a) + d_{\Gamma_0}(b) \left[d_{\overline{\Omega}}(b,z_a) - d_{\overline{\Omega}}(a,z_a)\right]}{\left[d_{\Gamma_0}(a) + d_{\overline{\Omega}}(a,z_a)\right] \left[d_{\Gamma_0}(b) + d_{\overline{\Omega}}(b,z_a)\right]}$$
$$\le d_{\overline{\Omega}}(a,b) \frac{d_{\Gamma_0}(z_a)}{d_{\Gamma_0}(a) + d_{\overline{\Omega}}(a,z_a)}$$

since $d_{\overline{\Omega}}(\cdot,\cdot)$ is a metric on Ω. Thus

$$\| \nabla\varphi_0 \|_{\infty,\Omega} = \sup_{a,b\in\Omega} \frac{|\varphi_0(a) - \varphi_0(b)|}{d_{\overline{\Omega}}(a,b)} = 1$$

and $\varphi_0 \in \mathcal{K}_\infty$. Finally,

$$\| \nabla\varphi_0 \|_{\infty,F_\varepsilon} = \sup_{a,b\in F_\varepsilon} \frac{|\varphi_0(a) - \varphi_0(b)|}{d_{\overline{\Omega}}(a,b)} \le 1 - \delta$$

where obviously

$$1 - \delta = \sup_{x\in F_\varepsilon} \sup_{z\in \operatorname{supp}\mu} \left[\frac{d_{\Gamma_0}(z)}{d_{\Gamma_0}(x) + d_{\overline{\Omega}}(x,z)} \right] < 1. \qquad \square$$

Now, we know from Corollary 3.4 that d_{Γ_0} is an extremal to the $\mathcal{P}_\infty$ problem. Clearly, the same holds for the function φ_0 defined in (5.2). Since $\varphi_0(x) < d_{\Gamma_0}(x)$ for every $x \in \Omega \setminus E$ (since E is closed) we conclude with the following corollary.

COROLLARY 5.2. *The $\mathcal{P}_\infty$ problem has a unique extremal if and only if $E = \overline{\Omega}$.*

REFERENCES

[Ad] R. A. ADAMS, *Sobolev Spaces,* Academic Press, London, New York, 1975.

[Ar1] G. ARONSSON, *Extension of functions satisfying Lipschitz conditions,* Ark. Mat., 6 (1967), pp. 551–561.

[Ar2] ———, *On the partial differential equation $u_x^2 u_{xx} + 2u_x u_y u_{xy} + u_y^2 u_{yy} = 0$,* Ark. Mat., 7 (1968), pp. 395–425.

[Ar3] ———, *Representation of a p-harmonic function near a critical point in the plane,* Manuscripta Math., 66 (1989), pp. 73–95.

[BDM] T. BHATTACHARYA, E. DIBENEDETTO, AND J. MANFREDI, *Limits as $p \to \infty$ of $\Delta_p u_p = f$ and related extremal problems,* Rend. Sem. Mat. Univ. Politec. Torino, (1989), pp. 15–68.

[B] H. BUSEMANN, *The Geometry of Geodesics,* Academic Press, New York, 1955.

[EH] W. D. EVANS AND D. J. HARRIS, *Sobolev embeddings for generalized ridged domains,* Proc. London Math. Soc., 54 (1987), pp. 141–175.

[ET] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems,* North–Holland, Amsterdam, Oxford, 1976.

[HKM] J. HEINONEN, T. KILPELÄINEN, AND O. MARTIO, *Nonlinear Potential Theory of Degenerate Elliptic Equations,* Oxford University Press, Oxford, 1993.

[Ho] I. HOLOPAINEN, *Nonlinear Potential Theory and Quasiregular Mappings on Riemannian Manifolds,* Ann. Acad. Sci. Fenn., Ser. A I Math. Dissertationes, 74, 1990.

[J] R. JENSEN, *Uniqueness of Lipschitz extensions: minimizing the sup norm of the gradient,* Arch. Rational Mech. Anal., 123 (1993), pp. 51–74.

[K] B. KAWOHL, *On a family of torsional creep problems,* J. Reine Angew. Math., 410 (1990), pp. 1–22.

[L] J. LEWIS, *Capacitary functions in convex rings,* Arch. Rational Mech. Anal., 66 (1977), pp. 201–224.

[M1] V. G. MAZ'YA, *Sobolev Spaces,* Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1985.

[M2] ———, *On the continuity at a boundary point of solutions of quasi-linear elliptic equations,* Vestnik Leningrad Univ. Math., 3 (1976), pp. 225–242 (in English); Vestnik Leningradskogo Universiteta, 25 (1970), pp. 42–55 (in Russian).

[S] J. SERRIN, *Applications of nonlinear partial differential equations in mathematical physics,* Proc. Sympos. Appl. Math, XVII (1965), pp. 68–88.

[T] P. TOLKSDORF, *Invariance properties and special structures near conical boundary points,* Lecture Notes in Mathematics 1121, Springer-Verlag, Berlin, Heidelberg, New York, 1985, pp. 308–318.

[Z] W. P. ZIEMER, *Weakly Differentiable Functions,* Springer-Verlag, Berlin, Heidelberg, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 27, No. 2, pp. 361–375, March 1996

DETERMINING LINEAR CRACKS BY BOUNDARY MEASUREMENTS: LIPSCHITZ STABILITY*

GIOVANNI ALESSANDRINI[†], ELENA BERETTA[‡], AND SERGIO VESSELLA[§]

Abstract. We consider the inverse boundary value problem of crack detection in a two-dimensional electrical conductor. We prove an estimate of Lipschitz type on the continuous dependence of an unknown linear crack from the boundary measurements.

Key words. inverse boundary value problem, stability, crack

AMS subject classifications. 35R30, 31A25

1. Introduction.

The inverse problem of detecting a crack in an electrically conducting body can be modeled as the determination of a curve σ in a planar domain Ω from boundary measurements of solutions u (potentials) to the problem

$$\Delta u = 0 \qquad \text{in } \Omega\backslash\sigma, \tag{1.1a}$$

$$u = \text{const.} \qquad \text{on } \sigma, \tag{1.1b}$$

$$\frac{\partial u}{\partial \nu} = \psi \qquad \text{on } \partial\Omega, \tag{1.1c}$$

when various profiles ψ (currents) are assigned.

Friedman and Vogelius have proved that a crack σ is uniquely determined when boundary measurements corresponding to two appropriate profiles ψ are known; see [F-V]. They also observe, by a duality argument, that a similar result holds when (1.1b) is replaced with

$$\frac{\partial u}{\partial \nu} = 0 \quad \text{on } \sigma. \tag{1.1b$'$}$$

When (1.1b) holds, the crack σ is said to be perfectly conducting, whereas in (1.1b′), it is said to be perfectly insulating. They also address the stability issue, discuss the relevance for the actual reconstruction of the crack, and give some initial results in this direction.

A stability estimate for perfectly conducting cracks has been obtained in [A1] and generalized by Diaz Valenzuela to the perfectly insulating case [DV]. Unfortunately, such estimates are of logarithmic type; see [A2] for a partial improvement. In [A1], [DV], and [A2], bounds on the smoothness and the size of the unknown crack are assumed, but they involve only a finite number of derivatives of the crack parametrization. One can expect that better stability estimates might be obtained when stronger

* Received by the editors February 4, 1994; accepted for publication (in revised form) August 2, 1994. This research was partially supported by the Fondi Ministero dell' Universitá della Ricerca Scientifica e Tecnologica and the Consiglio Nazionale della Richerche.

[†] Dipartimento di Scienze Matematiche, Universitá di Trieste, 34100, Trieste, Italy. The research of this author was partially supported by the Ministero Ricerca Scientifica e Tecnologica.

[‡] Istituto per le Applicazioni del Calcolo "Mauro Picone," Viale del Policlinico 137, 00161 Rome, Italy. The research of this author was supported by the Consiglio Nazionale delle Ricerche.

[§] Dipartimento di Matematica per le Decisioni Economiche, Finanziarie, Attuariali, e Sociali, Universitá di Firenze, 6/17, 50134 Firenze, Italy.

a priori information is prescribed. Linear cracks have been the object of interesting and successful numerical reconstruction algorithms by Santosa, Vogelius, Liepa, and Bryan [S-V], [L-S-V], [B-V]. Therefore, it seems interesting to complement such procedures with a theoretical study of the stability in this special case.

In this paper we prove a Lipschitz-type continuous dependence of the crack σ from the $L^2(\Gamma)$ traces of two solutions u^1 and u^2 of (1.1) when two appropriate profiles ψ^1 and ψ^2 are assigned and are as follows. Fixing three distinct points P_0, P_1, and P_2 on $\partial\Omega$, ψ^1 and ψ^2 are the two-electrode current configurations with electrodes at P_0, P_1, and P_0, P_2, respectively (see (2.3)–(2.5)). Here Γ is a fixed portion of the boundary $\partial\Omega$.

We have chosen to treat the perfectly conducting case as in [A1] and [A2] just for the sake of convenience, but we wish to stress that a completely analogous result could be obtained for the perfectly insulating case due to the results in [DV]. Let us also emphasize the fact that the measurements u^1 and u^2 need not to be taken on all of the boundary $\partial\Omega$, but just on a portion Γ of it.

Let us illustrate the main line of our argument. Considering $u^3 = u^2 - u^1$, that is, the solution with electrodes at P_1, P_2, we find that there exists at least one index $j = 1, 2, 3$ such that, setting $u = u^j$, $\partial_z u$ has a singularity of the type $z^{-1/2}$ at both endpoints V and W of σ. This fact suffices already to prove the unique determination of σ from the boundary data. This is a consequence of the unique harmonic continuation of the Cauchy data u, $\frac{\partial u}{\partial \nu}$ on Γ.

In order to prove our Lipschitz-type estimate, we study the Frechet derivative u' of $u = u(z)$ as a function of the crack σ. We write the asymptotic formula of $u' = u'(z)$ when z is near the endpoints of σ, for instance,

$$u'(z)\delta\sigma = Re(\alpha(z - V)^{-1/2}\delta V) + \text{lower-order terms} \quad \text{as } z \to V. \tag{1.2}$$

Here δV denotes the variation of the endpoint V and $\alpha \neq 0$ is a complex number. From (1.2) we deduce a lower bound on the derivative of the map

$$\sigma \to (u^1, u^2) \in L^2(\Gamma) \times L^2(\Gamma)$$

(see Proposition 4.1). By coupling this lower bound with upper bounds on the second directional derivative of the same map, we are able to estimate the derivative of the inverse mapping. This estimate, combined with the general stability estimate in [A1] (slightly adapted to our setting; see Lemma 4.6), yields the desired Lipschitz stability result.

Let us recall that Friedman and Vogelius already studied the stability problem for linear cracks and they proved a Lipschitz-type bound for the line containing the crack. Unfortunately, their approach does not give sufficient information about the location of the endpoints of the crack along the line. This difficulty is circumvented here by evaluating formula (1.2) on appropriate points z near V and also in the case when δV is parallel to σ.

The plan of the paper is as follows: In §2 we state our assumptions and the main theorem (Theorem 2.1). In §3 we describe, through a sequence of lemmas, the asymptotic behaviour of the solutions $u^j, j = 1, 2, 3$, near the endpoints of the crack. The main results of this section are contained in Proposition 3.4. Section 4 contains the main body of the proof of Theorem 2.1. The final section, §5, is devoted to the proof of two auxiliary results, Lemma 4.6 and Proposition 4.5.

2. The main result. Given positive constants L_1, L_2, M, and α, $0 < \alpha < 1$, which we shall name a priori data, we suppose that Ω is a bounded simply connected domain in $\mathbf{R}^2$ satisfying the following:

$$\text{perimeter of } \Omega \leq L_1, \tag{2.1a}$$

$$\begin{gathered}\forall z \in \partial\Omega \text{ there exist two circles of radius} L_2 \text{ tangent in } z, \\ \text{the first contained in } \overline{\Omega} \text{ and the second in } C\Omega.\end{gathered} \tag{2.1b}$$

If $z = z(s)$ is the arclength parametrization of $\partial\Omega$, we have

$$||z||_{C^{2,\alpha}} \leq M. \tag{2.1c}$$

We consider the class Σ of linear cracks in Ω which is made of the linear segments $\sigma \subset \Omega$ satisfying the following:

$$\text{length of } \sigma \geq L_2, \tag{2.2a}$$

$$\text{dist}(z, \partial\Omega) \geq L_2, \quad \forall z \in \sigma. \tag{2.2b}$$

Given three points $P_0, P_1, P_2 \in \partial\Omega$ such that

$$|P_i - P_j| \geq L_2 \quad \forall i, j, i \neq j, \tag{2.3}$$

we consider three smooth nonnegative functions η_0, η_1, and η_2 on $\partial\Omega$ satisfying

$$\int_{\partial\Omega} \eta_i \, ds = 1, \quad i = 0, 1, 2, \tag{2.4a}$$

and

$$\text{supp } \eta_i \subset \partial\Omega \cap B_h(P_i), \tag{2.4b}$$

where $B_h(P_i)$ denotes the ball centered at P_i with radius h, $0 < h < \frac{L_2}{2}$.

We denote

$$\begin{aligned} \psi^1 &= \eta_0 - \eta_1, \\ \psi^2 &= \eta_0 - \eta_2, \\ \psi^3 &= \eta_1 - \eta_2 = (\psi^2 - \psi^1). \end{aligned} \tag{2.5}$$

Let $\Gamma \subset \partial\Omega$ be a simple arc such that

$$\text{length of } \Gamma \geq L_2. \tag{2.6}$$

Given two cracks $\sigma_0, \sigma_1 \in \Sigma$, we consider $u_i^j \in H^1(\Omega)$, $i = 0, 1$, $j = 1, 2, 3$, as the unique solution of the following boundary value problem:

$$\Delta u_i^j = 0 \qquad \text{in } \Omega \backslash \sigma_i, \tag{2.7a}$$

$$u_i^j = \text{const.} \qquad \text{on } \sigma_i, \tag{2.7b}$$

$$\frac{\partial u_i^j}{\partial \nu} = \psi^j \qquad \text{on } \partial\Omega, \tag{2.7c}$$

$$\int_{\partial\Omega} u_i^j \, ds = 0. \tag{2.7d}$$

Here ν denotes the exterior unit normal to $\partial\Omega$. Notice that the constant value in (2.7b) is also unknown, but it is uniquely determined in view of (2.7d). Observe also that

$$u_i^3 = u_i^2 - u_i^1. \tag{2.8}$$

THEOREM 2.1. *Assume conditions* (2.1)–(2.6). *There exist* h_0, $C > 0$, *depending on the a priori data only, such that if* $h \leq h_0$ *and* $\sigma_0, \sigma_1 \in \Sigma$, *then we have*

$$d(\sigma_0, \sigma_1) \leq C \sum_{j=1}^{2} ||u_0^j - u_1^j||_{L^2(\Gamma)}. \tag{2.9}$$

Here $d(.,.)$ denotes the Hausdorff distance between bounded closed sets of the plane. If we denote by V_i and W_i the endpoints of σ_i, $i = 0, 1$, then we see that

$$d(\sigma_0, \sigma_1) = \min\{\max\{|V_0 - V_1|, |W_0 - W_1|\},\ \max\{|V_0 - W_1|, |W_0 - V_1|\}\}.$$

Hence, up to a renaming of the endpoints, we can assume that

$$d(\sigma_0, \sigma_1) = |V_0 - V_1|,$$

and

$$|W_0 - W_1| \leq d(\sigma_0, \sigma_1).$$

Remark. Observe that the hypothesis $h \leq h_0$ means that we require the electrodes to be concentrated. On the other hand, the constant C does not diverge as $h \to 0$ and thus, in the limit, (2.9) remains valid when the functions η_j are replaced with the Dirac deltas.

3. Some preliminary lemmas. When it is convenient we shall use the usual identification of $\mathbf{R}^2$ with the complex plane: $z = x + iy$, $\partial_z = \frac{1}{2}(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y})$. Also, we shall choose an orientation of coordinates in such a way that $W_0 - V_0$ is real and negative.

Let us denote

$$\begin{aligned} \tilde{\psi}^1 &= \delta_{P_0} - \delta_{P_1}, \\ \tilde{\psi}^2 &= \delta_{P_0} - \delta_{P_2}, \\ \tilde{\psi}^3 &= \delta_{P_1} - \delta_{P_2}, \end{aligned} \tag{3.1}$$

and let us name v_i^j the solution to (2.7) when ψ^j is replaced with $\tilde{\psi}^j$. Set $c_i^j = v_i^j|\sigma_i$.

From [A1, Thm. 2.1] we know that the level line $\{v_i^j = c_i^j\}$ in $\Omega\backslash\sigma_i$ is composed of two simple arcs, each of which has one endpoint on $\partial\Omega$ and the other on σ_i. The latter ones are called branching points. We also know that the branching points are distinct when considered as points on $\tilde{\sigma}_i$, the closed curve obtained by glueing at the endpoints two copies of σ_i (that is, if we distinguish one-sided limit points on σ_i). The following lemma gives a new formulation to arguments already used in [A1, Lem. 4.1].

LEMMA 3.1. *There exists* $C > 0$, *depending on the a priori data only, and at least one index* $j = 1, 2, 3$ *such that the branching points on* v_0^j *have distance larger than or equal to* C *from both the endpoints of* σ_0.

Proof. Let B_l^j, $l = 1, 2$, be the branching points of v_0^j on $\tilde{\sigma}_0$. From [A1, Thm. 2.1] we have that, up to a rearrangement of the points P_0, P_1, and P_2, the branching points have the following circular ordering on $\tilde{\sigma}_0$,

$$\cdots < B_1^1 < B_1^2 < B_1^3 < B_2^1 < B_2^2 < B_2^3 < \cdots,$$

and that the distance along $\tilde{\sigma_0}$ between any two consecutive branching points dominates a positive constant K, only depending on the a priori data.

Now we show that there exists $j = 1, 2, 3$ such that

$$|V_0 - B_l^j|, |W_0 - B_l^j| \geq \frac{K}{2}, \quad \forall l = 1, 2.$$

Suppose that for some $h = 1, 2, 3$, and $m = 1, 2$, we have

$$|V_0 - B_m^h| < \frac{K}{2};$$

therefore, we get

$$|V_0 - B_l^j| \geq \frac{K}{2} \quad \forall j \neq h, \ l = 1, 2.$$

Next, if for some $k \neq h$ and some $m = 1, 2$, we have

$$|W_0 - B_m^h| < \frac{K}{2},$$

then we are left with one index j, $j \neq h, k$, such that

$$|W_0 - B_l^j| \geq \frac{K}{2} \quad \forall l = 1, 2,$$

and our thesis follows with $C = \frac{K}{2}$. □

LEMMA 3.2. *Let $j = 1, 2, 3$ be as in Lemma* 3.1. *There exist ρ, $C > 0$, depending on the a priori data only, such that*

$$|\nabla v_0^j(z)| \geq C|z - V_0|^{-1/2}, \quad \forall z \in B_\rho(V_0) \backslash \sigma_0. \tag{3.2}$$

Proof. See [A1, Lem. 4.2].

LEMMA 3.3. *Let C, $\rho > 0$ be fixed. Let $u, v \in H^1(B_\rho(V_0))$ be harmonic in $B_\rho(V_0) \backslash \sigma_0$ and constant on $\sigma_0 \cap B_\rho(V_0)$.*

There exist $A, B, \delta_0 > 0$, depending on C and ρ only, such that, if we have

$$|\nabla v(z)| \geq C|z - V_0|^{-1/2}, \quad \forall z \in B_\rho(V_0) \backslash \sigma_0 \tag{3.3}$$

and

$$||v - u||_{H^1(B_\rho(V_0))} \leq \delta_0, \tag{3.4}$$

then we have

$$\partial_z u(z) = i\alpha(z - V_0)^{-1/2} + R(z), \quad \forall z \in B_\rho(V_0) \backslash \sigma_0, \tag{3.5}$$

where

$$\alpha \in \mathbf{R}, \qquad |\alpha| \geq A, \tag{3.6}$$

$$|\alpha| + |R(z)| \leq B||u||_{H^1(B_\rho(V_0))}, \quad \forall z \in B_\rho(V_0) \backslash \sigma_0. \tag{3.7}$$

We defer the proof of Lemma 3.3 until after Proposition 3.4, which contains the main results of this section.

Remark. Let us observe that an analogous statement could be obtained when V_0 is replaced with W_0.

PROPOSITION 3.4. *There exist constants A, B, ρ, $h_0 > 0$, depending on the a priori data only, and an index $j = 1, 2, 3$, such that if $0 < h < h_0$ and u_0^j is the solution of* (2.7) *when $i = 0$, then we have*

$$\partial_z u_0^j(z) = i\alpha(z - V_0)^{-1/2} + R(z), \quad \forall z \in B_\rho(V_0)\backslash\sigma_0, \tag{3.8}$$

with $\alpha \in \mathbf{R}$ and

$$|\alpha| \geq A; \ |\alpha| + |R(z)| \leq B||u||_{H^1(B_\rho(V_0))}, \quad \forall z \in B_\rho(V_0)\backslash\sigma_0. \tag{3.9}$$

Proof. We can fix two open subsets U_0 and U_1 of Ω such that

$$\sigma \subset U_0 \subset\subset U_1 \subset\subset \Omega, \qquad \forall \sigma \in \Sigma.$$

The weak formulation of (2.7) for $u = u_0^j - v_0^j$ gives

$$u \in H^1(\Omega), \qquad u = \text{const.} \quad \text{on } \sigma_0,$$

$$\int_\Omega \nabla u \cdot \nabla \phi = 0 \quad \forall \phi \in H_0^1(\Omega) \text{ such that } \phi = \text{const. on } \sigma_0.$$

Notice that the proof of the Caccioppoli inequality and of the weak maximum principle follow as in the usual case when no constraint is imposed on σ_0. Therefore, we have

$$\int_{U_0} |\nabla(u_0^j - v_0^j)|^2 \leq C \int_{U_1} |u_0^j - v_0^j|^2$$

and

$$\max_{U_1} |u_0^j - v_0^j| \leq \max_{\partial U_1} |u_0^j - v_0^j|$$

and hence

$$||u_0^j - v_0^j||_{H^1(U_0)} \leq C \max_{\partial U_1} |u_0^j - v_0^j|,$$

and by interior estimates in H^s-spaces,

$$\max_{\partial U_1} |u_0^j - v_0^j| \leq C||u_0^j - v_0^j||_{H^{1/2}(\Omega\backslash U_0)}.$$

This last norm can be bounded by the $H^{-1}(\partial\Omega)$ norm of the Neumann data for $u_0^j - v_0^j$. Therefore, we get

$$||u_0^j - v_0^j||_{H^1(U_0)} \leq C||\psi^j - \tilde{\psi}^j||_{H^{-1}(\partial\Omega)}.$$

Recalling (2.5) and (3.1), we have

$$||\psi^j - \tilde{\psi}^j||_{H^{-1}(\partial\Omega)} \leq 2 \max_{j=1,2,3} ||\eta_j - \delta_{P_j}||_{H^{-1}(\partial\Omega)}.$$

Now we see that

$$||\eta_j - \delta_{P_j}||_{H^{-1}(\partial\Omega)} = \sup_{||\phi||_{H^1(\partial\Omega)}=1} \left| \int_{\partial\Omega} \eta_j \phi \, ds - \phi(P_j) \right|.$$

By (2.4) and the standard estimate

$$||\phi||_{C^{1/2}(\partial\Omega)} \leq C||\phi||_{H^1(\partial\Omega)},$$

we obtain

$$||\eta_j - \delta_{P_j}||_{H^{-1}(\partial\Omega)} \leq Ch^{1/2}.$$

Therefore, we get

$$||u_0^j - v_0^j||_{H^1(U_0)} \leq Ch^{1/2},$$

and, by similar arguments,

$$||u_0^j||_{H^1(U_0)} \leq C.$$

Now we pick $j = 1, 2, 3$ as in Lemmas 3.1 and 3.2. The proposition follows from Lemma 3.3. □

Proof of Lemma 3.3. Without loss of generality, we may set $u = v = 0$ on $\sigma_0 \cap B_\rho(V_0)$. We may assume $V_0 = 0$; hence $\sigma_0 \subset \{y = 0, x < 0\}$.

Consider $\xi + i\eta = \zeta = \sqrt{z}$ as the analytic branch of the square root which maps $\mathbf{C}\backslash\{y = 0, x < 0\}$ onto $\{\xi = Re\zeta > 0\}$. That is, in polar coordinates $z = re^{i\theta}$, $|\theta| < \pi$, $\zeta = r^{1/2}e^{i\theta/2}$. By abuse of notation, we set $u(\zeta) = u(z)$ and $v(\zeta) = v(z)$.

By the conformal invariance of the Dirichlet integral, we get

$$\int_{|\zeta|<\rho^{1/2},\xi>0} |\partial_z(u - v)|^2 \, d\xi d\eta < \delta_0^2,$$

and we have $u = v = 0$ when $\xi = 0$. Let us continue u and v harmonically to $\{\xi < 0\}$ by an odd reflection. By standard interior bounds, we have

$$|\partial_\zeta(u - v)| \leq K\delta_0, \quad \forall\zeta \in B_{(\rho/2)^{1/2}}(0),$$

and therefore, when $\delta_0 < \frac{C}{2K}$

$$|\partial_\zeta u| \geq \frac{C}{2} \quad \forall\zeta, |\zeta| < \rho^{1/2}$$

and

$$\max_{|\zeta|<(\rho/2)^{1/2}} |\partial_\zeta u| \leq K \left(\frac{1}{\rho} \int_{|\zeta|<\rho^{1/2}} |\partial_\zeta u|^2 \right)^{1/2}.$$

Notice that

$$\frac{\partial u}{\partial \xi}(0) = 0, \qquad \frac{C}{2} \leq \left| \frac{\partial u}{\partial \eta}(0) \right| \leq K||\partial_\zeta u||_{L^2(B_\rho(0))}.$$

Therefore, the Taylor formula gives

$$\partial_\zeta u(\zeta) = -\frac{1}{2} i \frac{\partial u}{\partial \eta}(0) + R(\zeta), \quad \forall\zeta \in B_{(\rho/2)^{1/2}}(0)$$

with

$$|R(\zeta)| \leq |\zeta| \max_{|\zeta|<(\rho/2)^{1/2}} |\partial_\zeta^2 u| \leq |\zeta| \left(\frac{C}{\rho^2} \int_{|\zeta|<\rho^{1/2}} |\partial_\zeta u|^2 \right)^{1/2}.$$

Hence, setting $\alpha = -\frac{1}{4}\frac{\partial u}{\partial \eta}(0)$, by the chain rule $\partial_z u = \frac{1}{2\sqrt{z}}\partial_\zeta u$, we have

$$\left| \partial_z u - \frac{i\alpha}{\sqrt{z}} \right| \leq \frac{C}{\rho}||u||_{H^1(B_\rho(0))} \quad \forall z, |z| < \frac{\rho}{2},$$

and (3.5)–(3.7) follow. □

4. Proof of Theorem 2.1. From now on, we shall restrict our attention to the solutions u_i^j to (2.7) when j is the index found in Proposition 3.4. Therefore, we shall drop the superscript j from ψ^j and u_i^j.

Denote by σ_t, $0 \le t \le 1$, the line segment with endpoints

$$V_t = V_0 + t(V_1 - V_0), \qquad W_t = W_0 + t(W_1 - W_0).$$

Denote by $u_t \in H^1(\Omega)$, $0 \le t \le 1$, the weak solution to

$$\Delta u_t = 0 \quad \text{in } \Omega \backslash \sigma_t, \tag{4.1a}$$

$$u_t = \text{const.} \quad \text{on } \sigma_t, \tag{4.1b}$$

$$\frac{\partial u_t}{\partial \nu} = \psi \quad \text{on } \partial\Omega, \tag{4.1c}$$

$$\int_{\partial\Omega} u_t \, ds = 0. \tag{4.1d}$$

Let us also consider $v_t \in H^1(\Omega)$, given as the solution to

$$\Delta v_t = 0 \quad \text{in } \Omega \backslash \sigma_t, \tag{4.2a}$$

$$v_t = 0 \quad \text{on } \sigma_t, \tag{4.2b}$$

$$\frac{\partial v_t}{\partial \nu} = \psi \quad \text{on } \partial\Omega. \tag{4.2c}$$

If we set $c_t = u_t|\sigma_t$, then we have

$$u_t = v_t + c_t, \; c_t = -\frac{1}{|\partial\Omega|} \int_{\partial\Omega} v_t \, ds. \tag{4.3}$$

LEMMA 4.1. *There exist $C, \delta_0 > 0$, depending on the a priori data only, such that if $d(\sigma_0, \sigma_1) \le \delta_0$, then there exists a one parameter family of invertible C^∞ mappings $\zeta_t : \Omega \to \Omega$, $0 \le t \le 1$, which satisfies the following properties:*

(i) *There exists a neighborhood U of $\cup_{0 \le t \le 1} \sigma_t$, $U \subset\subset \Omega$, such that*

$$U \ni z \to \zeta_t(z) \in \Omega$$

is a complex linear function for all t, $0 \le t \le 1$.

(ii)

$$\zeta_t(\sigma_t) = \sigma_0 \quad \forall t, 0 \le t \le 1.$$

(iii) *There is a neighborhood V of $\partial\Omega$, $V \subset\subset \overline{\Omega} \backslash U$, such that*

$$\zeta_t(z) = z \quad \forall z \in V, \; \forall t, 0 \le t \le 1. \tag{4.4}$$

(iv) *Denoting $\zeta_t(z) = \xi_t(x, y) + i\eta_t(x, y)$, we have*

$$\left| \frac{\partial(\xi_t, \eta_t)}{\partial(x, y)} - I \right| \le C d(\sigma_0, \sigma_1) t, \; \forall z \in \Omega, \quad \forall t \in [0, 1]. \tag{4.5}$$

(v) *ζ_t is twice continuously differentiable with respect to t, and we have*

$$|\zeta_t'(z)| \le C d(\sigma_0, \sigma_1), \tag{4.6a}$$

$$|\zeta_t''(z)| \le C d^2(\sigma_0, \sigma_1). \tag{4.6b}$$

Proof. Let T_1 be the linear mapping

$$T_1 z = \frac{V_1 - W_1}{V_0 - W_0} z + \frac{V_0 W_1 - V_1 W_0}{V_0 - W_0}. \tag{4.7}$$

We have $T_1\sigma_0 = \sigma_1$. Hence, if we set $T_t = I + t(T_1 - I)$, we obtain $T_t^{-1}\sigma_t = \sigma_0$ for every $t \in [0, 1]$. Observe that

$$||T_t^{-1} - I|| \leq Cd(\sigma_0, \sigma_1)t \tag{4.8a}$$

and

$$\frac{d}{dt}T_t^{-1} = T_t^{-1}(I - T_1)T_t^{-1}, \quad 0 \leq t \leq 1. \tag{4.8b}$$

Next, fixing U and V in such a way that

$$\cup_{0 \leq t \leq 1}\sigma_t \subset U \subset\subset \Omega, \qquad \partial\Omega \subset V \subset\subset \overline{\Omega}\backslash U,$$

consider a $C_0^\infty(\Omega)$ function ϕ such that $0 \leq \phi \leq 1$ in Ω, $\phi = 0$ in V, and $\phi = 1$ in U. We define

$$\zeta_t(z) = (1 - \phi(z))z + \phi(z)T_t^{-1}z. \tag{4.9}$$

The estimates (4.5)–(4.6) follow by direct computation. The invertibility of ζ_t follows from (4.5) provided that $d(\sigma_0, \sigma_1) \leq \delta_0$ with $\delta_0 > 0$ sufficiently small depending on the a priori data only. Statements (i)–(iii) follow from the construction. □

Let us introduce $w_t = v_t \circ \zeta_t^{-1}$; setting $J_t = \frac{\partial(\xi_t, \eta_t)}{\partial(x,y)}$ and $A_t = \frac{J_t J_t^T}{\det J_t}$, we obtain from (4.2) that $w_t \in H^1(\Omega)$ is the unique solution to the following problem:

$$\text{div}(A_t \nabla w_t) = 0 \quad \text{in } \Omega\backslash\sigma_0, \tag{4.10a}$$

$$w_t = 0 \quad \text{on } \sigma_0, \tag{4.10b}$$

$$\frac{\partial w_t}{\partial \nu} = \psi \quad \text{on}\partial\Omega. \tag{4.10c}$$

Notice that, if the hypotheses of Lemma 4.1 are fulfilled, then the matrix A_t satisfies a uniform ellipticity condition which depends on the a priori data only. Let us denote by $K = \frac{\Omega}{\text{V}}$ (V as in (iii) of Lemma 4.1).

LEMMA 4.2. *Let $d(\sigma_0, \sigma_1) \leq \delta_0$. Then the solution $w_t \in H^1(\Omega)$ to* (4.10) *satisfies the following estimates:*

$$||w_t||_{L^2(\Omega)} \leq C, \tag{4.11}$$

$$||w_t||_{H^1(K)} \leq C. \tag{4.12}$$

Here $C > 0$ depends on the a priori data only.

Remark. Obviously we have

$$||w_t||_{H^1(\Omega)} \leq C||\psi||_{H^{-1/2}(\partial\Omega)}.$$

However, the right-hand side cannot be bounded uniformly with respect to h, the size of the support of ψ.

Proof. We introduce the Robin function for problem (4.10). For every $y \in \Omega\backslash\sigma_0$, $R_t(x, y)$ is given as the distributional solution to

$$\text{div}(A_t \nabla R_t(., y)) = -\delta_y \quad \text{in } \Omega\backslash\sigma_0,$$

$$R_t(., y) = 0 \quad \text{on } \sigma_0,$$

$$\frac{\partial R_t}{\partial \nu}(., y) = 0 \quad \text{on } \partial\Omega.$$

Observe that for $x \in \Omega$, $y \to R_t(x,y)$ extends continuously to $\partial\Omega$. Moreover, the asymptotic behaviour near $\partial\Omega$ is the same as that of the Neumann function for the Laplacian in Ω, and therefore we have the bound

$$|R_t(x,y)| \leq C(1 + |\ln|x-y||) \quad \forall x \in \Omega,\ y \in \partial\Omega.$$

We have

$$w_t(x) = \int_{\partial\Omega} R_t(x,y)\psi(y)\, ds(y);$$

thus (4.11) follows easily by recalling that $||\psi||_{L^1(\partial\Omega)} \leq 2$. Finally, (4.12) follows from the Caccioppoli inequality. □

LEMMA 4.3. *Let $d(\sigma_0, \sigma_1) \leq \delta_0$. The mapping $[0,1] \ni t \to w_t \in H^1(\Omega)$ is differentiable, and its derivative w_t' is Lipschitz continuous with respect to t. Moreover, there exists $C > 0$, depending on the a priori data only, such that*

$$||w_t'||_{H^1(\Omega)} \leq Cd(\sigma_0,\sigma_1), \quad \forall t \in [0,1], \tag{4.13}$$

$$||w_t''||_{H^1(\Omega)} \leq Cd^2(\sigma_0,\sigma_1), \quad \textit{for almost every } t \in [0,1]. \tag{4.14}$$

Proof. From Lemma 4.1 we have that $t \to A_t$ is C^2. First, we show that $t \to w_t$ is Lipschitz continuous by taking finite differences of w_t. Using (4.10) and (4.12), we prove that such finite differences are uniformly bounded in $H^1(\Omega)$. Therefore, for almost every t, w_t' exists and satisfies

$$\begin{aligned} \operatorname{div}(A_t \nabla w_t' + A_t' \nabla w_t) &= 0 && \text{in } \Omega \backslash \sigma_0, \\ w_t' &= 0 && \text{on } \sigma_0, \\ \frac{\partial w_t'}{\partial \nu} &= 0 && \text{on } \partial\Omega. \end{aligned} \tag{4.15}$$

Notice that $A_t = I$ in $U \cup V$. Therefore $A_t' = 0$ there. Hence

$$||w_t'||_{H^1(\Omega)} \leq C||A_t' \nabla w_t||_{L^2(K)}$$

and therefore, by (4.5) and (4.12), (4.13) follows for almost every t. We can repeat the argument to prove that w_t' is Lipschitz continuous. Hence (4.13) holds for every t and, for almost every t, the second derivative satisfies

$$\begin{aligned} \operatorname{div}(A_t \nabla w_t'' + 2A_t' \nabla w_t' + A_t'' \nabla w_t) &= 0 && \text{in } \Omega \backslash \sigma_0, \\ w_t'' &= 0 && \text{on } \sigma_0, \\ \frac{\partial w_t''}{\partial \nu} &= 0 && \text{on } \partial\Omega. \end{aligned} \tag{4.16}$$

By arguments similar to those above and recalling (4.6), (4.12), and (4.13), we obtain (4.14). □

Now we can use the chain rule to differentiate $v_t = w_t \circ \zeta_t$. We obtain

$$v_t' = w_t'(\zeta_t) + (\nabla w_t)(\zeta_t) \cdot \begin{pmatrix} \xi_t' \\ \eta_t' \end{pmatrix}, \tag{4.17}$$

and also

$$v_t'' = w_t''(\zeta_t) + 2(\nabla w_t')(\zeta_t) \cdot \begin{pmatrix} \xi_t' \\ \eta_t' \end{pmatrix} + (\nabla w_t)(\zeta_t) \cdot \begin{pmatrix} \xi_t'' \\ \eta_t'' \end{pmatrix} + (D^2 w_t)(\zeta_t)\begin{pmatrix} \xi_t' \\ \eta_t' \end{pmatrix} \cdot \begin{pmatrix} \xi_t' \\ \eta_t' \end{pmatrix}. \tag{4.18}$$

Remark. We easily see that

$$\begin{aligned} &\Delta v_t' = \Delta v_t'' = 0 && \text{in } \Omega\backslash\sigma_t, \\ &\frac{\partial v_t'}{\partial\nu} = \frac{\partial v_t''}{\partial\nu} = 0 && \text{on } \partial\Omega, \end{aligned} \tag{4.19}$$

but on the other hand, v_t' and v_t'' verify somewhat complicated boundary conditions on σ_t. For instance, v_t' satisfies

$$(v_t')^{+}_{-} = \left((\nabla w_t)(\zeta_t) \cdot \begin{pmatrix} \xi_t' \\ \eta_t' \end{pmatrix}\right)^{+}_{-} \quad \text{on } \sigma^{+}_{-}.$$

Here the $^{+}_{-}$ sign distinguishes the one-sided traces on σ_t; moreover, the right-hand side must be interpreted in a delicate distributional sense. We avoid this approach by deriving estimates on v_t' and v_t'' directly from those on w_t' and w_t''.

LEMMA 4.4. *Assume that $d(\sigma_0, \sigma_1) \leq \delta_0$. There exists $C > 0$, depending on the a priori data only, such that*

$$||u_t'||_{L^2(\partial\Omega)} \leq Cd(\sigma_0, \sigma_1), \tag{4.20}$$

$$||u_t''||_{L^2(\partial\Omega)} \leq Cd^2(\sigma_0, \sigma_1). \tag{4.21}$$

Proof. We start by proving the following estimates on v_t' and v_t'':

$$||v_t'||_{L^2(\Omega)} \leq Cd(\sigma_0, \sigma_1), \tag{4.22}$$

$$||v_t''||_{H^{-1}(\Omega)} \leq Cd^2(\sigma_0, \sigma_1). \tag{4.23}$$

(4.22) follows from (4.17) by using (4.12) and (4.13) and by noticing that ζ_t' has support in K. (4.23) is a consequence of (4.18) via the estimates (4.13) and (4.14). Observe that the first three summands on the right-hand side of (4.18) can be bounded in $L^2(\Omega)$, whereas the last one can be estimated in $H^{-1}(\Omega)$ by an integration by parts.

Next, recalling the smoothness of $\partial\Omega$ and (4.19), we have a higher regularity near $\partial\Omega$, for instance

$$||v_t'||_{H^1(V)} \leq Cd(\sigma_0, \sigma_1),$$

$$||v_t''||_{H^1(V)} \leq Cd^2(\sigma_0, \sigma_1).$$

Consequently, by the trace theorem, we derive

$$||v_t'||_{L^2(\partial\Omega)} \leq Cd(\sigma_0, \sigma_1), \tag{4.24}$$

$$||v_t''||_{L^2(\partial\Omega)} \leq Cd^2(\sigma_0, \sigma_1). \tag{4.25}$$

Observe that (4.24) and (4.25) also imply that $c_t = -\frac{1}{|\partial\Omega|}\int_{\partial\Omega} v_t$ is twice differentiable:

$$\begin{aligned} c_t' &= -\frac{1}{|\partial\Omega|}\int_{\partial\Omega} v_t' \quad \forall t \in [0,1], \\ c_t'' &= -\frac{1}{|\partial\Omega|}\int_{\partial\Omega} v_t'' \quad \text{for almost every } t \in [0,1]. \end{aligned} \tag{4.26}$$

Therefore, we also obtain that $t \rightarrow u_t \in L^2(\partial\Omega)$ is twice differentiable, and we obtain (4.20) and (4.21). Finally, notice also that by (4.19) and the regularity estimates at the boundary, we have

$$\|u_t'\|_{H^2(\partial\Omega)} \le Cd(\sigma_0,\sigma_1). \qquad \square \tag{4.27}$$

The Taylor formula gives us

$$u_1 - u_0 = u_0' + R \quad \text{on } \Gamma, \tag{4.28}$$

where $R = \int_0^1 (t-1)u_t''\,dt$ satisfies

$$\|R\|_{L^2(\Gamma)} \le Cd^2(\sigma_0,\sigma_1). \tag{4.29}$$

The following statements, whose proof we defer until §5, allow us to conclude the proof of Theorem 2.1.

PROPOSITION 4.5. *Let $d(\sigma_0,\sigma_1) \le \delta_0$. There exists $c > 0$, depending on the a priori data only, such that*

$$\|u_0'\|_{L^2(\Gamma)} \ge cd(\sigma_0,\sigma_1). \tag{4.30}$$

LEMMA 4.6. *There exists a continuous, increasing function $\omega : [0,\infty) \to [0,\infty)$, with $\omega(0) = 0$, such that*

$$d(\sigma_0,\sigma_1) \le \omega\left(\sum_{j=1}^{2} \|u_1^j - u_0^j\|_{L^2(\Gamma)}\right). \tag{4.31}$$

Let us conclude now the proof of Theorem 2.1. In fact, if $d(\sigma_0,\sigma_1) \le \delta_0$, then from (4.28)–(4.30), we obtain

$$\|u_1 - u_0\|_{L^2(\Gamma)} \ge d(\sigma_0,\sigma_1)(c - Cd(\sigma_0,\sigma_1)).$$

Let $\delta_1 = \min\{\delta_0, \frac{c}{2C}\}$. If $d(\sigma_0,\sigma_1) \le \delta_1$, then (2.9) follows (recall (2.8)). On the other hand, by (4.31) of Lemma 4.6, when $d(\sigma_0,\sigma_1) \ge \delta_1$, we have that

$$d(\sigma_0,\sigma_1) \le \text{diam}\ \Omega \le \frac{\text{diam}\Omega}{\omega^{-1}(\delta_1)} \sum_{j=1}^{2} \|u_1^j - u_0^j\|_{L^2(\Gamma)}$$

and (2.9) again follows. $\square$

5. Proofs of Lemma 4.6 and Proposition 4.5.

Proof of Lemma 4.6. The estimate (4.31) is a slight variation of the estimate in [A1, Thm. 1.4]. In [A1] only the Dirac-type boundary data (3.1) are considered, and the $L^2(\Gamma)$ norm is replaced by the $L^\infty(\Gamma)$ norm.

In view of the considerations made in §3, Theorem 1.4 in [A1] also applies when the boundary data are of the type in (2.5) when $h \le h_0$, the number appearing in Proposition 3.4.

The change of the norm on the right-hand side of (4.31) can also be easily adjusted. In fact, by (4.27) we have

$$\|u_1 - u_0\|_{H^2(\partial\Omega)} \le C,$$

and therefore we can use the interpolation inequality

$$\|u_1 - u_0\|_{L^\infty(\Gamma)} \le C(\|u_1 - u_0\|_{H^2(\Gamma)})^{1/2}(\|u_1 - u_0\|_{L^2(\Gamma)})^{1/2}. \qquad \square$$

Proof of Proposition 4.5. Our proof will be based on the following two steps:

(I) There exist $r_0, c > 0$, depending on the a priori data only, and a point z_0, such that $B_{r_0}(z_0) \subset \Omega \backslash \sigma_0$ and

$$|u_0'(z_0)| \geq c d(\sigma_0, \sigma_1). \tag{5.1}$$

(II) There exist C, δ, $C > 0$, $0 < \delta < 1$, depending on the a priori data only such that

$$|u_0'(z_0)| \leq C ||u_0'||^{1-\delta}_{L^2(\Omega)} ||u_0'||^{\delta}_{L^2(\Gamma)}. \tag{5.2}$$

In fact, by (4.22), (4.24), and (4.26), we deduce

$$||u_0'||_{L^2(\Omega)} \leq C d(\sigma_0, \sigma_1), \tag{5.3}$$

and therefore, by combining (5.1)–(5.3), we get (4.30).

Proof of step (I). We use formula (4.17) near V_0. We have

$$|u_0'(z_0)| \geq \left| \nabla w_0(\zeta_0(z) \cdot \begin{pmatrix} \xi_0' \\ \eta_0' \end{pmatrix} \right| - |w_0'(\zeta_0(z))| - |c_0'|.$$

Recalling (4.9) and (4.8b), we have that

$$\zeta_0'(z) = (I - T_1)z, \quad \forall z \in U.$$

Therefore,

$$|\zeta_0'(z) - \zeta_0'(V_0)| = \left| \frac{(z - V_0)(V_0 - V_1 + W_1 - W_0)}{V_0 - W_0} \right| \leq |z - V_0| C d(\sigma_0, \sigma_1).$$

Let $\varphi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ be such that $e^{i\varphi} = \pm \frac{V_1 - V_0}{|V_1 - V_0|}$. Now, recall that ζ_0 is the identity mapping and, therefore, $w_0 = v_0 = u_0 - c_0$. Hence

$$\begin{aligned} \left| \nabla w_0(\zeta_0) \cdot \begin{pmatrix} \xi_0' \\ \eta_0' \end{pmatrix} \right| &= 2 |Re(\partial_z u_0 \zeta_0')| \\ &\geq 2 d(\sigma_0, \sigma_1) |Re(\partial_z u_0 e^{i\varphi})| - |\partial_z u_0| |\zeta_0'(z) - \zeta_0'(V_0)| \\ &\geq d(\sigma_0, \sigma_1) \left(2 |Re(\partial_z u_0)| - C |\partial_z u_0| |z - V_0| \right). \end{aligned}$$

Let $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ to be chosen later on, and consider $z_r = V_0 + 2re^{i\theta}$, with $0 < r < r_1$, where r_1, depending on the a priori data only, is such that $B_{3r_1}(V_0) \subset U$ and $3r_1 < \rho$ (with ρ as in Proposition 3.4). By (3.8) and (3.9) on $z = z_r$, we have

$$\left| Re(\partial_z u_0 e^{i\varphi}) + \alpha (2r)^{-1/2} \sin \left(\varphi - \frac{\theta}{2} \right) \right| \leq B.$$

Notice that, as $|\theta| \leq \frac{\pi}{2}$, $\varphi - \frac{\theta}{2}$ spans an interval of length $\frac{\pi}{2}$; therefore, we shall fix θ in such a way that $|\sin(\varphi - \frac{\theta}{2})| \geq \frac{1}{\sqrt{2}}$. We obtain, on $z = z_r$,

$$\left| \nabla w_0(\zeta_0) \cdot \begin{pmatrix} \xi_0' \\ \eta_0' \end{pmatrix} \right| \geq d(\sigma_0, \sigma_1) \left(\frac{A}{2} r^{-1/2} - B - C r^{1/2} \right).$$

Hence

$$|u_0'(z_r)| \geq d(\sigma_0, \sigma_1) \left(\frac{A}{2} r^{-1/2} - B - C r^{1/2} \right) - |w_0'(z_r)| - |c_0'|.$$

Now, by recalling (4.24) and (4.26) and observing that by (4.13) we have

$$|w_0'(z_r)| \leq C ||w_0'||_{H^1(U)} \leq C d(\sigma_0, \sigma_1),$$

we get

$$|u_0'(z_r)| \geq d(\sigma_0, \sigma_1)(cr^{-1/2} - K).$$

Hence, fixing $r_0 = \min\{r_1, (\frac{c}{2K})^2\}$, we obtain (5.1), with $z_0 = z_{r_0}$.

Proof of step (II). (5.2) consists of a stability estimate for a Cauchy problem for the Laplace equation. From (4.19) we have that $u_0' = v_0' + c_0'$ is harmonic in $\Omega \backslash \sigma_0$ and satisfies a homogeneous Neumann condition on $\partial\Omega$. The estimate (5.2) can be derived from two well-known estimates for harmonic functions (see [P]),

$$\int_{B_r^+(0)} |u|^2 \leq C_1 \left(\int_{B_{2r}^+(0)} |u|^2\right)^{1-\eta} \left(\int_{S_{2r}(0)} (|u|^2 + |\nabla u|^2)\right)^{\eta}, \tag{5.4}$$

$$\int_{B_{2r}(0)} |u|^2 \leq C_2 \left(\int_{B_{4r}(0)} |u|^2\right)^{1-\theta} \left(\int_{B_r(0)} |u|^2\right)^{\theta}. \tag{5.5}$$

Here $B_r^+(0)$ is the half disk $\{z : |z| < r, Imz > 0\}$, $S_r(0) = \{z = x : |x| < r\}$, and $C_1, C_2 > 0$, $0 < \eta, \theta < 1$ are absolute constants. We may conformally map a neighborhood of Γ in Ω onto $B_2^+(0)$ in such a way that the C^2-norm of this map and its inverse are controlled by the a priori data and Γ is mapped onto $S_2(0)$. Now

$$\int_{S_2(0)} (|u_0'|^2 + |\nabla u_0'|^2) = \int_{S_2(0)} (|u_0'|^2 + |u_{0,x}'|^2)$$
$$\leq C\left(\int_{S_2(0)} |u_0'|^2 + \left(\int_{S_2(0)} |u_{0,xx}'|^2\right)^{1/2} \left(\int_{S_2(0)} |u_0'|^2\right)^{1/2}\right).$$

By regularity estimates at the boundary, we see that

$$\int_{S_2(0)} |u_{0,xx}'|^2 \leq C \int_\Omega |u_0'|^2;$$

recall that here we are assuming $u_{0,y}' = 0$ on $S_2(0)$. Therefore, by (5.4) and returning to the original coordinates, we find a disk $B_\rho(w) \subset \Omega \backslash \sigma_0$ near Γ such that

$$\int_{B_\rho(w)} |u_0'|^2 \leq C\left(\int_\Omega |u_0'|^2\right)^{1-\eta/2} \left(\int_\Gamma |u_0'|^2\right)^{\eta/2};$$

here ρ and C depend on the a priori data only. Now we can form a chain of disks $\{B_r(w_j)\}_{j=1}^M$ such that $w_{j+1} \subset B_r(w_j)$ and $B_{4r}(w_j) \subset \Omega \backslash \sigma_0$ for all j, $B_r(w_0) \subset B_\rho(w)$, and $w_M = z_0$, and in such a way that the numbers r and M depend only on the a priori data. By a repeated use of (5.5), we arrive at

$$\int_{B_\rho(z_0)} |u_0'|^2 \leq C\left(\int_\Omega |u_0'|^2\right)^{1-\delta} \left(\int_\Gamma |u_0'|^2\right)^{\delta},$$

where C and δ depend only on η, θ, and the a priori data. The local boundedness estimate

$$|u_0'(z_0)|^2 \leq \frac{C}{r^2} \int_{B_r(z_0)} |u_0'|^2$$

yields (5.2). □

REFERENCES

[A1] G. Alessandrini, *Stable determination of a crack from boundary measurement*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 497–516.

[A2] ———, *Stability for the crack determination problem*, in Inverse Problems in Mathematical Physics, L. Päivärinta and E. Somersalo, eds., Springer-Verlag, Berlin, 1993, pp. 1–8.

[B-V] K. Bryan and M. Vogelius, *A computational algorithm to determine crack locations from electrostatic boundary measurements: The case of multiple cracks*, Internat. J. Engrg. Sci., to appear.

[DV] A. Diaz Valenzuela, *Unicitá e stabilitá per il problema inverso del crack perfettamente isolante, thesis*, Universitá di Trieste, Trieste, Italy, 1993.

[F-V] A. Friedman and M. Vogelius, *Determining cracks by boundary measurements*, Indiana Math. J., 38 (1989), pp. 527–556.

[L-S-V] V. Liepa, F. Santosa, and M. Vogelius, *Crack determination from boundary measurements: Computational reconstruction from laboratory measurements*, J. Nondestructive Evaluation, 53 (1993), pp. 163–176.

[P] L. E. Payne, *Improperly Posed Problems in Partial Differential Equations*, Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[S-V] F. Santosa and M. Vogelius, *A computational algorithm to determine cracks from electrostatic boundary measurements*, Internat. J. Engrg. Sci., 29 (1991), pp. 917–937.

SIAM J. MATH. ANAL.
Vol. 27, No. 2, pp. 376–403, March 1996

004

YOUNG MEASURE SOLUTIONS FOR A NONLINEAR PARABOLIC EQUATION OF FORWARD-BACKWARD TYPE*

SOPHIA DEMOULINI†

Abstract. The scope is to study the nonlinear parabolic evolution of forward-backward type

$$u_t \;=\; \nabla \cdot q(\nabla u) \qquad \text{on } Q_\infty \equiv \Omega \times \mathbb{R}^+$$

with initial data u_0 given in $H_0^1(\Omega)$, where $\Omega \subset \mathbb{R}^N$ is open, bounded, and $q \in C(\mathbb{R}^N; \mathbb{R}^N)$, an analogue to heat flux, satisfies $q = \nabla\phi$ with $\phi \in C^1(\mathbb{R}^N)$ of suitable growth. When ϕ is not convex classical solutions do not exist in general; the problem admits *Young measure* solutions. By that is meant a function u in a suitable Sobolev space and a gradient-generated family of probability measures $\nu = (\nu_{x,t})_{(x,t)\in Q_\infty}$ related by $\nabla u = \langle \nu, id\rangle$ almost everywhere (a.e.) (the identity integrated against ν) and such that the equation can be interpreted distributionally in H^{-1} : $\int_0^{+\infty}\int_\Omega \langle \nu, q\rangle \cdot \nabla\zeta + u_t\zeta \, dxdt$ for all $\zeta \in H_0^1(Q_\infty)$. The family ν is not unique, but through its first moment some of the classical properties are preserved: uniqueness of the function u is true; stability is reflected in a maximum principle and a comparison result. The asymptotic analysis yields, as time tends to infinity, a unique limit z and an associated Young measure ν^∞ such that the pair (z, ν^∞) is a Young measure solution of the steady-state problem $\nabla \cdot q(\nabla z) = 0$. The relevant energy function is shown to be monotone decreasing and asymptotically tending to its minimum, globally and locally in space.

Key words. Young measures, forward-backward heat equation, weak convergence, calculus of variations

AMS subject classification. 35K15

1. Introduction.

We study the nonlinear evolution problem

$$u_t = \nabla \cdot q(\nabla u) \qquad \text{on } Q_\infty \equiv \Omega \times \mathbb{R}^+, \tag{1}$$

$$u(\cdot, 0) = u_0 \qquad \text{on } \Omega, \tag{2}$$

$$u = 0 \qquad \text{on } \partial\Omega \times \mathbb{R}^+, \tag{3}$$

which will be denoted by $\mathcal{P}$. Here Ω is an open, bounded subset of $\mathbb{R}^N$ such that the cone or the segment property is satisfied on the boundary (as for example in the case of a Lipschitz boundary), and $q : \mathbb{R}^N \to \mathbb{R}^N$, a nonlinear, continuous, potential gradient function, an analogue to heat flux, satisfying $q = \nabla\phi$, where $\phi \in C^1(\mathbb{R}^N)$ (the space of continuously differentiable functions on $\mathbb{R}^N$) is of suitable growth. The initial data function u_0 is given in $H_0^1(\Omega)$ and the zero boundary data (3) can be taken to be a general time-homogeneous function $g \in H^1(\Omega)$. (Here $H^1(\Omega)$ is the Sobolev space of functions on Ω which together with their first-order weak derivatives are in $L^2(\Omega)$, and $H_0^1(\Omega)$ is its subset consisting of the functions with zero trace on the boundary of Ω.)

When ϕ is not convex, in which case the monotonicity condition $(q(x) - q(y)) \cdot (x - y) \geq 0$ for $x, y \in \mathbb{R}^N$ is violated on subsets of $\mathbb{R}^N$, equation (1) then constitutes a *forward-backward* parabolic equation which generally admits no classical strong or distributional solutions. The nonconvexity of the potential is compatible with the usual requirement that $q(\lambda) \cdot \lambda \geq 0$ be imposed on a theory of thermal conductors by the Clausius–Duhem inequality.

* Received by the editors January 18, 1994; accepted for publication (in revised form) June 17, 1994. This research was supported by the National Science Foundation and the Army Research Office.

† Department of Mathematics, University of California, Davis, CA 95616.

A notion of a solution appropriate for the study of $\mathcal{P}$ is that of a *measure-valued* or *Young measure* solution. By that is meant a function u in a Sobolev space and a parametrized family of probability measures $\boldsymbol{\nu} = (\nu_{x,t})_{(x,t)\in Q_\infty}$ generated by the spatial gradients of a sequence in the same space and satisfying

$$\int_0^{+\infty}\int_\Omega \left(\langle \boldsymbol{\nu}, q\rangle \nabla\zeta + \frac{\partial u}{\partial t}\zeta\right) dx\, dt = 0,$$

where

$$\langle \boldsymbol{\nu}, q\rangle \;=\; \int_{\mathbb{R}^N} q(\lambda)\, \boldsymbol{\nu}(d\lambda) \qquad \text{a.e. in } Q_\infty$$

for all ζ in an appropriate subspace of $H^1(Q_\infty)$. In addition, u and $\boldsymbol{\nu}$ satisfy

$$\nabla u \;=\; \langle \boldsymbol{\nu}, \mathrm{id}\rangle \;=\; \int_{\mathbb{R}^N} \lambda\, \boldsymbol{\nu}(d\lambda) \qquad \text{a.e. in } Q_\infty,$$

where $id(\lambda) = \lambda$. So to each point (x,t) in the domain Q_∞ is associated a probability measure $\nu_{x,t}$ on $\mathbb{R}^N$; via this parametrized measure the nonlinearity of $q(\nabla u)$ is replaced by the expected value of q, while the first moment of the measure is the gradient of the solution. To date the term "Young measure solution," although strictly derived from the fundamental theorem of Young measures described in §2, admits slightly different definitions by different authors; the decision of what is a Young measure solution of a problem must necessarily accommodate the way the generating sequence is chosen. In this case, the definition of Young measure solutions for $\mathcal{P}$ appears in §3.

The approach we shall assume in this paper for the study of (1)–(3) is the one employed by Kinderlehrer and Pedregal in [KP1] to establish existence. The method incorporates the explicit methods for solutions of evolution equations (cf. [BC], [HK]) with variational methods used to accommodate and describe the oscillatory behavior (cf. [Ba], [E], [KP1], etc). The combination of the two methods leads to the existence of Young measure solutions of evolution problems that may be of forward-backward type.

The analytical context of our approach to obtain existence is to approximate the dynamics of equation (1) with a sequence of stationary problems the solutions of which are in turn interpreted as minimizers of variational principles. More precisely, the time-discretized version of (1) is the Euler equation of a nonconvex variational principle which at each time step (of size h) is minimized. The minimizer solves the stationary problem and approximates the solution of $\mathcal{P}$ within time h. By taking arbitrarily small time steps we pass from the stationary to the evolution problem. The method is well known in the study of semigroups. It has been implemented by Horihata and Kikuchi in [HK] to construct weak solutions to a quasilinear parabolic problem associated with a convex variational principle. Further, this method has also been employed by Bethuel, Coron, Ghidaglia, and Soyeur in [BC] to establish existence of weak solutions for a nonlinear heat equation associated with weakly harmonic maps in a Sobolev-type space of functions of the unit ball into the sphere in $\mathbb{R}^3$.

In applying this method to treat $\mathcal{P}$ the difficulty that arises is twofold: first, the nonconvexity of the potential ϕ implicating the minimization of a nonconvex variational principle; second, the unwieldiness of the nonlinear dependence of the heat flow q on the gradient of the solution. Both situations call for sensible generalizations, in the former case that of a "minimizer" of a variational principle and in the latter, that of a "weak solution" of a differential equation. Respectively, the pertinent themes implemented in [KP1] to overcome these impediments are first, the *relaxation* of a nonconvex functional and second, replacing the nonlinearity of $q(\nabla u)$ with the expected value of q against a *Young measure.* The sense in which $\mathcal{P}$ has a solution is then that of a Young measure solution.

We review the variational method of Kinderlehrer and Pedregal in [KP1] to obtain existence and following that we investigate deeper the properties of the Young measure solution and establish a *uniqueness* result. It should be remarked that as a rule, nonuniqueness results appear in the literature (for example in [BC] or [HN]) regarding nonlinear parabolic problems, especially of a forward-backward nature. In our case, the uniqueness of the Young measure solution, although not directly dependent on the particular construction scheme of the solution, is contingent upon an *independence* property, namely that the heat flux q and the solution u be independent with respect to the Young measure $\boldsymbol{\nu}$, and furthermore, on a condition regarding the support of the Young measure $\boldsymbol{\nu}$, namely $\langle \boldsymbol{\nu}, q\rangle = \langle \boldsymbol{\nu}, p\rangle$, where $p = \nabla\phi^{**}$ and ϕ^{**} is the convexification of the potential ϕ. The function u is unique but the Young measure is not.

The solution u is also *continuously dependent on the initial data* in the L^2 norm. In addition, $u(\cdot,t)$ satisfies continuity properties in the L^2 norm, both as $t \to 0^+$ (monotone decreasing) and as $t \to +\infty$. Stability of the solution is reflected in the fact that it satisfies a *maximum principle* and a *comparison lemma.*

In §4 we investigate the time-asymptotics of $\mathcal{P}$. As time tends to infinity, the solution $u(\cdot,t)$ converges to z strongly in L^2 (monotonically decreasing) and weakly in H^1, and the measure $\boldsymbol{\nu}$ has a (weak) asymptotic limit $\boldsymbol{\nu}^\infty$ such that the pair $(z, \boldsymbol{\nu}^\infty)$ constitutes a Young measure solution to the steady-state version of $\mathcal{P}$, $\nabla\cdot q(\nabla z) = 0$. This is achieved by showing that the set of all weak limit points of $(u(\cdot,t))_{t\geq 0}$ in H^1 is invariant under the operator $\mathcal{P}$ and further, that there exists exactly one such weak limit point z. The asymptotic Young measure $\boldsymbol{\nu}^\infty$ has restricted support satisfying $supp\, \boldsymbol{\nu}^\infty \subseteq \{q(\lambda)\cdot\lambda = 0\} \cap \{\phi^{**} = \phi\}$.

In §5 we introduce the relevant *energy* function $E(t) = \int_0^\infty \phi^{**}(\nabla u)(x,t)\,dx$. As time tends to infinity, the energy converges to zero, monotonically decreasing globally in space. We show that it also vanishes locally in space, that is, on any subdomain $\omega \subseteq \Omega$ (although not monotonically on ω).

The forward-backward heat equation has also been studied by Höllig [H], Höllig and Nohel [HN], and Slemrod [Sl]. The treatment in [H] and [HN] concerns the Neumann initial value problem in the case of one spatial dimension ($\Omega = [0,1]$). It establishes in the model case of a piecewise linear heat flux q, decreasing on an interval $[a,b] \subset \Omega$, that a continuum of solutions exists for finite time satisfying (1) weakly in the sense of L^2. Each such solution is obtained as the sum of an explicitly constructed oscillating function and a smooth function that solves (weakly in L^2) an inhomogeneous heat equation.

The treatment in [Sl] involves Young measures but the spirit is different from that assumed here. $\mathcal{P}$ with Dirichlet or Neumann boundary conditions is approximated by a sequence of regular, singularly perturbed problems whose solutions are used to extract the Young measure solution. The differences between the Young measure solutions obtained in [Sl] and [KP1] are subtle. In [Sl] the heat flux q and the initial data u_0 are required to be sufficiently smooth, q must have strictly subquadratic growth, and equation (1) is satisfied in the sense of distributions. In [KP1] q is continuous, of linear growth, $u_0 \in H_0^1$, and equation (1) is satisfied in H^{-1}.

A one-dimensional convex analogue to $\mathcal{P}$ associated with a potential of linear growth has been studied by Zhou [Z]. The approach in [Z] differs from a Young measure viewpoint but here also a variational technique is developed to solve the stationary and evolution problems.

2. A result on sequences of gradient-generated Young measures. Lemma 2.4 describes limit points of sequences of parametrized measures. The question asked is whether the limit of a sequence of $W^{1,p}$ gradient Young measures is itself a gradient Young measure. The result applied in the next section shows that the parameterized measure solution derived in [KP1] can further be chosen to have the properties of a gradient Young measure. These properties will be particularly useful in the asymptotic analysis in §4.

We start with the notion of a $W^{1,p}$-gradient Young measure introduced and fully analyzed in [KP2] and state a characterization of such measures in Theorem 2.1. Most statements appear in vectorial formulation although it is their scalar version which we shall make use of subsequently in this paper.

DEFINITION. *A family of probability measures $\boldsymbol{\nu} = (\nu_x)_{x\in\Omega}$ on $\mathbf{M}$, where Ω is an open set in $\mathbb{R}^N$, and is a* $W^{1,p}$-gradient Young measure *for some $p \in [1,\infty]$ if*

(i) *$x \in \Omega \mapsto \int_{\mathbf{M}} f(A)\,\nu_x(dA) \in \mathbb{R}$ is a Lebesgue measurable function for all f bounded continuous on $\mathbf{M}$, the vector space $\mathbb{R}^{M\times N}$ of $M \times N$ matrices over the reals*;

(ii) *there is a sequence of functions $(u^k)_{k>0} \subset W^{1,p}(\Omega;\mathbb{R}^M)$ for which the representation formula*

$$\lim_{k\to+\infty}\int_E \phi(\nabla u^k)(x)\,dx = \int_E\int_{\mathbf{M}} \phi(A)\,\nu_x(dA)\,dx \tag{4}$$

holds for all measurable $E \subseteq \Omega$ and all ϕ in the space

$$\mathcal{E}_0^p(\mathbf{M}) := \left\{\phi \in C(\mathbf{M}) : \lim_{|A|\to+\infty}\frac{\phi(A)}{1+|A|^p} \text{ exists}\right\}$$

for $p < +\infty$, and for all functions ϕ continuous on $\mathbf{M}$ when $p = +\infty$. In the above, $C(\mathbf{M})$ denotes the space of continuous real-valued functions on $\mathbf{M}$. We shall use the notation

$$\overline{\phi} = \langle\boldsymbol{\nu},\phi\rangle : = \int_{\mathbf{M}} \phi(A)\,\boldsymbol{\nu}(dA).$$

Property (i) above is equivalent to weak$*$ measurability of $x \in \Omega \mapsto \nu_x \in$ *Prob*$(\mathbb{R}^N)$ (the set of probability measures on $\mathbf{M}$), that is, measurability with respect to the weak$*$ topology on *Prob*$(\mathbb{R}^N)$. Strong measurability usually will not be true. Property (ii) implies that there exists a sequence of functions $(u^k)_{k>0} \subset W^{1,p}(\Omega;\mathbb{R}^M)$ such that

$$\phi(\nabla u^k) \rightharpoonup \langle\boldsymbol{\nu},\phi\rangle \qquad \text{in } L^1(\Omega;\mathbb{R}^M) \text{ as } k \to +\infty$$

for all $\phi \in \mathcal{E}_0^p$, (where the notation $\rightharpoonup$ is used to denote weak convergence in the space indicated). In particular,

$$|\nabla u^k|^p \rightharpoonup \langle\boldsymbol{\nu},|A|^p\rangle \qquad \text{in } L^1(\Omega) \text{ as } k \to +\infty$$

(a condition not guaranteed for any subsequence by the uniform boundedness of the $(\|u^k\|_{W^{1,p}})_{k>0}$ alone).

As noted in [KP2] the space $\mathcal{E}_0^p(\mathbb{R}^N)$ is a separable Banach space in the norm

$$\|\phi\|_{\mathcal{E}^p} = \sup_{A\in\mathbf{M}} \frac{\phi(A)}{1+|A|^p} = \left\|\frac{\phi}{1+|\cdot|^p}\right\|_{L^\infty(\Omega)}.$$

Separability is desirable when the duals of the spaces such as $L^1(\Omega,\mathcal{E}_0^p(\mathbf{M}))$ are considered and the representation formula (4) remains valid if $\mathcal{E}_0^p(\mathbf{M})$ is replaced by the inseparable space

$$\mathcal{E}^p(\mathbf{M}) = \left\{\phi\in C(\mathbf{M}) : \sup_{A\in\mathbf{M}} \frac{\phi(A)}{1+|A|^p} < +\infty\right\}.$$

There is also the notion of *biting Young measure* defined in [KP2]. Recall that a bounded sequence $(z^k)_k \in L^1(\Omega)$ does not necessarily possess L^1 weak limit points. However, it admits *biting* limit points. That is, there is a decreasing sequence of subsets $E_{j+1} \subset E_j$ of Ω with $meas(E_j) \to 0$ and a subsequence of the $(z^k)_k$ that is convergent weakly in $L^1(\Omega\setminus E_j)$ to $z \in L^1(\Omega)$ for all j. This is Chacon's biting lemma. For details see for example [BM]. This motivates the following definition (cf., [BZ] and [KP2]).

DEFINITION. *A family of probability measures $\boldsymbol{\nu} = (\nu_x)_{x\in\Omega}$ is a* biting Young measure *provided there is a sequence $(z^k)_k \in L^p(\Omega)$ and $z\in L^1(\Omega)$ such that $|z^k|^p \rightharpoonup z$ and $\psi(z^k) \rightharpoonup \langle\nu_x,\psi\rangle$ in the biting sense as $k\to\infty$ for all $\psi\in\mathcal{E}_0^p$ (or $\mathcal{E}^p$).*

When the z^k are gradients of functions in $W^{1,p}$ it is straightforward to establish that a form of Jensen's inequality holds for biting Young measures, a property that characterizes $W^{1,p}$- gradient Young measures as described in the following theorem.

THEOREM 2.1. *Let $\boldsymbol{\nu} = (\nu_x)_{x\in\Omega}$ be a family of probability measures in $C(\mathbf{M})'$ (the dual space of the bounded continuous functions on $\mathbf{M}$). Then $\boldsymbol{\nu}$ is an $W^{1,p}$-gradient Young measure if and only if it is true that*

(i) *there exists $u\in W^{1,p}(\Omega;\mathbb{R}^M)$ such that*

$$\nabla u(x) = \int_{\mathbf{M}} A\,\nu_x(dA) \qquad x \text{ a.e. in } \Omega;$$

(ii) *Jensen's inequality,*

$$\phi(\nabla u(x)) \le \int_{\mathbf{M}} \phi(A)\,\nu_x(dA),$$

holds for all ϕ continuous in the case $p=+\infty$ and all $\phi\in\mathcal{E}^p(\mathbf{M})$ continuous, quasiconvex, and bounded below if $1\le p<+\infty$;

(iii) *the function*

$$x \mapsto \int_{\mathbf{M}} |A|^p\,\nu_x(dA)$$

is in $L^1(\Omega)$ if $1\le p<\infty$ and when $p=\infty$

$$\operatorname{supp}\nu_\alpha \subseteq K \qquad \text{for } \alpha \text{ a.e. in } \Omega,$$

where K is a compact independent of α.

For the proof see [KP2].

Remark. The authors also note that as a consequence *the $W^{1,p}$-biting Young measures are the same as the $W^{1,p}$-gradient Young measures* (but the sequences that will give rise to the measure as a gradient Young measure differ from the one that generates it as a biting Young measure).

The guarantee of the existence of Young measures draws upon a theorem originally proved by Tartar [T] and built on ideas developed by Young[Y]; a version appears in Ball [Ba] and an extension was proved by Schonbek [Sc]. The theorem describes weak limit points of sequences $(f(z^k))_k$, where f is continuous and $(z^k)_k$ is a sequence of measurable functions. Accordingly, the sequence $(z^k)_k$ defined on a $\mathcal{L}^N$-measurable subset $S \subseteq \mathbb{R}^N$ into $\mathbb{R}^M$ gives rise to a subsequence $(z^j)_{j\geq 1}$ and a parametrized family of measures $\boldsymbol{\nu} = (\nu_x)_{x\in S}$ on $\mathbb{R}^M$ such that $f(z^j)$ converge to $\langle \nu_x, f\rangle = \int_{\mathbb{R}^M} f(\lambda)\, \nu_x(d\lambda)$ weakly$*$ in $L^\infty(S)$ for all $f \in C_0(\mathbb{R}^M)$ (the set of continuous functions on $\mathbb{R}^M$ which vanish at infinity). With improved boundedness conditions on the generating sequence $(z^j)_{j\geq 1}$, the convergence of the $(f(z^j))_{j\geq 1}$ is obtained for a larger class of functions f and the measures $\boldsymbol{\nu}$ are probability measures on $\mathbb{R}^M$. For example, if $(z^j)_j$ is bounded in L^∞, then $f(z^j) \rightharpoonup \langle \boldsymbol{\nu}, f\rangle$ in weakly$*$ in L^∞ for any continuous f.

The representation is also valid if instead $(z^j)_j$ is bounded in L^p (actually a milder boundedness condition suffices); in this case, the measures are probability measures and for any measurable $E \subset S$ the sequence $\big(f(z^j)\big)_{j\geq 1}$ converges to $\langle \nu_x, f\rangle$ weakly in $L^1(E)$ for all f continuous such that $(f(z^j))_j$ is weakly sequentially precompact in $L^1(E)$. Hence it is important to establish criteria for the weak sequential precompactness of the $(f(z^j))_{j\geq 1}$ in L^1. When the domain is bounded, a general criterion is provided by de la Vallée Poussin: the $(f(z^j))_{j\geq 1}$ are weakly sequentially precompact in $L^1(E)$ for $E \subseteq \mathbb{R}^N$ bounded if and only if there exists $\psi : [0, +\infty) \mapsto \mathbb{R}$ with superlinear growth at infinity and such that

$$\sup_j \int_E \psi(f(z^j))\, dx \; < \; +\infty.$$

The following theorem of Ascerbi and Fusco in [AF] and Kinderlehrer and Pedregal in [KP1] also serves to characterize weak sequential precompactness in L^1 in a variational setting. It has an important application to minimization problems in variational calculus. The existence of (local) minimizers of a functional of the form

$$I(u) := \int_\Omega f(x, u, \nabla u)\, dx$$

over $W^{1,p}(\Omega; \mathbb{R}^M)$ is very closely related to the lower semicontinuity properties of I, which in turn are reflected in the quasiconvexity properties of f in the last argument. We recall the notion of quasiconvexity introduced by Morrey [Mo]: a Borel measurable locally integrable function $f : \mathbf{M} \longrightarrow \mathbb{R}$ is *quasiconvex* if for all $A \in \mathbf{M}$,

$$f(A) \; \leq \; \frac{1}{\mathcal{L}^N(D)} \int_D f(A + \nabla\zeta)\, dx$$

for all $\zeta \in W_0^{1,\infty}(D; \mathbb{R}^M)$ (in fact, it suffices to consider $\zeta \in C_0^\infty$) and for all D open bounded sets in $\mathbb{R}^N$ with $\mathcal{L}^N(\partial D) = 0$. In general, convexity is a stronger condition than quasiconvexity but in the scalar case, that is, when either $M = 1$ or $N = 1$, the two conditions are equivalent (cf. [D]).

THEOREM 2.2. *Suppose $f \in \mathcal{E}^p(\mathbf{M})$, for some $1 \leq p \leq +\infty$, is quasiconvex and bounded below and let $u^k \rightharpoonup u$ in $W^{1,p}(\Omega; \mathbb{R}^M)$. Then*

(i) *For all measurable $E \subseteq \Omega$,*

$$\int_E f(\nabla u)\, dx \; \leq \; \liminf_{k\to+\infty} \int_E f(\nabla u^k).$$

(ii) *If, in addition,*

$$\int_\Omega f(\nabla u^k)\,dx \overset{k\to+\infty}{\longrightarrow} \int_\Omega f(\nabla u)\,dx,$$

then the $(f(\nabla u^k))_{j>0}$ are weakly sequentially precompact in $L^1(\Omega)$ and the sequence converges (weakly) to $f(\nabla u)$.

The proof can be found in [KP1]. Part (ii) is a consequence of (i) and it implies that if

$$f(\nabla u^{k_j}) \overset{j\to+\infty}{\longrightarrow} f(\nabla u) \qquad \text{in } L^1(\Omega),$$

then the $W^{1,p}$-gradient Young measure $\boldsymbol{\nu} = (\nu_x)_{x\in\Omega}$ generated by $(\nabla u^{k_j})_{j>0}$ satisfies

$$\langle \boldsymbol{\nu}, f\rangle = f(\nabla u) \qquad x \text{ a.e. in } \Omega.$$

The consequence of Theorem 2.2, which we will have occasion to use directly in this paper, occurs when a p-growth condition of the function f from below allows one to obtain information on the L^p norm of the gradients. This is described in the next result and provides a sufficient (but not necessary) condition for a sequence of functions in $W^{1,p}$ to generate a $W^{1,p}$-gradient Young measure.

THEOREM 2.3. *Let f and $(u^k)_{k\geq 1}$ be as in Theorem 2.2(ii) and assume in addition that*

$$(c|A|^p - 1)^+ \leq \phi(A) \leq C|A|^p + 1$$

for $0 < c \leq C$. Let $\boldsymbol{\nu} = (\nu_x)_{x\in\Omega}$ be generated by the gradients $(\nabla u^k)_{k\geq 1}$. Then $\boldsymbol{\nu}$ is a $W^{1,p}$- gradient Young measure.

The proof can be found in [KP1].

We now state and prove a result on the sequences of gradient-generated Young measures. Given such a bounded sequence we extract a (weakly) convergent subsequence using duality so that the representation formula holds for functions of sublinear growth; then we extend the representation to hold for functions of strictly subquadratic growth; finally we show that the limiting measure is itself a biting Young measure and so a gradient-generated Young measure.

We use the following notation in the remainder of this section:

$$\mathcal{E}_0 \equiv \mathcal{E}_0^2(\mathbb{R}^N), \qquad \mathcal{F}_0 \equiv \mathcal{E}_0^1(\mathbb{R}^N), \qquad \text{and } \mathcal{G}_p \equiv \mathcal{E}_0^p(\mathbb{R}^N) \text{ for } 1 \leq p < 2,$$

where functions on $\mathbb{R}^N$ can be replaced with vector-valued functions. We also let $Q_\infty = \Omega \times \mathbb{R}^+$ with Ω an open bounded set in $\mathbb{R}^N$.

LEMMA 2.4. *Suppose that $(\boldsymbol{\nu}^\alpha)_{\alpha\geq 0}$ with $\boldsymbol{\nu}^\alpha = (\nu^\alpha_{x,t})_{(x,t)\in Q_\infty}$ is a sequence of $H^1_{loc}(Q_\infty)$-gradient Young measures and each $\boldsymbol{\nu}^\alpha$ is generated by $(\nabla v^{\alpha,m})_{m\geq 0}$, where $(v^{\alpha,m})_{m\geq 0}$ is a sequence in $H^1_{loc}(Q_\infty)$ uniformly bounded in α and m. Then a subsequence (not relabeled) of the $(\boldsymbol{\nu}^\alpha)_{\alpha>0}$ and an $H^1_{loc}(Q_\infty)$-gradient Young measure $\boldsymbol{\nu} = (\nu_{x,t})_{(x,t)\in Q_\infty}$ exist such that*

$$\boldsymbol{\nu}^\alpha \overset{\alpha\to 0}{\longrightarrow} \boldsymbol{\nu} \tag{5}$$

weakly in $L^1(Q_T;\, \mathcal{G}_p')$, weakly in $L^2(Q_T;\, \mathcal{F}_0')$, and weakly in $L^\infty(Q_T;\, M(\mathbb{R}^N))$ for each $T \geq 0$. That is,*

$$\langle \boldsymbol{\nu}^\alpha, \psi\rangle \overset{\alpha\to 0}{\longrightarrow} \langle \boldsymbol{\nu}, \psi\rangle$$

weakly in $L^1(Q_T)$ for $\psi \in \mathcal{G}_p$, weakly in $L^2(Q_T)$ for $\psi \in \mathcal{F}_0$, and weakly∗ in $L^\infty(Q_T)$ for $\psi \in C_0(\mathbb{R}^N)$. In addition, this convergence also holds in the biting sense for $\psi \in \mathcal{E}_0$.

Remark 1. Recall that the assumption on the $(\boldsymbol{\nu}^\alpha)_{\alpha \geq 0}$ implies that the representation formula

$$\int_0^T \int_\Omega \psi(\nabla v^{\alpha,m})(x,t)\theta(x,t)\,dx\,dt \overset{m\to\infty}{\longrightarrow} \int_0^T \int_\Omega \int_{\mathbf{M}} \psi(A)\,d\nu^\alpha_{x,t}(A)\,\theta(x,t)\,dx\,dt$$

holds for all $\psi \in \mathcal{E}_0$ and $\mathcal{E}^2$, for $\theta \in \mathrm{L}^1(Q_T)$, and for each $\alpha \geq 0$ (not necessarily uniformly in α). This in turn implies that the representation formula also holds for $\psi \in \mathcal{F}$ or $C_0(\mathbb{R}^N)$ weakly in L^2 or weakly∗ in L^∞, respectively.

Remark 2. Assume that a sequence of Young measures is bounded in $L^1_{loc}(Q_\infty;\, \mathcal{E}_0^{'})$. Duality cannot be used here to ensure a limit point. However, we are able to reduce to the case of Lemma 2.4 as follows:

Suppose $(\boldsymbol{\nu}^\alpha)_{\alpha>0}$, with $\boldsymbol{\nu}^\alpha = (\nu^\alpha_{x,t})_{(x,t)\in Q_\infty}$, is a sequence of Young measures bounded in $L^1_{loc}(Q_\infty; \mathcal{E}_0^{'})$. For each α let $(\nabla v^{\alpha,k})_{k>0}$ be the generating gradients, where $v^{\alpha,k} \in L^2_{loc}(\mathbb{R}^+, H^1_0(\Omega))$. Then $(\boldsymbol{\nu}^\alpha)_{\alpha>0}$ is bounded in $L^2_{loc}(Q_\infty; \mathcal{F}_0^{'}) \cap L^\infty_{loc}(Q_\infty; M(\mathbb{R}^N))$ and a diagonal subsequence of the $(v^{\alpha,k})_{\alpha,k}$ is bounded in $L^2_{loc}(\mathbb{R}^+, H^1_0(\Omega))$ uniformly in α and k (and can be taken to be the new generating sequence). Thus Lemma 2.4 applies.

The proof of this remark is straightforward and is omitted.

Proof of Lemma 2.4.

Step 1. Here we extract the subsequence of the measures satisfying (5). Fix $T > 0$. It is straightforward to see that $(\boldsymbol{\nu}^\alpha)_{\alpha \geq 0}$ is bounded in the spaces $L^2(Q_T; \mathcal{F}_0^{'})$ and $L^\infty(Q_T; M(\mathbb{R}^N))$, which are isomorphic to the dual spaces of $L^2(Q_T;\, \mathcal{F}_0)$ and $L^1(Q_T;\, C_0(\mathbb{R}^N))$, respectively. Using this we can extract a subsequence (not relabeled) $(\boldsymbol{\nu}^\alpha)_{\alpha \geq q0}$ and a parametrized probability measure $\boldsymbol{\nu} = (\nu_{x,t})_{(x,t)\in Q_\infty}$ such that

$$\boldsymbol{\nu}^\alpha \overset{\alpha\to 0}{\longrightarrow} \boldsymbol{\nu} \qquad \text{weakly in } L^2(Q_T;\, \mathcal{F}_0^{'}) \text{ and weakly} * \text{ in } L^\infty(Q_T;\, M(\mathbb{R}^N)).$$

We now show that the convergence remains valid if we allow ψ to have higher growth, provided we compensate by higher integrability on the test functions.

Claim. $\langle \nu^\alpha, \psi\rangle \overset{\alpha\to 0}{\longrightarrow} \langle \nu, \psi\rangle$ *weakly in $L^1(Q_T)$ for all $\psi \in \mathcal{G}_p$, $1 \leq p < 2$.*

Proof of Claim. The key idea here is that the sequence $(|\nabla v^{\alpha,m}|^p)_{\alpha,m}$ is weakly precompact in L^1 *uniformly* in α,m for each $1 \leq p \leq 2$. (This is not true for $p = 2$; the sequences $(|\nabla v^{\alpha,m}|^2)_m$ are by assumption weakly precompact in L^1 but not uniformly in α.) We use the same cutoff functions used in Ball [Ba] and Slemrod [Sl]. Set

$$\eta^k(\lambda) := \begin{cases} 1 & \text{if } |\lambda| \leq k-1, \\ k - |\lambda| & \text{if } k-1 \leq |\lambda| \leq k, \\ 0 & \text{if } |\lambda| \geq k. \end{cases}$$

Let $\psi \in \mathcal{G}_p$ and let $\theta \in L^\infty(Q_T)$. Define

$$\psi^k(\lambda) := \psi(\lambda)\eta^k(\lambda), \qquad \psi^k \in C_0(\mathbb{R}^N).$$

$$\left| \int_0^T \int_\Omega \langle \nu^\alpha_{x,t}, \psi\rangle \theta(x,t)\,dx\,dt \;-\; \int_0^T \int_\Omega \langle \nu_{x,t}, \psi\rangle \theta(x,t)\,dx\,dt \right|$$

$$\begin{aligned}
\le\quad & \|\theta\|_{L^\infty(Q_T)} \int_0^T \int_\Omega \left|\langle \boldsymbol{\nu}^\alpha,\, \psi - \psi^k\rangle\right| \,dx\,dt \\
+\quad & \left| \int_0^T \int_\Omega \left(\langle \boldsymbol{\nu}^\alpha,\, \psi^k\rangle - \langle \boldsymbol{\nu},\, \psi^k\rangle\right) dx\,dt \right| \\
+\quad & \|\theta\|_{L^\infty(Q_T)} \int_0^T \int_\Omega \left|\langle \boldsymbol{\nu},\, \psi - \psi^k\rangle\right| \,dx \\
=\quad & I + II + III.
\end{aligned}$$

Fix $\epsilon > 0$. It is a consequence of the Dunford–Pettis theorem that

$$\begin{aligned}
I &= \lim_{m\to+\infty} \|\theta\|_{L^\infty(Q_T)} \int_0^T \int_\Omega |\psi - \psi^k|(\nabla v^{\alpha,m})\,dx\,dt \\
&\le c \lim_{m\to+\infty} \int_{\{(x,t):|\nabla v^{\alpha,m}|\ge k-1\}} |\psi|(\nabla v^{\alpha,m})\,dx\,dt \\
&\le c \sup_{\alpha,m} \int_{\{(x,t):|\nabla v^{\alpha,m}|\ge k-1\}} (1 + c'|\nabla v^{\alpha,m}|^p)\,dx\,dt \\
&\longrightarrow 0
\end{aligned}$$

as $k \to +\infty$ uniformly in α, by the L^1 precompactness mentioned above for $1 \le p < 2$, and using that

$$meas\{(x,t) : |\nabla v^{\alpha,m}| \ge k\} \overset{k\to+\infty}{\longrightarrow} 0,$$

since $\|\nabla v^{\alpha,m}\|_{L^2(Q_T)}$ is uniformly bounded in α, m. For each k, $\psi^k \in C_0(\mathbb{R}^N)$ and so $\exists \delta(k,\epsilon)$ such that $II < \epsilon$ $\forall |\alpha| < \delta(k,\epsilon)$ (but not necessarily uniformly in k).

For III, assume $\psi \ge 0$. Then $0 \le \psi^k \nearrow \psi$ pointwise and so $\langle \nu_{x,t},\, \psi - \psi^k\rangle \to 0$ as $k \to +\infty$. (For general ψ, write $\psi = \psi^+ - \psi^-$ and $(\psi^+)^k = \psi^+\eta^k$, $(\psi^-)^k = \psi^-\eta^k$, and use the monotone convergence of each term.) So $\exists K(\epsilon)$ such that $III < \epsilon$ $\forall k \ge K(\epsilon)$.

We choose k for I and III which is independent of α; using this k we then find $\delta(\epsilon, k)$ for II. This shows that the sequence of $\boldsymbol{\nu}^\alpha$ converges in $L^\infty(Q_T; M(\mathbb{R}^N))$ to $\boldsymbol{\nu}$ and proves the claim.

Step 2. We now show that for each $1 \le p < 2$ the limit point $\boldsymbol{\nu}$ is a gradient Young measure. For this we fix such a p and find a sequence of gradients for which the representation formula holds for all functions in a dense set of $\mathcal{G}_p$ and show that the same sequence works for all ψ in $\mathcal{G}_p$. (It is obvious that for this argument one must work with the separable space $\mathcal{G}_p$ rather than the inseparable space $\mathcal{E}^p$.)

Fix $T > 0$. Let $(\phi_n)_{n\ge1}$ be dense in $\mathcal{G}_p$. For each $n \ge 1$ we have, by Step 1,

$$\phi_n(\nabla v^{\alpha,m}) \quad \overset{m\to+\infty}{\rightharpoonup} \quad \langle \boldsymbol{\nu}^\alpha, \phi_n\rangle \quad \text{in } L^1(Q_T),$$

and also

$$\langle \boldsymbol{\nu}^\alpha, \phi_n\rangle \quad \overset{\alpha\to0}{\rightharpoonup} \quad \langle \boldsymbol{\nu}, \phi_n\rangle \quad \text{in } L^1(Q_T).$$

Therefore a diagonal subsequence indexed by $\mu(n)$ exists such that

$$\phi_n(\nabla v^{\mu(n)}) \quad \overset{\mu(n)\to+\infty}{\rightharpoonup} \quad \langle \boldsymbol{\nu}, \phi^n\rangle \quad \text{in } L^1(Q_T).$$

This way we obtain the sequences $(\nabla v^{\mu(n)})_{\mu(n)\geq 1}$ for each n which we Cantor-diagonalize to obtain a single sequence $(\nabla v^{\mu})_{\mu\geq 1}$ such that the representation formula holds for each ϕ_n, i.e.,

$$\phi_n(\nabla v^{\mu}) \quad \overset{\mu\to+\infty}{\rightharpoonup} \quad \langle \boldsymbol{\nu}, \phi^n\rangle \qquad \text{in } L^1(Q_T) \text{ for all } n.$$

Now using density we show that the sequence of gradients just obtained is a generating sequence for the parametrized measure $\boldsymbol{\nu}$ obtained in Step 1. For this, let $\phi \in \mathcal{G}_p$ and $\epsilon > 0$ be given. Find $N(\epsilon)$ such that $\|\phi - \phi_n\|_{\mathcal{G}_p} < \epsilon \ \forall n \geq N(\epsilon)$. Let $\theta \in \mathrm{L}^{\infty}(Q_T)$. We have

$$\begin{aligned}
&\left| \int_0^T \int_\Omega \theta(x,t)\phi(\nabla v^{\mu})(x,t)\,dx\,dt \; - \; \int_0^T \int_\Omega \theta(x,t)\langle \nu_{x,t}, \phi\rangle\,dx\,dt \right| \\
&\leq \quad \|\,\theta\,\|_{L^\infty(\Omega)} \int_0^T \int_\Omega |\phi(\nabla v^{\mu}) - \phi_n(\nabla v^{\mu})|\,dx\,dt \\
&+ \quad \|\,\theta\,\|_{L^\infty(\Omega)} \int_0^T \int_\Omega |\phi_n(\nabla v^{\mu}) - \langle \boldsymbol{\nu}, \phi_n\rangle|\,dx\,dt \\
&+ \quad \|\,\theta\,\|_{L^\infty(\Omega)} \int_0^T \int_\Omega |\langle \boldsymbol{\nu},\, \phi_n - \phi\rangle|\,dx\,dt \\
&= \quad I + II + III.
\end{aligned}$$

For each term we have

$$\begin{aligned}
&I \leq c\,\|\phi - \phi_n\|_{\mathcal{G}_p} \int_0^T \int_\Omega (1 + |\nabla v^{\mu}|^2)dx\,dt \;\; \leq \;\; c\,\epsilon\,\forall n \geq N(\epsilon), \text{ uniformly in } \mu, \\
&II \leq c\,\epsilon \qquad \forall \mu \geq M(\epsilon, n), \\
&III \leq c\,\|\phi - \phi_n\|_{\mathcal{G}_p} \int_0^T \int_\Omega \langle \boldsymbol{\nu}, 1 + |\mathrm{id}|^p\rangle\,dx\,dt \;\; \leq \;\; c\,\epsilon.
\end{aligned}$$

Thus we may choose n for I and III which is independent of μ and for this n we find M for II.

We conclude that

$$\phi(\nabla v^{\mu}) \quad \overset{\mu\to+\infty}{\rightharpoonup} \quad \langle \nu_{x,t}, \phi\rangle \qquad \text{in } L^1(Q_T) \qquad \forall \phi \in \mathcal{G}_p$$

and by Remark 1 to Lemma 2.4 this finishes the proof of (5).

Step 3. We now extend $\boldsymbol{\nu}$ to $\mathcal{E}_0'$ and show it is a biting gradient Young measure. Then by the Remark to Theorem 2.1 it is an $H^1_{loc}(Q_\infty)$-gradient Young measure. First note that by assumption

$$\sup_{\alpha} \|\langle \boldsymbol{\nu}^{\alpha}, \mathrm{id}^2\rangle\|_{L^1(Q_T)} \leq \infty,$$

so that there exists a subsequence converging in biting. We use its limit to extend $\boldsymbol{\nu}$; there is a decreasing sequence of subsets $E_{j+1} \subset E_j$ of Q_T with $meas(E_j) \to 0$ and a subsequence of the $\boldsymbol{\nu}^{\alpha}$, not relabeled, such that

$$\text{(6)} \qquad \langle \boldsymbol{\nu}^{\alpha}, \mathrm{id}^2\rangle \; \rightharpoonup \; \langle \boldsymbol{\nu}, \mathrm{id}^2\rangle \text{ in } L^1(Q_T \setminus E_j) \text{ for all } j.$$

Accordingly, we may assume henceforth that in Step 1 the subsequence extracted to satisfy (5) also satisfies the above biting convergence. Using the Dunford–Pettis theorem and the growth condition we extend $\boldsymbol{\nu}$ for $\psi \in \mathcal{E}_0$ c anonically; for the same sequence of biting sets E_j and the subsequence as in (6) we see that $(\langle \boldsymbol{\nu}^\alpha, \psi \rangle)_\alpha$ are $L^1(Q_T \setminus E_j)$ weakly convergent for each j, so we denote this biting limit by $\langle n, \psi \rangle$:

$$\langle \boldsymbol{\nu}^\alpha, \psi \rangle \rightharpoonup \langle \boldsymbol{\nu}, \psi \rangle \text{ in } L^1(Q_T \setminus E_j), \text{ for all } j.$$

To see this, fix j, choose $A \subset Q_T$, and assume $\psi \geq 0$, or otherwise consider ψ^+ and ψ^-. Then as $meas(A) \to 0$,

$$\int_A \langle \boldsymbol{\nu}^\alpha, \psi \rangle \, dx \, dt \leq \int_A c\langle \boldsymbol{\nu}^\alpha, \mathrm{id}^2 \rangle + c' \, dx \, dt \longrightarrow 0$$

uniformly in α. In addition, by the growth condition on ψ, the sequence $(\langle \boldsymbol{\nu}^\alpha, \psi \rangle)_\alpha$ is weakly Cauchy in $L^1(Q_T \setminus E_j)$.

We now produce a generating sequence for the measure $\boldsymbol{\nu}$. Recall that

$$|\nabla v^{\alpha,m}|^2 \overset{m \to \infty}{\rightharpoonup} \langle \boldsymbol{\nu}^\alpha, \mathrm{id}^2 \rangle \qquad \text{in } L^1(Q_T)$$

and also

$$\langle \boldsymbol{\nu}^\alpha, \mathrm{id}^2 \rangle \qquad \overset{\alpha \to 0}{\rightharpoonup} \qquad \langle \boldsymbol{\nu}, \mathrm{id}^2 \rangle \qquad \text{in biting.}$$

So for a diagonal subsequence we have

$$|\nabla v^\mu|^2 \qquad \overset{\mu \to \infty}{\rightharpoonup} \qquad \langle \boldsymbol{\nu}, \mathrm{id}^2 \rangle$$

so that $\psi(\nabla v^\mu) \rightharpoonup \langle \boldsymbol{\nu}, \psi \rangle$ in $L^1(Q_T)$ for all $\psi \in \mathcal{E}_0$. Hence $\boldsymbol{\nu}$ is a biting Young measure. Letting ∇v be the $L^2(Q_T)$ weak limit of $(\nabla v^\mu)_\mu$ and using Theorem 2.1, we conclude that $\boldsymbol{\nu}$ is a gradient Young measure. This completes the proof of the lemma. □

3. The variational treatment and the existence of a Young measure solution.

Assumptions. We define the two separable Banach spaces

$$\mathcal{E}_0 \equiv \mathcal{E}_0^2(\mathbb{R}^N) := \left\{ \psi \in C(\mathbb{R}^N) : \lim_{|A| \to +\infty} \frac{|\psi(A)|}{1 + |A|^2} \text{ exists} \right\}$$

and

$$\mathcal{F}_0 \equiv \mathcal{E}_0^1(\mathbb{R}^N; \mathbb{R}^N) := \left\{ \psi \in C(\mathbb{R}^N; \mathbb{R}^N) : \lim_{|A| \to +\infty} \frac{|\psi(A)|}{1 + |A|} \text{ exists} \right\}.$$

We assume the heat flux satisfies $q = \nabla \phi$ on $\mathbb{R}^N$ with $\phi \in C^1(\mathbb{R}^N)$. We impose the growth conditions $\phi \in \mathcal{E}_0$, $q \in \mathcal{F}_0$, and furthermore,

$$(c|a|^2 - 1)^+ \leq \phi(a) \leq C|a|^2 + 1 \qquad \forall a \in \mathbb{R}^N, \tag{7}$$

$$|q(a)| \leq C|a| \qquad \forall a \in \mathbb{R}^N. \tag{8}$$

We let ϕ^{**} denote the *convexification* of ϕ, that is,

$$\phi^{**} = \sup\{f(x) : f \leq \phi, \ f \text{ convex}\}.$$

Since ϕ is in $C^1(\mathbb{R}^N)$, so is ϕ^{**}, and we set

$$p := \nabla\phi^{**}.$$

We note that $q = p$ on the set $\{\phi = \phi^{**}\}$ and that ϕ^{**} and p satisfy the same growth conditions as ϕ and q, respectively. We assume

$$\phi^{**}(0) = 0$$

and hence $p(0) = 0$ follows and by the convexity of ϕ^{**} implies

$$p(\lambda)\cdot\lambda \geq 0 \qquad \forall\lambda \in \mathbb{R}^N.$$

Under these hypotheses we fix ideas by agreeing on the following definition.

DEFINITION. *A* measure solution *to $\mathcal{P}$,*

$$\begin{aligned} u_t &= \nabla\cdot q(\nabla u) \qquad \textit{in } \; Q_\infty := \Omega\times\mathbb{R}^+, \\ u(x,0) &= u_0(x) \qquad \textit{for } x \textit{ in } \; \Omega, \\ u &= 0 \qquad \textit{on } \partial\Omega\times\mathbb{R}^+, \end{aligned}$$

with $u_0 \in H_0^1(\Omega)$ a given function, is a pair $(u, \boldsymbol{\nu})$, where $u \in H^1_{loc}(Q_\infty)\cap L^\infty\left(\mathbb{R}^+;\, H_0^1(\Omega)\right)$ and $\boldsymbol{\nu} = (\nu_{x,t})_{(x,t)\in Q_\infty}$ is a parametrized family of probability measures on $\mathbb{R}^N$ such that the equation

$$\int_0^{+\infty}\int_\Omega \left(\langle\boldsymbol{\nu}, q\rangle\cdot\nabla\zeta + \frac{\partial u}{\partial t}\zeta\right) dx\,dt = 0 \qquad \forall\zeta \in H_0^1(Q_\infty) \tag{9}$$

(where $\langle\nu_{x,t}, q\rangle = \int_{\mathbb{R}^N} q(\lambda)\nu_{x,t}(d\lambda)$) and

$$\langle\nu_{x,t}, \mathrm{id}\rangle = \nabla u(x,t) \qquad (x,t) \textit{ a.e. in } Q_\infty \tag{10}$$

hold. Equivalently stated, equation (9) *is*

$$u_t \;=\; \nabla\cdot\langle\boldsymbol{\nu}, q\rangle \qquad \textit{in } H^{-1}(Q_\infty). \tag{11}$$

If in addition $\boldsymbol{\nu}$ is a Young measure generated by the gradients with respect to x of a sequence in $L^2_{loc}(\mathbb{R}^+; H_0^1(\Omega))$, then the pair $(u, \boldsymbol{\nu})$ is called a Young measure solution to $\mathcal{P}$. We will say that $(u, \boldsymbol{\nu})$ is a solution or solves to indicate that the pair is a Young measure solution *to $\mathcal{P}$.*

The function u is also called a solution and the use of the term is clear from the context. In the above, $\Omega \subset \mathbb{R}^N$ is an open, bounded set, $\partial\Omega\times\mathbb{R}^+$ is the lateral boundary of the parabolic cylinder Q_∞, and *id* stands for the identity function. The notation H^1 stands for the Sobolev space $W^{1,2}$, and $W_0^{1,2}$ for its subspace of functions with zero trace on the boundary. As usual $H^{-1} = W^{-1,2}$ denotes the dual of $W_0^{1,2}$. The space H^1_{loc} stands for the subspace of H^1 of functions which, together with their first weak derivatives, are locally square-integrable and $H^1_{0,loc}(Q_\infty)$ is the space of functions $\zeta \in H^1_{loc}(Q_\infty)$ with t-slices $\zeta(\cdot, t) \in H_0^1(\Omega)$ for $t \geq 0$ a.e.

Remark 1. By the existence proof below there is a Young measure solution to $\mathcal{P}$ such that $u_t \in L^2(Q_\infty)$ and (9) is satisfied also locally in time, that is, for test functions $\zeta \in H_0^1(Q_\infty)$. In particular, the solution u is an admissible test function.

Remark 2. A consequence of the above definition is that an equilibrium equation is also satisfied pointwise in time in $H^{-1}(\Omega)$; indeed, for t a.e. in $[0, T]$ and for all $\zeta \in H_0^1(\Omega)$ we have

$$\int_0^T\int_\Omega \langle\nu_{x,t}, q\rangle\cdot\nabla\zeta(x)\,dx\,dt \;=\; -\int_0^T\int_\Omega \frac{d}{dt}u(x,t)\zeta(x)\,dx\,dt;$$

differentiating in time we obtain

$$\int_\Omega \langle \nu_{x,t}, q\rangle \cdot \nabla\zeta(x)\, dx \;=\; -\int_\Omega u_t(x,t)\zeta(x)\, dx \qquad t \text{ a.e. in } \mathbb{R}^+ \ \forall\zeta \in H_0^1(\Omega).$$

Remark 3. A classical solution to $\mathcal{P}$ which is bounded in time is a measure solution with $\boldsymbol{\nu} = \delta_{\nabla u}$.

THEOREM 3.1 (Existence). *Under the assumptions stated above there exists a Young measure solution $(u, \boldsymbol{\nu})$ to $\mathcal{P}$. In addition, $u_t \in L^2(Q_\infty)$,*

$$\operatorname{supp} \nu_{x,t} \subset \{a \in \mathbb{R}^N : \phi(a) = \phi^{**}(a)\} \qquad (x,t) \text{ a.e. in } Q_\infty,$$

and $(u, \boldsymbol{\nu})$ is also a Young measure solution of the relaxed problem

$$u_t \;=\; \nabla \cdot p(\nabla u) \qquad \text{in } H^{-1}(Q_\infty)$$

with the same initial-boundary data.

Proof. The following existence proof is due to Kinderlehrer and Pedregal [KP1].

Step 1. Let $h > 0$ be fixed and for each $j \geq 0$ consider the functionals

$$\Phi_h(v; u^{h,j-1}) \;:=\; \int_\Omega \phi(\nabla v) + \frac{1}{2h}(v - u^{h,j-1})^2\, dx \qquad \text{for } v \in H_0^1(\Omega)$$

and

$$\Phi_h^{**}(v; u^{h,j-1}) \;:=\; \int_\Omega \phi^{**}(\nabla v) + \frac{1}{2h}(v - u^{h,j-1})^2\, dx \qquad \text{for } v \in H_0^1(\Omega).$$

We drop the explicit dependence on h. By relaxation,

$$I \;:=\; \inf\left\{\Phi(v; u^{h,j-1}) : v \in H_0^1(\Omega)\right\} \;=\; \inf\left\{\Phi^{**}(v; u^{h,j-1}) : v \in H_0^1(\Omega)\right\}.$$

Let $(u^{h,j,k})_{k\geq 1} \subset H_0^1(\Omega)$ be a minimizing sequence for Φ (and Φ^{**}). By the growth condition (7) and the Rellich theorem, together with the H^1-weak sequential lower semicontinuity of Φ^{**}, there exist $u^{h,j} \in H_0^1(\Omega)$ and a subsequence,[1] not relabeled, such that

$$u^{h,j,k} \overset{k\to\infty}{\longrightarrow} u^{h,j} \qquad \text{weakly in } H_0^1(\Omega) \text{ and strongly in } L^2(\Omega),$$

$$I \;=\; \Phi^{**}(u^{h,j}; u^{h,j-1}),$$

and therefore,

$$(12)\quad \int_\Omega \phi^{**}(\nabla u^{h,j})\, dx \;=\; \lim_{k\to+\infty} \int_\Omega \phi^{**}(\nabla u^{h,j,k})\, dx \;=\; \lim_{k\to+\infty} \int_\Omega \phi(\nabla u^{h,j,k})\, dx.$$

Then by Theorem 2.2

$$(13)\qquad \phi^{**}(\nabla u^{h,j,k}) \overset{k\to\infty}{\longrightarrow} \phi^{**}(\nabla u^{h,j}) \qquad \text{in } L^1(\Omega).$$

Let $\boldsymbol{\nu}^{h,j} = (\nu_x^{h,j})_{x\in\Omega}$ be the Young measure generated by the $(\nabla u^{h,j,k})_{k\geq 1}$. By Theorem 2.3 $\boldsymbol{\nu}^{h,j}$ is an $H^1(\Omega)$-gradient Young measure and the sequence $(\phi(\nabla u^{h,j,k}))_k$

[1] In fact, the whole sequence converges (weakly) as the minimizer $u^{h,j}$ is unique.

is also weakly convergent in L^1; the representation formula (4) describes the weak L^1 limits and by (12) and (13) we obtain

$$\int_\Omega \phi^{**}(\nabla u^{h,j})\,dx \;=\; \int_\Omega \langle \boldsymbol{\nu}^{h,j}, \phi^{**}\rangle\,dx \;=\; \int_\Omega \langle \boldsymbol{\nu}^{h,j}, \phi\rangle\,dx,$$

which together with $\phi^{**} \leq \phi$, implies

$$\operatorname{supp} \boldsymbol{\nu}^{h,j} \subseteq \{\phi = \phi^{**}\} \tag{14}$$

and therefore

$$\langle \boldsymbol{\nu}^{h,j}, \phi\rangle \;=\; \langle \boldsymbol{\nu}^{h,j}, \phi^{**}\rangle \;=\; \phi^{**}(\nabla u^{h,j}) \qquad x \text{ a.e. in } \Omega, \tag{15}$$

$$\nabla u^{h,j} \;=\; \langle \boldsymbol{\nu}^{h,j}, id\rangle \qquad x \text{ a.e. in } \Omega. \tag{16}$$

In addition,

$$\langle \boldsymbol{\nu}^{h,j}, q\rangle \;=\; \langle \boldsymbol{\nu}^{h,j}, p\rangle \qquad x \text{ a.e. in } \Omega, \tag{17}$$

which follows from (14). Setting the Gâteaux derivative of $\Phi^{**}(\cdot; u^{h,j-1})$ to zero at the minimizer $u^{h,j}$, we obtain the equilibrium equation

$$\int_\Omega p(\nabla u^{h,j})\cdot\nabla\zeta + \frac{1}{h}(u^{h,j} - u^{h,j-1})\zeta\,dx \;=\; 0 \qquad \forall \zeta \in H_0^1(\Omega),$$

or equivalently the Euler–Lagrange equation $\nabla \cdot p(\nabla u^{h,j}) = \frac{1}{h}(u^{h,j} - u^{h,j-1})$ in $H^{-1}(\Omega)$. By the stability of Young measure minimizers discussed in [KP1], the equilibrium equation holds with $\langle \boldsymbol{\nu}^{h,j}, q\rangle = \langle \boldsymbol{\nu}^{h,j}, p\rangle$ in place of $p(\nabla u^{h,j})$ and thus

$$\nabla \cdot \langle \boldsymbol{\nu}^{h,j}, q\rangle \;=\; \nabla \cdot \langle \boldsymbol{\nu}^{h,j}, p\rangle \;=\; \nabla \cdot p(\nabla u^{h,j}) \qquad \text{in } H^{-1}(\Omega) \tag{18}$$

hold.

Let $I^{h,j} = [hj, h(j+1))$ and $\chi^{h,j}$ be the indicator function of $I^{h,j}$ and for $t > 0$ set

$$\lambda^{h,j}(t) \;=\; \begin{cases} \frac{t}{h} - j & \text{if } hj \leq t < h(j+1), \\ 0 & \text{otherwise.} \end{cases}$$

Define for x a.e. in Ω each $t \in \mathbb{R}^+$,

$$u^h(x,t) :\,= \sum_j \chi^{h,j}(t)\left\{u^{h,j}(x) + \lambda^{h,j}(t)(u^{h,j+1}(x) - u^{h,j}(x))\right\} \tag{19}$$

so that $u^h \in L^\infty(\mathbb{R}^+; H_0^1(\Omega))$ and also

$$u^h = 0 \text{ on } \partial\Omega \times \mathbb{R}^+.$$

We let

$$w^h :\,= \sum_j \chi^{h,j} u^{h,j} \qquad \in L^\infty(\mathbb{R}^+; H_0^1(\Omega)), \tag{20}$$

$$\boldsymbol{\nu}^h = (\nu^h_{x,t})_{(x,t)\in Q_\infty} := \sum_j \chi^{h,j}\boldsymbol{\nu}^{h,j} \quad \in \mathrm{L}^1_{loc}(Q_\infty;\mathcal{E}_0') \tag{21}$$

be probability measures on $\mathbb{R}^N$ (where $\mathcal{E}_0'$ is the dual space of $\mathcal{E}_0$). By (16) we know

$$\nabla w^h = \langle \boldsymbol{\nu}^h, \mathrm{id}\rangle \qquad (x,t) \text{ a.e. in } Q_\infty. \tag{22}$$

We also let

$$\bar{q}^h := \langle \boldsymbol{\nu}^h, q\rangle = \sum_j \chi^{h,j}\langle \boldsymbol{\nu}^{h,j}, q\rangle \quad \in L^2(Q_\infty).$$

Differentiating (19) almost everywhere in time we have

$$\frac{\partial u^h}{\partial t}(x,t) = \sum_j \chi^{h,j}(t)\frac{1}{h}(u^{h,j+1} - u^{h,j})(x) \quad \in L^2(Q_\infty).$$

Then for almost every $t > 0$,

$$\frac{\partial u^h}{\partial t} = \nabla\cdot\langle \boldsymbol{\nu}^h, q\rangle \qquad \text{in } H^{-1}(\Omega)$$

or, equivalently, the equation

$$\int_\Omega \left(\bar{q}^h\cdot\nabla\zeta + \frac{\partial u^h}{\partial t}\zeta\right) dx = 0 \quad \forall\zeta\in H^1_0(\Omega) \tag{23}$$

holds. From (23) it is easy to deduce that

$$\int_0^\tau\int_\Omega \langle \boldsymbol{\nu}^h, q\rangle\cdot\nabla\zeta + \frac{\partial u^h}{\partial t}\zeta\,dx\,dt = 0 \quad \forall\zeta\in H^1_0(Q_\tau)\,\forall\tau\in[0,+\infty] \tag{24}$$

(also for $\zeta(\cdot,t)\in H^1_{0,loc}(Q_\infty)$), that is,

$$\frac{\partial u^h}{\partial t} = \nabla\cdot\langle \boldsymbol{\nu}^h, q\rangle \qquad \text{in } H^{-1}(Q_\tau)\,\forall\tau\in[0,+\infty].$$

By (17), (18), and (20),

$$\langle \boldsymbol{\nu}^h, q\rangle = \langle \boldsymbol{\nu}^h, p\rangle \qquad \text{for each } t>0 \text{ and } x \text{ a.e. in } \Omega, \tag{25}$$

$$\nabla\cdot\langle \boldsymbol{\nu}^h, q\rangle = \nabla\cdot\langle \boldsymbol{\nu}^h, p\rangle = \nabla\cdot p(\nabla w^h), \tag{26}$$

both in $H^{-1}(\Omega)$ for each $t\geq 0$ and in $H^{-1}(Q_\tau)\,\forall\tau\in[0,+\infty]$.

Step 2. Using the growth conditions (7) and (8) on ϕ and q we obtain uniform estimates in h for the $(u^h)_{h>0}$ and $(w^h)_{h>0}$ in $L^\infty\left(\mathbb{R}^+;\, H^1_0(\Omega)\right)$ and $(\bar{q}^h)_{h>0}$ in $L^2(Q_\infty)$. Further we obtain that $(\frac{\partial u^h}{\partial t})_{h>0}\in L^2(Q_\infty)$ is bounded in h and the Young measures $(\boldsymbol{\nu}^h)$ are bounded in $\mathrm{L}^\infty(\mathbb{R}^+;\mathcal{E}_0')$. Using weak compactness we may therefore extract weakly convergent subsequences indexed by $h'\to 0$ and a pair $(u,\boldsymbol{\nu})$ satisfying (9) and (10). By [KP1, Lem. 6.3] u is obtained as the common weak limit in $L^2_{loc}(Q_\infty)$ of $(u^{h'})_{h'>0}$ and $(w^{h'})_{h'>0}$ and it satisfies $\frac{\partial u}{\partial t}\in L^2(Q_\infty)$. Along h' (24) yields the equation

$$\int_0^\tau\int_\Omega \left(\langle \boldsymbol{\nu}, q\rangle\cdot\nabla\zeta + \frac{\partial u}{\partial t}\zeta\right) dx\,dt = 0 \quad \forall\zeta\in H^1_0(Q_\tau)\;\forall\tau\in[0,+\infty] \tag{27}$$

(and as in (24), all $\zeta \in H^1_{loc}(Q_\infty)$ with $\zeta(\cdot,t) \in H^1_0(\Omega)$ for almost all $t > 0$ are admissible). Further, by (25),

$$(28) \qquad \langle \boldsymbol{\nu}, q\rangle = \langle \boldsymbol{\nu}, p\rangle \qquad (x,t) \text{ a.e. in } Q_\infty.$$

In addition, by (14) we have

$$(29) \qquad \operatorname{supp} \boldsymbol{\nu} \subseteq \{\phi = \phi^{**}\}.$$

In [KP1] the measure $\boldsymbol{\nu}$ is extracted as the weak limit of the sequence of Young measures $(\boldsymbol{\nu}^h)_{h>0}$ in $L^\infty(\mathbb{R}^+;\mathcal{E}_0')$. We next establish, in accordance with the definition, that $\boldsymbol{\nu}$ is a Young measure generated by the spatial gradients of a sequence and related to u via

$$\langle \nu_{x,t}, \mathrm{id}\rangle = \nabla u(x,t) \qquad (x,t) \text{ a.e. in } Q_\infty.$$

This is a direct consequence of Lemma 2.4 of the previous section and is proved in the following corollary.

Corollary 3.2. *The measure $\boldsymbol{\nu}$ obtained in the proof of the existence theorem (Theorem 3.1) is a Young measure generated by the gradients with respect to x of a sequence in $L^2_{loc}(\mathbb{R}^+;H^1_0(\Omega))$ and so the pair $(u,\boldsymbol{\nu})$ is a Young measure solution of $\mathcal{P}$ (in the sense of the definition).*

Proof. In the notation of Theorem 3.1 recall that

$$\nu^h_{x,t} := \sum_{j\geq 0} \chi_{[hj,h(j+1))}(t)\nu^{h,j}_x \qquad \text{for } x \text{ a.e. in } \Omega \text{ and } \forall t \geq 0.$$

Then for each $h > 0$ the sequence $(\nabla u^{h,k})_{k\geq 0}$ defined by

$$u^{h,k} := \sum_{j\geq 0} \chi_{[hj,h(j+1))}(t) u^{h,j,k}(x) \qquad \in L^2_{loc}(\mathbb{R}^+;H^1_0(\Omega))$$

generates $\boldsymbol{\nu}^h$. We apply Remark 2 following Lemma 2.4 on $\boldsymbol{\nu}^h$ to extract a subsequence indexed by $h' \to 0$ along which the $\boldsymbol{\nu}^{h'}$ converge to a parametrized measure $\boldsymbol{\nu} = (\nu_{x,t})_{(x,t)\in Q_\infty}$, which is generated by a subsequence of $(\nabla u^{h,k})_{h,k}$. In particular,

$$\langle \boldsymbol{\nu}^{h'}, \mathrm{id}\rangle \rightharpoonup \langle \boldsymbol{\nu}, \mathrm{id}\rangle \qquad \text{in } L^2_{loc}(Q_\infty)$$

(because $\mathrm{id} \in \mathcal{F}_0$), and (22) then gives

$$\langle \nu_{x,t}, \mathrm{id}\rangle = \nabla u(x,t) \qquad (x,t) \text{ a.e. in } Q_\infty,$$

since $\nabla w^{h'} \rightharpoonup \nabla u$ in $L^2(\Omega)$. □

4. Uniqueness and properties of the Young measure solution.

4.1. Uniqueness. The following lemma describes a property of the solution upon which the uniqueness proof relies.

Lemma 4.1 (Independence). *For $(u,\boldsymbol{\nu})$ a solution of (1) and (2) the equality*

$$(30) \qquad \langle \nu_{x,t}, q\cdot \mathrm{id}\rangle = \langle \nu_{x,t}, q\rangle \cdot \langle \nu_{x,t}, \mathrm{id}\rangle \qquad (x,t) \text{ a.e. in } Q_\infty$$

holds, i.e., q and ∇u are independent with respect to the Young measure $\boldsymbol{\nu}$.

Proof.

Step 1 (the time-discretized case). Fix $h > 0$.

CLAIM. $\langle \nu_x^{h,j}, q \cdot \mathrm{id}\rangle = \langle \nu_x^{h,j}, q\rangle \cdot \langle \nu_x^{h,j}, \mathrm{id}\rangle$ *x a.e. in Ω for $j = 0, 1, \ldots$.*

Proof of Claim. Let $(u^{h,j,k})_{k=1}^{\infty}$ be the minimizing sequence to the variational principle $\Phi^{**}(v; u^{h,j-1})$ converging to $u^{h,j}$ weakly in $H_0^1(\Omega)$ and strongly in $L^2(\Omega)$. Recall that $(\nabla u^{h,j,k})_{k=1}^{\infty}$ generates $\boldsymbol{\nu}^{h,j}$. For all $\zeta \in H_0^1(\Omega)$, we have

$$\begin{aligned}\int_\Omega p(\nabla u^{h,j,k}) \cdot \nabla\zeta + \frac{u^{h,j,k} - u^{h,j-1}}{h}\zeta \, dx \overset{k\to\infty}{\longrightarrow} & \int_\Omega \langle \nu_x^{h,j}, p\rangle \cdot \nabla\zeta + \frac{u^{h,j} - u^{h,j-1}}{h}\zeta \, dx \\ = & \int_\Omega p(\nabla u^{h,j}) \cdot \nabla\zeta + \frac{u^{h,j} - u^{h,j-1}}{h}\zeta \, dx \\ = & \ 0,\end{aligned}$$

because $p(\nabla u^{h,j,k}) \overset{k\to\infty}{\rightharpoonup} \langle \nu_x^{h,j}, p\rangle$ in $L^2(\Omega)$ and $\nabla\cdot\langle \nu_x^{h,j}, p\rangle = \nabla\cdot p(\nabla u^{h,j})$ in $H^{-1}(\Omega)$. It follows easily that

(i) $\nabla \cdot p(\nabla u^{h,j,k}) \overset{k\to\infty}{\longrightarrow} \nabla \cdot p(\nabla u^{h,j})$ in $H^{-1}(\Omega)$

by the estimate

$$\begin{aligned}\| \nabla \cdot p(\nabla u^{h,j,k}) - \nabla \cdot p(\nabla u^{h,j}) \|_{H^{-1}(\Omega)} &= \sup_{\|\zeta\|_{H_0^1(\Omega)}=1} \left| \int_\Omega [p(\nabla u^{h,j,k}) - p(\nabla u^{h,j})] \cdot \nabla\zeta \, dx \right| \\ \text{(for all sufficiently large } k) &\le \sup_{\|\zeta\|_{H_0^1(\Omega)}=1} \left| \int_\Omega \frac{u^{h,j,k} - u^{h,j}}{h}\zeta \, dx \right| + \epsilon \\ &\le \frac{1}{h} \| u^{h,j,k} - u^{h,j} \|_{L^2(\Omega)} + \epsilon \\ &\overset{k\to\infty}{\longrightarrow} \epsilon \qquad \text{for any } \epsilon > 0,\end{aligned}$$

since $u^{h,j,k} \overset{k\to\infty}{\longrightarrow} u^{h,j}$ in $L^2(\Omega)$ strongly. Recalling Remark 1 to Lemma 2.4 and noting that p, id $\in \mathcal{F}_0$ and $p \cdot \mathrm{id} \in \mathcal{E}_0$, we have as $k \to \infty$,

(ii) $p(\nabla u^{h,j,k}) \cdot \nabla u^{h,j,k} \rightharpoonup \langle \nu_x^{h,j}, p \cdot \mathrm{id}\rangle$ in $L^1(\Omega)$,

(iii) $p(\nabla u^{h,j,k}) \rightharpoonup \langle \nu_x^{h,j}, p\rangle$ in $L^2(\Omega)$,

(iv) $\nabla u^{h,j,k} \rightharpoonup \langle \nu_x^{h,j}, \mathrm{id}\rangle$ in $L^2(\Omega)$.

Now by the div-curl lemma (see [E], [T], or [Mu]), or by direct computation and using the H^1-strong convergence in (i), we have from (i), (iii), (iv),

$$p(\nabla u^{h,j,k}) \cdot \nabla u^{h,j,k} \overset{k\to+\infty}{\rightharpoonup} \langle \nu_x^{h,j}, p\rangle \cdot \langle \nu_x^{h,j}, \mathrm{id}\rangle$$

in the sense of distributions; by (ii) above and recalling (17) we have the claim.

Step 2 (passing to the limit). By (21) and Step 1 we have

$$\langle \boldsymbol{\nu}^h, q \cdot \mathrm{id}\rangle = \langle \boldsymbol{\nu}^h, q\rangle \cdot \langle \boldsymbol{\nu}^h, \mathrm{id}\rangle \qquad x \text{ a.e. in } \Omega \ \forall t \ge 0.$$

We may apply Lemma 2.4 to $(\boldsymbol{\nu}^h)_{h>0}$ to pass to a limit point as $h \to 0$. We obtain a subsequence, not relabeled, such that for each $T \ge 0$,

(v) $\langle \boldsymbol{\nu}^h, q \cdot \mathrm{id}\rangle \rightharpoonup \langle \boldsymbol{\nu}, q \cdot id\rangle$ in $L^1(Q_T)$,

(vi) $\langle \boldsymbol{\nu}^h, q\rangle \rightharpoonup \langle \boldsymbol{\nu}, q\rangle$ in $L^2(Q_T)$,

(vii) $\langle \boldsymbol{\nu}^h, \mathrm{id}\rangle \rightharpoonup \langle \boldsymbol{\nu}, \mathrm{id}\rangle$ in $L^2(Q_T)$.

Using the div-curl lemma as in Step 1 we obtain (30). □

THEOREM 4.2 (Uniqueness and continuity with respect to initial data). *There is a unique function* $u : Q_\infty \longrightarrow \mathbb{R}$ *with* $u \in H^1_{0,loc}(Q_\infty)$ *and* $u(\cdot, 0) = u_0$ *for which there exists a parametrized probability measure* $\boldsymbol{\nu} = (\nu_{x,t})_{(x,t)\in Q_\infty}$ *so that* (9), (10), (28), *and* (30) *are true. Under the same conditions,* $u_0 \mapsto u(\cdot, t)$ *is continuous from* $L^2(\Omega)$ *into* $L^2(\Omega)$ *for each* $t \geq 0$ (*and also into* $L^2(Q_T)$ *for each* $T > 0$).

Proof. Suppose $(u, \boldsymbol{\nu})$ and $(w, \boldsymbol{\mu})$ are two Young measure solutions to $\mathcal{P}$ with initial data u_0 and w_0, respectively. Apply equation (9) using $(u - w)\chi_{[0,T]}$ as the test function[2] in the previous section and against $\overline{q}^\nu$ and $\overline{q}^\mu$ and subtract to obtain (where the shorthand notation $\overline{q}^\nu$ is used for $\langle \boldsymbol{\nu}, q\rangle$ and similarly for $\overline{q}^\mu$)

$$\int_0^T \int_\Omega (\overline{q}^\nu - \overline{q}^\mu) \cdot (\overline{\mathrm{id}}^\nu - \overline{\mathrm{id}}^\mu)\, dx\, dt = -\int_0^T \int_\Omega \frac{\partial(u-w)}{\partial t}(u - w)\, dx\, dt$$

$$(31) \qquad = -\frac{1}{2}\Big(\| (u(\cdot, T) - w(\cdot, T) \|^2_{L^2(\Omega)} - \|u_0 - w_0\|^2_{L^2(\Omega)}\Big).$$

By Lemma 4.1 and (28),

$$\text{left-hand side (l.h.s.) (31)} = \int_0^T \int_\Omega \Big(\overline{p \cdot \mathrm{id}}^\nu + \overline{p \cdot \mathrm{id}}^\mu - \overline{p}^\nu \cdot \overline{\mathrm{id}}^\mu - \overline{p}^\mu \cdot \overline{\mathrm{id}}^\nu\Big)\, dx\, dt$$
$$\geq 0,$$

because the integrand above is precisely the quantity

$$\int_{\mathbb{R}^N} \int_{\mathbb{R}^N} (\nabla\phi^{**}(\alpha) - \nabla\phi^{**}(\beta)) \cdot (\alpha - \beta)\ \nu_{x,t}(d\alpha)\, \mu_{x,t}(d\beta),$$

which is nonnegative by the convexity of ϕ^{**}. This implies for (31)

$$(32) \qquad \| (u(\cdot, T) - w(\cdot, T) \|^2_{L^2(\Omega)} \leq \|u_0 - w_0\|^2_{L^2(\Omega)} \qquad \forall T > 0,$$

which is the continuity with respect to initial data. When $u_0 = w_0$ is used in (32) we have

$$u(\cdot, T) = w(\cdot, T) \qquad x \text{ a.e. in } \Omega\ \forall T > 0$$

and this shows uniqueness. □

Remark 1 (Comparison between classical and Young measure solutions). The statement of uniqueness does not depend on the method of extracting a Young measure solution for $\mathcal{P}$ and does not require that $\boldsymbol{\nu}$ be a Young measure, only that $\overline{q} = \overline{p}$ and the independence property of Lemma 4.1 hold. In particular, if $(u, \delta_{\nabla u})$ is a classical solution to $\mathcal{P}$ satisfying $q(\nabla u) = p(\nabla u)$, a weaker condition than (29), by uniqueness it coincides with the Young measure solution provided by Theorem 3.1 and (29) follows (the independence property is automatically satisfied by classical solutions). We note that there is no claim that the parametrized measure ν is unique; this is false in general.

[2] This is allowed by Remark 1 to the definition of Young measure solutions to $\mathcal{P}$ and (27).

The following lemma gives some properties of the solution $(u, \boldsymbol{\nu})$ which are consequences of the convexity of ϕ^{**} and the independence property. Most will be useful in establishing the uniqueness of the asymptotic limit (in §3).

LEMMA 4.3 (Further properties of the Young measure solution). *Let $(u, \boldsymbol{\nu})$ be the solution to $\mathcal{P}$ and $(u^h)_{h>0}$ as in the proof of the existence theorem (Theorem 3.1). Then the following are true:*

1. $\nabla \cdot p(\nabla u) = \nabla \cdot \overline{p} = \nabla \cdot \overline{q}$ *in $H^{-1}(Q_\infty)$. These equalities also hold in $L^2(Q_\infty)$ since the existence theorem yields $u_t \in L^2(Q_\infty)$ (recall that by (28) $\overline{p} = \overline{q}$ (x,t) a.e. in Q_τ $\forall \tau \in \mathbb{R}^+$).*
2. (i) *For each $T \geq 0$, $u^h \in C\left([0,T];\, L^2(\Omega)\right)$ and $(u^h)_{h>0}$ is Cauchy in $C\left([0,T];\, L^2(\Omega)\right)$.*
   (ii) *$u \in C\left([0,T];\, L^2(\Omega)\right)$; that is, $u(\cdot,t) \to u(\cdot,t_0)$ in $L^2(\Omega)$ as $t \to t_0$ for each $t_0 \geq 0$. In particular, $u(\cdot,t) \to u_0$ in $L^2(\Omega)$ as $t \to 0$.*
3. (i) *$t \mapsto \|u(\cdot,t)\|_{L^2(\Omega)}$ is decreasing (and therefore has a limit as $t \nearrow +\infty$).*
   (ii) *$t \mapsto \|u(\cdot,\delta+t) - u(\cdot,t)\|_{L^2(\Omega)}$ is decreasing for each $\delta \geq 0$.*
   (iii) *The integral*

$$\int_0^{+\infty}\int_\Omega \overline{q}\cdot \nabla u \, dx\, dt$$

exists.

Proof. 1. Fix $T > 0$, let w^h be given by (20), and let $\zeta \in H^1(Q_T)$ with $\zeta(\cdot,t) \in H_0^1(\Omega)$, for t a.e. in $[0,T]$. By Remark 1 to the definition (see §3) $\zeta - w^h$ is an admissible test function in equation (24). Using the convexity of ϕ^{**} and (24) we know that

$$\int_0^T\int_\Omega p(\nabla\zeta)\cdot\nabla(\zeta - w^h)\,dx\,dt \geq \int_0^T\int_\Omega p(\nabla w^h)\cdot\nabla(\zeta - w^h)\,dx\,dt$$
$$= -\int_0^T\int_\Omega \frac{\partial u^h}{\partial t}(\zeta - w^h)\,dx\,dt.$$

Letting $h \to 0$ we obtain

$$\int_0^T\int_\Omega p(\nabla\zeta)\cdot\nabla(\zeta - u) \;\geq\; -\int_0^T\int_\Omega \frac{\partial u}{\partial t}(\zeta - u) \;=\; \int_0^T\int_\Omega \overline{q}\cdot\nabla(\zeta - u).$$

Choosing $\zeta = u + \lambda(\theta - u)$ for $\theta \in H_0^1(Q_T)$ and letting $\lambda \to 0^+$ we obtain

$$\int_0^T\int_\Omega p(\nabla u)\cdot\nabla(\theta - u) \;\geq\; -\int_0^T\int_\Omega \overline{q}\cdot\nabla(\theta - u) \qquad \forall\theta \in H_0^1(Q_T).$$

Replacing $\theta - u$ with its negative we obtain equality above and this proves 1.

2. Fix $T > 0$. Recall that

$$u^h(x,t) = u^{h,j}(x) + \left(\frac{t}{h} - j\right)(u^{h,j+1} - u^{h,j})(x)$$

for $hj \leq t < h(j+1)$. When $hj \leq t, s < h(j+1)$,

$$\|u^h(\cdot,t) - u^h(\cdot,s)\|_{L^2(\Omega)} \;=\; \frac{|t-s|}{h}\,\|u^{h,j+1} - u^{h,j}\|_{L^2(\Omega)}.$$

Also, by the uniform estimates in [KP1],

$$\sup_{t\geq 0}\|u^h(\cdot,t)\|_{L^2(\Omega)} = \sup_{j\geq 0}\|u^{h,j}\|_{L^2(\Omega)} \leq M,$$

which shows that $t \mapsto u^h(\cdot,t)$ is (uniformly) continuous and bounded on $\mathbb{R}^+$ into $L^2(\Omega)$.

Set $U^{h,h'} := u^h - u^{h'} \in H^1(Q_T)$; we have

$$\begin{aligned}\|U^{h,h'}\|_{L^2(\Omega)}(T) &= \int_0^T\int_\Omega 2U^{h,h'}U_t^{h,h'}\,dx\,dt\\ &\leq \|U^{h,h'}\|_{L^2(Q_T)}\|U_t^{h,h'}\|_{H^1(Q_T)},\end{aligned}$$

and since $(u^h)_{h>0}$ converges in $L^2(Q_T)$ and is bounded in $H^1(Q_T)$, we see

$$\lim_{h,h'\to 0}\ \sup_{t\geq 0}\|u^h - u^{h'}\|_{L^2(\Omega)}(t) \ = \ 0.$$

Therefore, $(u^h)_{h>0}$ is Cauchy in $C\left(\mathbb{R}^+;\ L^2(\Omega)\right)$. This shows 2.

3. For $0 \leq s \leq t$ apply (9) with $u\chi_{[s,t]}$ as the test function and obtain

$$\begin{aligned}\int_s^t\int_\Omega \overline{q}\cdot\nabla u &= -\int_s^t\int_\Omega \frac{\partial u}{\partial t}u\\ &= -\frac{1}{2}\left(\|u(\cdot,t)\|^2_{L^2(\Omega)} - \|u(\cdot,s)\|^2_{L^2(\Omega)}\right).\end{aligned}\tag{33}$$

Using (28), the convexity of ϕ^{**}, and the independence relation (30) as in the proof of Theorem 4.2, together with the assumption $p(0) = 0$, we conclude that the l.h.s. of (33) is nonnegative and (i) follows. Letting $s = 0$ and $t \to +\infty$ in the l.h.s. of (33), we then obtain (iii).

Notice that by the uniqueness of solution the pair $(u^\delta, \boldsymbol{\nu}^\delta) \equiv \left(u(\cdot,\delta+\cdot),\ \boldsymbol{\nu}_{(\cdot,\delta+\cdot)}\right)$ solves $\mathcal{P}$ with initial data $u(\cdot,\delta)$. For fixed $0 \leq s \leq t$ we apply (9) to each of the solution pairs $(u^\delta,\boldsymbol{\nu}^\delta)$ and $(u,\boldsymbol{\nu})$ using $(u^\delta - u)\chi_{[s,t]}$ as a test function and subtract the two equations. Arguing as in (i) yields (ii). □

4.2. Stability: Maximum and comparison principles. We investigate the stability of the Young measure solution. We show that a maximum principle and a comparison result are satisfied. We conclude the section with a localization property of the solution $(u,\boldsymbol{\nu})$, a corollary of the comparison principle.

THEOREM 4.4 (Maximum principle). *Let $(u,\boldsymbol{\nu})$ and $(w,\boldsymbol{\mu})$ solve with initial data u_0 and w_0, respectively. Then (x,t) a.e. in $\overline{Q_\infty}$,*

$$-\operatorname{ess\,sup}_{x\in\overline{\Omega}}(u_0 - w_0)^- \leq u(x,t) - w(x,t) \leq \operatorname{ess\,sup}_{x\in\overline{\Omega}}(u_0 - w_0)^+.\tag{34}$$

Proof. Set

$$K := \operatorname{ess\,sup}_{x\in\overline{\Omega}}(u_0 - w_0)^+.$$

We introduce auxiliary functions as in the proof for a maximum principle for the solution of the heat equation with H^1 data (cf. [Br]). Fix $G \in C^1(\mathbb{R})$ such that $G = 0$

on $(-\infty, 0]$ and is strictly increasing with $0 < G^{'} \leq M$ on $(0, +\infty)$. For $t \geq 0$ define the functions

$$H(t) := \int_0^t G(s)\, ds$$

and

$$\psi(t) := \int_\Omega H\left(u(x,t) - w(x,t) - K\right)\, dx.$$

Then $\psi \in C(\mathbb{R}^+) \cap H^1(\mathbb{R}^+)$, $\psi(0) = 0$ and $\psi \geq 0$ on $\mathbb{R}^+$. Note that $G(u - w - K) \in H_0^1(\Omega)$ so that it is an admissible test function in (9); for all $T \geq 0$,

$$\begin{aligned}\int_0^T \psi'(t)\, dt &= \int_0^T \int_\Omega G\left(u(x,t) - w(x,t) - K\right) \frac{\partial(u-w)}{\partial t}\, dx\, dt \\ &= -\int_0^T \int_\Omega (\overline{q}^\nu - \overline{q}^\mu) \cdot \nabla(u - w)\, G^{'}\left(u - w - K\right) dx\, dt \\ &\leq 0,\end{aligned}$$

because $(\overline{q}^\nu - \overline{q}^\mu) \cdot \nabla(u - w) \geq 0$ (as in the proof of the Uniqueness Theorem 4.2), and $G^{'} \geq 0$. Hence $\psi \equiv 0$ and $H(u - w - K) \;=\; 0$ (x,t) a.e. in $\overline{Q_\infty}$, or

$$\int_0^{u(x,t)-w(x,t)-K} G(s)\, ds \;=\; 0 \qquad (x,t) \text{ a.e. in } \overline{Q_\infty},$$

which by the choice of G implies

$$u(x,t) - w(x,t) - K \;\leq\; 0 \qquad (x,t) \text{ a.e. in } \overline{Q_\infty}.$$

Reversing the roles of u and w we obtain the lower bound in (34). □

LEMMA 4.5 (Comparison principle). *Assume $(u, \boldsymbol{\nu})$ and $(v, \boldsymbol{\mu})$ are the solutions to $\mathcal{P}$ with to initial data u_0 and v_0, respectively. Assume further that*

$$u_0 \;\geq\; v_0 \qquad a.e. \text{ } in\ \Omega.$$

Then

$$u \;\geq\; v \qquad (x,t) \text{ } a.e. \text{ } in\ Q_\infty.$$

Proof. Although this follows directly from the maximum principle (34) we give an independent proof which can also be modified to prove (34). Let $w = \max(u, v)$ in Q_∞. It suffices to show that

$$(v - u)^+ \;=\; 0 \qquad (x,t) \text{ a.e. in } Q_\infty.$$

We apply (9) for each solution noting that $w - u = (v - u)^+$ is admissible as a test function.

$$\int_0^T \int_\Omega \overline{q}^\nu \cdot \nabla(w - u) + u_t(w - u)\ dx\, dt \;=\; 0,$$

$$\int_0^T \int_\Omega \overline{q}^\mu \cdot \nabla(w - u) + v_t(w - u)\ dx\, dt \;=\; 0.$$

By subtraction we obtain

$$\begin{aligned}\int_0^T \int_\Omega (\overline{q}^\mu - \overline{q}^\nu) \cdot \nabla (v-u)^+ &= -\frac{1}{2} \int_0^T \int_\Omega (v_t - u_t)(v-u)^+ \, dx\, dt \\ &= \int_0^T \frac{d}{dt} \int_\Omega (v-u)^+ \, dx\, dt \\ &= -\frac{1}{2} \{ \|(v-u)^+\|^2_{L^2(\Omega)}(T) - \|(v_0 - u_0)^+\|^2_{L^2(\Omega)} \} \\ &= -\frac{1}{2} \|(v-u)^+\|^2_{L^2(\Omega)}(T).\end{aligned}$$

Since $(\overline{q}^\mu - \overline{q}^\nu) \cdot \nabla(v-u)^+ \geq 0$, we conclude that

$$\|(v-u)^+\|^2_{L^2(\Omega)}(T) \;=\; 0 \qquad \forall T \geq 0;$$

that is, $v \leq u$, (x,t) a.e. in Q_∞. □

COROLLARY 4.6 (localization). *Assume $\omega \subseteq \Omega$ is open with Lipschitz boundary and let $(u, \boldsymbol{\nu})$ be the solution to $\mathcal{P}$. Let v be the restriction of u to ω. Then v is a solution with respect to initial data v_0, the restriction of u_0 to ω.*

Proof. Suppose χ is the solution on ω (with v_0 initial data). Apply the comparison result to the differences $\chi - v$ and $v - \chi$. □

5. Asymptotic analysis and the equilibrium Young measure solution.

We investigate the asymptotic behavior of the solution as $t \longrightarrow +\infty$ and establish the following theorem.

THEOREM 5.1. *Let $(u, \boldsymbol{\nu})$ be the unique solution of $\mathcal{P}$; there exists a unique $z \in H^1_0(\Omega)$ and a Young measure $\boldsymbol{\nu}^\infty = (\nu^\infty_{x,t})_{(x,t) \in Q_\infty}$ such that*

$$(35) \qquad u(\cdot, t) \;\longrightarrow\; z \qquad \textit{weakly in } H^1_0(\Omega) \textit{ and strongly in } L^2(\Omega) \textit{ as } t \longrightarrow +\infty$$

(these limits exist without restriction to a subsequence in time).

$$(36) \qquad t \mapsto \|u(\cdot,t) - z\|_{L^2(\Omega)} \qquad \textit{is decreasing,}$$

$$(37) \qquad \nabla \cdot \langle \boldsymbol{\nu}^\infty, q \rangle \;=\; 0 \qquad \textit{in } H^{-1}(Q_\infty),$$

$$(38) \qquad \nabla z = \langle \boldsymbol{\nu}^\infty, \mathrm{id} \rangle \qquad \textit{a.e. in } \Omega \ (\textit{independent of time})$$

so $(z, \boldsymbol{\nu}^\infty)$ is an equilibrium Young measure solution of the steady-state version of $\mathcal{P}$.

$$(39) \qquad \langle \nu^\infty_{x,t}, q \cdot \mathrm{id} \rangle = 0 \qquad \textit{a.e. in } Q_T,$$

$$(40) \qquad \operatorname{supp} \boldsymbol{\nu}^\infty \;\subseteq\; \{\lambda : q(\lambda) \cdot \lambda = 0\} \cap \{\phi^{**} = \phi = 0\}.$$

DEFINITION. *With z, $\boldsymbol{\nu}^\infty$ as in the theorem, we call z the* asymptotic limit *of u and $\boldsymbol{\nu}^\infty$ the* asymptotic Young measure; *we call the pair $(z, \boldsymbol{\nu}^\infty)$ the* equilibrium (Young measure) solution *of $\mathcal{P}$.*

We define the set of weak limit points of $(u(\cdot,t))_{t \geq 0}$ in $H^1_0(\Omega)$

$$W_\omega(u_0) := \{ z \in H^1_0(\Omega) \,|\, \exists (t_n)_{n \geq 1} \nearrow +\infty \text{ with } u(\cdot, t_n) \overset{w-s}{\longrightarrow} z \}.$$

The notation

$$u(\cdot,t_n) \overset{w-s}{\longrightarrow} z \qquad \text{in } H^1 - L^2$$

indicates that the sequence converges weakly in H^1 and strongly in L^2, which we may always achieve by reducing to a subsequence using the Rellich theorem. Note that $W_\omega(u_0)$ is nonempty since $u \in L^\infty\left(\mathbb{R}^+; H_0^1(\Omega)\right)$. Theorem 5.1 establishes that $W_\omega(u_0)$ consists of exactly one function.

We begin by describing some properties of all functions in $W_\omega(u_0)$.

LEMMA 5.2. *Let $z \in W_\omega(u_0)$ and $t_n \to +\infty$ along which $u(\cdot,t_n) \overset{w-s}{\longrightarrow} z$. Then $\forall t \geq 0$*

$$u(\cdot,t_n+t) \overset{w-s}{\longrightarrow} z \qquad in\ H^1(\Omega) - L^2(\Omega), \tag{41}$$

$$u(\cdot,t_n+\cdot) \overset{w-s}{\longrightarrow} z \qquad in\ H^1_{loc}(Q_\infty) - L^2_{loc}(Q_\infty) \tag{42}$$

as $n \to +\infty$ (without restricting to subsequences).

Proof. Fix $t \geq 0$. Since $u \in L^\infty\left(\mathbb{R}^+; H_0^1(\Omega)\right)$, the sequence $(u(\cdot,t_n+t))_{n\geq 1}$ is bounded in $H^1(\Omega)$; hence for a subsequence $(n_j)_{j\geq 1}$ there exists $y(\cdot,t) \in H^1(\Omega)$ such that as $j \to +\infty$,

$$u(\cdot,t_{n_j}+t) \overset{w-s}{\longrightarrow} y(\cdot,t) \qquad \text{in } H^1(\Omega) - L^2(\Omega)$$

and, of course,

$$u(\cdot,t_{n_j}) \overset{w-s}{\longrightarrow} z \qquad \text{in } H^1(\Omega) - L^2(\Omega).$$

Note that $y(\cdot,t) \in W_\omega(u_0)$; we show that $y(\cdot,t) = z$. For all $\zeta \in \mathrm{L}^2(\Omega)$,

$$\begin{aligned}\int_\Omega (y(x,t) - z(x))\zeta(x)\,dx &= \lim_{j\to+\infty} \int_\Omega (u(x,t_{n_j}+t) - u(x,t_{n_j}))\zeta(x)dx \\ &= \lim_{j\to+\infty} \int_\Omega \int_{t_{n_j}}^{t_{n_j}+t} u_t(x,s)\zeta(x)\,ds\,dx \\ &\leq \|\zeta\|_{L^2(\Omega)} \limsup_{j\to+\infty} \|u_t\|_{L^2([t_{n_j},t_{n_j}+t]\times\Omega)}\sqrt{t} \\ &= 0,\end{aligned}$$

since $u_t \in L^2(Q_\infty)$. This shows $y(\cdot,t) = z$ for x a.e. in Ω; as a result the whole sequence converges to z and (41) is true.

Fix $T > 0$. The sequence $(u(\cdot,t_n+\cdot))_{t\geq 0}$ is bounded in $H^1(Q_T)$ and thus we can find $\chi \in H^1(Q_T)$ and a subsequence $(n_j)_{j\geq 1}$ along which

$$u(\cdot,t_{n_j}+\cdot) \overset{w-s}{\longrightarrow} \chi \qquad \text{in } H^1(Q_T) - L^2(Q_T).$$

Choose $\zeta(x)\eta(t) \in H^1(Q_T)$ with $\eta(t)$ defined for each t. By (41),

$$\int_\Omega u(x,t_{n_j}+t)\zeta(x)\,dx \overset{j\to+\infty}{\longrightarrow} \int_\Omega z(x)\zeta(x)\,dx$$

for each $t > 0$. Thus

$$\eta(t)\int_\Omega u(x,t_{n_j}+t)\zeta(x)\,dx \overset{j\to+\infty}{\longrightarrow} \eta(t)\int_\Omega z(x)\zeta(x)\,dx \tag{43}$$

pointwise in t and the Lebesgue dominated convergence theorem applies to (43) to give

$$\int_0^T\int_\Omega \eta(t)\zeta(x)u(x,t_{n_j}+t)\,dx\,dt \overset{n\to+\infty}{\longrightarrow} \int_0^T\int_\Omega \eta(t)\zeta(x)z(x)\,dx\,dt$$

$$\text{(by assumption)} \quad = \quad \int_0^T\int_\Omega \eta(t)\zeta(x)\chi(x,t)\,dx\,dt.$$

By the density of separable functions in L^2 this implies

$$\chi(\cdot,t)=z(\cdot) \quad \forall t>0 \quad x \text{ a.e. in } \Omega.$$

We conclude that no reduction to a subsequence is necessary and (42) obtains. □

Proof of Theorem 5.1. Fix $z\in W_w(u_0)$ and $t_n\to+\infty$ along which $u(\cdot,t_n)\overset{w-s}{\longrightarrow} z$. We define

$$u^n(\cdot,\cdot):=u(\cdot,t_n+\cdot),$$

$$\boldsymbol{\nu}^n \;=\; (\nu^n_{x,t})_{(x,t)\in Q_\infty} \;:=\; (\nu_{x,t_n+t})_{(x,t)\in Q_\infty}.$$

Then $(u^n,\boldsymbol{\nu}^n)$ is the solution with respect to initial data $u(\cdot,t_n)$ for $n\geq 0$. By Lemma 5.2 we know that as $n\to+\infty$,

$$u^n(\cdot,t)\overset{w-s}{\longrightarrow} z \qquad \text{in } H^1_0(\Omega) \;-\; L^2(\Omega)\;\forall t\in\mathbb{R}^+,$$

$$u^n\overset{w-s}{\longrightarrow} z \qquad \text{in } H^1_{loc}(Q_\infty) \;-\; L^2_{loc}(Q_\infty).$$

Since z is independent of t it follows that

$$\frac{\partial u^n}{\partial t} \overset{n\to+\infty}{\longrightarrow} 0 \qquad \text{in } L^2(Q_\infty).$$

In addition, note that $(\boldsymbol{\nu}^n)_{n\geq0}$ is bounded in $L^1(Q_\infty,\mathcal{E}_0')$. By Remark 2 to Lemma 2.4 there exists a Young measure $\boldsymbol{\nu}^\infty=(\nu^\infty_{x,t})_{(x,t)\in Q_\infty}$ (generated by spatial gradients as in the remark) satisfying

$$\nu^\infty \;\in\; L^\infty(Q_\infty;\,M(\mathbb{R}^N)) \;\cap\; L^2_{loc}(Q_\infty;\,\mathcal{F}_0') \;\cap\; L^1_{loc}(Q_\infty;\,\mathcal{E}_0').$$

For each n,

$$\int_0^\tau\int_\Omega \langle\boldsymbol{\nu}^n,q\rangle\cdot\nabla\zeta+\frac{\partial u^n}{\partial t}\zeta\,dx\,dt \;=\; 0 \qquad \forall\zeta\in H^1_0(Q_\tau)\,\forall\tau\in[0,+\infty],$$

and we may pass to the limit as $n\to\infty$. (Note that the nonlocal convergence in n is not guaranteed by Lemma 2.4 but by the boundedness of $(\frac{\partial u^n}{\partial t})_{n\geq0}$ in $L^2(Q_\infty)$); we obtain

$$\int_0^\tau\int_\Omega \langle\boldsymbol{\nu}^\infty,q\rangle\cdot\nabla\zeta\,dx\,dt \;=\; 0 \qquad \forall\zeta\in H^1_0(Q_\tau)\,\forall\tau\in[0,+\infty]$$

or, equivalently,

$$\nabla\cdot\langle\boldsymbol{\nu}^\infty,q\rangle \;=\; 0 \qquad \text{in } H^{-1}(Q_\infty).$$

Recall that $\langle \boldsymbol{\nu}^n, \mathrm{id}\rangle = \nabla u^n$ and converges to ∇z weakly in $L^2_{loc}(Q_\infty)$; in addition, by Lemma 2.4,

$$\langle \boldsymbol{\nu}^n, \mathrm{id}\rangle \rightharpoonup \langle \boldsymbol{\nu}^\infty, \mathrm{id}\rangle \qquad \text{in } L^2_{loc}(Q_\infty).$$

Thus,

$$\nabla z \;=\; \langle \boldsymbol{\nu}^\infty, \mathrm{id}\rangle \qquad x \text{ a.e. in } \Omega.$$

From (37) and (38) we infer that $(z, \boldsymbol{\nu}^\infty)$ solves the stationary problem associated to $\mathcal{P}$ and so the independence lemma (Lemma 4.1) implies

$$\langle \boldsymbol{\nu}^\infty, q\cdot \mathrm{id}\rangle = \langle \boldsymbol{\nu}^\infty, q\rangle \cdot \langle \boldsymbol{\nu}^\infty, \mathrm{id}\rangle \qquad (x,t) \text{ a.e. in } Q_\infty$$

and, as before,

$$\langle \boldsymbol{\nu}^\infty, q\cdot \mathrm{id}\rangle \ge 0 \qquad \text{a.e. in } \Omega. \tag{44}$$

On the other hand, for each $T \ge 0$, by Lemma 4.3 (3i),

$$\begin{aligned}\int_0^T \int_\Omega \langle \nu^n_{x,t}, q\rangle \cdot \nabla u^n(x,t)\, dx\, dt \;&=\; -\frac{1}{2}\left(\|u(\cdot, T+t_n)\|^2 - \|u(\cdot, t_n)\|^2\right)\\ &\overset{n\to+\infty}{\longrightarrow} 0\end{aligned}$$

and applying the independence lemma (Lemma 4.1) on the l.h.s. we obtain

$$\int_0^T \int_\Omega \langle \boldsymbol{\nu}^n, q\cdot \mathrm{id}\rangle\, dx\, dt \;\overset{n\to+\infty}{\longrightarrow}\; 0$$

or

$$\int_0^T \int_\Omega \langle \boldsymbol{\nu}^\infty, q\cdot \mathrm{id}\rangle\, dx\, dt \;=\; 0.$$

This proves (39). By (29) and (44) we conclude

$$\operatorname{supp} \boldsymbol{\nu}^\infty \subseteq \{\lambda : q(\lambda)\cdot\lambda = 0\} \cap \{\phi^{**} = \phi\}.$$

The proof of (40) will be completed by Corollary 6.2 in the next section.

It remains to show (35) and (36). Let $0 \le s \le t$ and apply (9) to the solutions of $\mathcal{P}$ $(u, \boldsymbol{\nu})$ and $(z, \boldsymbol{\nu}^\infty)$ (corresponding to initial data u_0 and z, respectively); we have (using the notation $\bar{q}^\infty = \langle \boldsymbol{\nu}^\infty, q\rangle$),

$$\begin{aligned}\int_s^t \int_\Omega (\bar{q} - \bar{q}^\infty)\cdot \nabla(u - z)\, dx\, dt &= -\int_s^t \int_\Omega \frac{\partial(u-z)}{\partial t}(u - z)\, dx\, dt\\ &= -\frac{1}{2}\left(\|u(\cdot,t) - z\|^2_{L^2(\Omega)} - \|u(\cdot,s) - z\|^2_{L^2(\Omega)}\right)\\ &\ge 0,\end{aligned}$$

so that

$$t \mapsto \|u(\cdot,t) - z\|_{L^2(\Omega)}$$

is decreasing and its limit as $t \to +\infty$ exists; it is zero because $z \in W_w(u_0)$ and by the Rellich theorem a subsequence exists along which $u(\cdot, t_j) \overset{j\to\infty}{\longrightarrow} z$ in $L^2(\Omega)$. This finishes the proof of the theorem. □

CONCLUSION. *Let $(u, \boldsymbol{\nu})$ be the Young measure solution to $\mathcal{P}$ with initial data u_0. Then $W_w(u_0) = \{z\}$, i.e., the $L^2(\Omega)$ asymptotic limit of u is unique and the equilibrium solution $(z, \boldsymbol{\nu}^\infty)$ solves the steady-state problem*

$$\nabla \cdot \langle \boldsymbol{\nu}^\infty, q\rangle = 0 \quad \text{in } H^{-1}(Q_\infty).$$

6. Energy. Define the *energy function*

$$E(t) := \int_\Omega \phi^{**}(\nabla u)(x,t)\,dx \qquad \text{for } t \geq 0. \tag{45}$$

The results in this section serve to justify the term *energy* for the function in (45) and show that it vanishes at infinity. Throughout this section $(u, \boldsymbol{\nu})$ is the solution of $\mathcal{P}$ and $(z, \boldsymbol{\nu}^\infty)$ the equilibrium solution.

THEOREM 6.1. *Let E be given by* (45). *Then $E \in L^1(\mathbb{R}^+)$ and E is a decreasing function of t. Moreover,*

$$\int_\Omega \phi^{**}(\nabla u)(x,t)\,dx \searrow 0 \qquad \text{as } t \nearrow +\infty. \tag{46}$$

Proof. For $0 \leq s \leq t$,

$$\begin{aligned} \int_s^t E(\tau)\,d\tau &= \int_s^t \int_\Omega \phi^{**}(\nabla u)(x,\tau)\,dx\,d\tau \\ \text{(since } \phi^{**} \text{ is convex and } \phi^{**}(0) = 0) \quad &\leq \int_s^t \int_\Omega p(\nabla u)\cdot \nabla u\,dx\,d\tau \\ &= -\int_s^t \int_\Omega u_t u\,dx\,d\tau \\ &= -\frac{1}{2}\left(\|u\|^2_{L^2(\Omega)}(t) - \|u\|^2_{L^2(\Omega)}(s)\right) \\ &\longrightarrow 0^+ \quad \text{as } s,t \to +\infty, \end{aligned}$$

by Lemma 4.3 (3i) and (3iii). Therefore,

$$\int_0^{+\infty} \int_\Omega \phi^{**}(\nabla u)(x,t)\,dx\,dt < +\infty$$

and this shows that the energy is integrable.

Next we give the proof due to P. Pedregal that the energy is decreasing. For $T \geq 0$ fixed we have for all $t > 0$,

$$\begin{aligned} \int_T^{T+t} (E(s) - E(T))\,ds &= \int_T^{T+t} \int_\Omega (\phi^{**}(\nabla u(x,s)) - \phi^{**}(\nabla u(x,T)))\,dx\,ds \\ &\leq \int_T^{T+t} \int_\Omega p(\nabla u(x,s))\cdot(\nabla u(x,s) - \nabla u(x,T))\,dx\,ds \\ &= -\int_T^{T+t} \int_\Omega \frac{\partial(u(x,s) - u(x,T))}{\partial s}(u(x,s) - u(x,T))\,dx\,ds \end{aligned}$$

$$= -\frac{1}{2}\|u(\cdot, T+t) - u(\cdot, T)\|^2_{L^2(\Omega)}$$
$$\leq 0.$$

By the continuity of E this implies

$$E(T+t) \ \leq \ E(T)$$

for all $t > 0$ sufficiently small. Since T is arbitrary, this shows that E is in $L^1(\mathbb{R}^+)$ and decreasing, so (46) follows. □

COROLLARY 6.2. *The energy converges as* $t \to +\infty$ *and attains its minimum, i.e.,*

$$\lim_{t\to+\infty} \int_\Omega \phi^{**}(\nabla u)(x,t)\,dx \ = \ \int_\Omega \phi^{**}(\nabla z)(x)\,dx \ = \ 0. \tag{47}$$

Consequently, the asymptotic Young measure satisfies

$$\operatorname{supp} \boldsymbol{\nu}^\infty \subseteq \{\phi^{**} = \phi = 0\} \tag{48}$$

(*which completes the proof of* (40)).

Proof. Since ϕ^{**} is convex, the functional

$$u \mapsto \int_\Omega \phi^{**}(\nabla u)\,dx$$

is sequentially lower semicontinuous with respect to weak convergence in $H^1_0(\Omega)$. By Theorem 5.1 we have $u(\cdot,t) \rightharpoonup z$ in $H^1(\Omega)$ as $t \to +\infty$ and

$$\begin{aligned} 0 \ \leq \ \int_\Omega \phi^{**}(\nabla z)(x)\,dx &\leq \liminf_{t\to+\infty} \int_\Omega \phi^{**}(\nabla u)(x,t)\,dx \\ &= \lim_{t\to+\infty} \int_\Omega \phi^{**}(\nabla u)(x,t)\,dx \\ &= 0, \end{aligned}$$

because $E \in L^1(\mathbb{R}^+)$ and $\phi^{**} \geq 0$. From Jensen's inequality and (47) we obtain (48). □

The energy is also minimized asymptotically locally in the sense of the following lemma.

LEMMA 6.3. *For all* $A \subseteq \Omega$, *measurable*

$$\lim_{t\to+\infty} \int_A \phi^{**}(\nabla u)(x,t)\,dx \ = \ \int_A \phi^{**}(\nabla z)(x)\,dx \tag{49}$$

(but this limit is not necessarily monotone decreasing).

Proof. By (47) and by Theorem 2.2(ii) we conclude that $((\phi^{**}(\nabla u)(\cdot,t))_{t\geq 0}$ is weakly sequentially precompact in $L^1(\Omega)$ and (49) follows. □

Acknowledgments. The results reported here have also been submitted to the Graduate School of the University of Minnesota in partial fulfillment of the requirements for the degree of Doctor of Philosophy. I wish to express deep gratitude to my thesis supervisor, David Kinderlehrer, whose professional guidance and collaboration played a crucial role in the development of this work. In addition, I want to truly thank Irene Fonseca, Wilfrid Gangbo, William Hrusa, Norman Meyers, Pablo Pedregal, Luc Tartar, Thanos Tzavaras, Noel Walkington, to mention a few, whose assistance I have sought, received, and used repeatedly.

REFERENCES

[AF] E. Acerbi and N. Fusco, *Semicontinuity problems in the calculus of variations*, Arch. Rational Mech. Anal., 86 (1984), pp. 125–145.

[Ba] J. Ball, *A version of the fundamental theorem of Young measures*, in PDEs and Continuum Models of Phase Transitions, Rascle, Serre, and Slemrod, eds., Lecture Notes of Physics 344, Springer-Verlag, Berlin, New York, 1989, pp. 207–215.

[BM] J. Ball and F. Murat, *Remarks on Chacon's biting lemma*, Proc. Amer. Math. Soc., 107 (1989), pp. 655–663.

[BC] F. Bethuel, J. Coron, J. Ghidaglia, and A. Soyeur, *Heat flows and relaxed energies for harmonic maps*, in Nonlinear Diffusion Equations and Their Equilibrium States 3, Lloyd, Ni, Peletier, and Serrin, eds., Progress in Nonlinear Differential Equations, Vol. 7, Birkhäuser, Boston, 1992, pp. 99–109.

[Br] H. Brezis, *Analyse Fonctionnelle*, Masson, Paris, 1983.

[BZ] J. Ball and K. Zhang, *Lower semicontinuity of multiple integrals and the biting lemma*, Proc. Roy. Soc. Edinburgh Sect. A, 114A (1990), pp. 367–379.

[D] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, New York, 1989.

[E] L. Evans, *Weak convergence methods for nonlinear partial differential equations*, CBMS 74, American Mathematical Society, Providence, RI, 1990.

[H] K. Höllig, *Existence of infinitely many solutions for a forward backward heat equation*, Trans. Amer. Math. Soc., 278 (1983), pp. 299–316.

[HK] K. Horihata and N. Kikuchi, *A construction of solutions satisfying a Cacciopoli inequality for non-linear parabolic equations associated to a variational functional of harmonic type*, Boll. Un. Mat. Ital. A (7), 3 (1989), pp. 199–207.

[HN] K. Höllig and J. Nohel, *A diffusion equation with a nonmonotone constitutive function*, in Systems of Nonlinear Partial Differential Equations, J. Ball, ed., NATO ASI Series C, D. Reidel, Boston, MA, 1983, pp. 409–422.

[KP1] D. Kinderlehrer and P. Pedregal, *Weak convergence of integrands and the Young measure representation*, SIAM J. Math. Anal., 23 (1992), pp. 1–19.

[KP2] ———, *Remarks about the analysis of Gradient Young measures*, J. Geom. Anal., 4 (1994), pp. 59–90.

[Mo] C. Morrey, *Quasiconvexity and the semicontinuity of multiple integrals*, Pacific J. Math., 2 (1952), pp. 25–33.

[Mu] F. Murat, *Compacité par compensation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) (1978), pp. 490–507.

[Sc] M. Schonbek, *Convergence of solutions to nonlinear dispersive equations*, Comm. Partial Differential Equations, 7 (1982), pp. 959–1000.

[Sl] M. Slemrod, *Dynamics of measure valued solutions to a backward-forward heat equation*, J. Dynamics Differential Equations, 3 (1991), pp. 1–28.

[T] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics, Knops, ed., Heriot–Watt Symposium, Vol. IV, Pitman Research Notes in Mathematics, Pitman, Boston, 1979, pp. 136–192.

[Y] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969 (reprint by Chelsea, New York, 1980).

[Z] X. Zhou, *An evolution problem for plastic antiplanar shear*, Appl. Math. Optim., 25 (1992), pp. 263–285.

SIAM J. MATH. ANAL.
Vol. 27, No. 2, pp. 404–423, March 1996

005

ON THE CAHN–HILLIARD EQUATION WITH DEGENERATE MOBILITY[*]

CHARLES M. ELLIOTT[†] AND HARALD GARCKE[‡]

Abstract. An existence result for the Cahn–Hilliard equation with a concentration dependent diffusional mobility is presented. In particular, the mobility is allowed to vanish when the scaled concentration takes the values ±1, and it is shown that the solution is bounded by 1 in magnitude. Finally, applications of our method to other degenerate fourth-order parabolic equations are discussed.

Key words. Cahn–Hilliard equations, degenerate parabolic equations, nonlinear diffusion, phase transitions.

AMS subject classifications. 35K55, 35K65, 82C26

1. Introduction.

The Cahn–Hilliard equation

$$\begin{aligned} u_t &= -\nabla \cdot \mathbf{J}, \\ \mathbf{J} &= -B(u)\nabla w, \\ w &= -\gamma\Delta u + \Psi'(u), \qquad \gamma \in \mathbb{R}^+, \end{aligned} \tag{1.1}$$

was introduced to study phase separation in binary alloys (see Cahn and Hilliard [8, 9]). Although the Cahn–Hilliard equation has been intensively studied, little mathematical analysis has been done for a diffusional mobility B which depends on u (where u is the difference of the mass density of the two components of the alloy). A concentration-dependent mobility appeared in the original derivation of the Cahn–Hilliard equation (see [9]), and a thermodynamically reasonable choice is $B(u) = 1 - u^2$ (see [10, 11, 18]). The mathematical difficulty in studying the Cahn–Hilliard equation with a mobility like this lies in the degeneracy of B. On the other hand, there is hope that solutions which initially take values in the interval $[-1,1]$ will do so for all positive time (which is not true for fourth-order parabolic equations without degeneracy). We remark that only values in the interval $[-1,1]$ are physically meaningful.

The function Ψ represents the homogeneous free energy in the energy functional

$$\mathcal{E}(u) = \int_\Omega \Big(\frac{\gamma}{2} \mid \nabla u \mid^2 + \Psi(u)\Big)\, dx,$$

where $\Omega \subset \mathbb{R}^n$ ($n \in \mathbb{N}$) is a bounded domain with sufficiently smooth boundary. Possible choices for Ψ are

$$\begin{aligned} &\Psi(u) = (1-u^2)^2 \quad \text{and} \\ &\Psi(u) = \frac{\theta}{2}\left((1+u)\ln(1+u) + (1-u)\ln(1-u)\right) + F_0(u) \end{aligned} \tag{1.2}$$

[*] Received by the editors May 13, 1994; accepted for publication (in revised form) August 2, 1994.

[†] Centre for Mathematical Analysis and Its Applications, School of Mathematical and Physical Sciences, University of Sussex, Falmer, Brighton BN1 9QH, United Kingdom.

[‡] Institut für Angewandte Mathematik, Wegelerstraße 6, 53115 Bonn, Germany. The research of this author was supported by a Research Fellowship of the European Science Foundation held at the Centre for Mathematical Analysis and Its Applications, School of Mathematical and Physical Sciences, University of Sussex, United Kingdom.

with a smooth function F_0. In the case $F_0(u) = 1 - u^2$, one gets in the limit as $\theta \searrow 0$ the double obstacle potential (see the papers of Blowey and Elliott [5, 6] and Elliott and Luckhaus [14])

$$\Psi(u) = \begin{cases} 1 - u^2 & \text{if } |u| \leq 1, \\ \infty & \text{otherwise.} \end{cases}$$

In order to formulate an existence result for (1.1) in a general situation we make the following assumptions.

Let

$$\Psi(u) := \Psi_1(u) + \Psi_2(u)$$

with functions Ψ_1 and Ψ_2 such that

$$\|\Psi_2\|_{C^2[-1,1]} \leq C$$

and

$$\Psi_1 : (-1,1) \to \mathbb{R}$$

is convex and of the form

$$\Psi_1''(u) = (1 - u^2)^{-m} F(u) \quad (m \geq 1)$$

with a C^1-function $F : [-1,1] \to \mathbb{R}_0^+$. This means that we allow Ψ to be singular in the convex part as $|u| \to 1$. In particular, the logarithmic form (1.2) is a possible choice. Furthermore, we assume that the mobility is of the form

$$B(u) = (1 - u^2)^m \bar{B}(u)$$

with a C^1-function

$$\bar{B} : [-1,1] \to \mathbb{R}$$

which satisfies

$$b_0 \leq \bar{B}(u) \leq B_0 \qquad (B_0, b_0 > 0)$$

for $u \in [-1,1]$. We extend the defintion of B to all of $\mathbb{R}$ by $B(u) = 0$ for $|u| > 1$. Let

$$\Phi : (-1,1) \to \mathbb{R}_0^+$$

be defined by

$$\Phi''(u) = \frac{1}{B(u)}, \quad \Phi'(0) = 0, \quad \text{and } \Phi(0) = 0.$$

The following theorem states the existence of a weak solution to the Cahn–Hilliard equation with a nonconstant mobility as above on an arbitrary time interval $[0,T]$ ($T \in \mathbb{R}^+$) which fulfills the boundary conditions

$$\mathbf{n} \cdot \mathbf{J} = 0 \quad \text{and} \quad \mathbf{n} \cdot \nabla u = 0 \quad \text{on} \quad \partial\Omega \times (0,T),$$

where $\mathbf{n}$ is the outer normal to $\partial\Omega$.

THEOREM 1. *Let either $\partial\Omega \in C^{1,1}$ or Ω be convex and suppose that $u_0 \in H^1(\Omega)$ with $|u_0| \leq 1$ a.e. and*

$$\int_\Omega (\Psi(u_0) + \Phi(u_0)) \leq C, \qquad C \in \mathbb{R}^+.$$

Then there exists a pair $(u, \mathbf{J})$ such that

(a) $u \in L^2(0,T; H^2(\Omega)) \cap L^\infty(0,T; H^1(\Omega)) \cap C([0,T]; L^2(\Omega))$,

(b) $u_t \in L^2(0,T; (H^1(\Omega))')$,

(c) $u(0) = u_0$ *and* $\nabla u \cdot \mathbf{n} = 0$ *on* $\partial\Omega \times (0,T)$,

(d) $|u| \leq 1$ *a.e in* $\Omega_T := \Omega \times (0,T)$,

(e) $\mathbf{J} \in L^2(\Omega_T, \mathbb{R}^n)$

which satisfies $u_t = -\nabla \cdot \mathbf{J}$ in $L^2(0,T; (H^1(\Omega))')$, i.e.,

$$\int_0^T \langle \zeta(t), u_t(t) \rangle_{H^1,(H^1)'} = \int_{\Omega_T} \mathbf{J} \cdot \nabla\zeta$$

for all $\zeta \in L^2(0,T; H^1(\Omega))$ and

$$\mathbf{J} = -B(u)\nabla \cdot (-\gamma\Delta u + \Psi'(u))$$

in the following weak sense:

$$\int_{\Omega_T} \mathbf{J} \cdot \boldsymbol{\eta} = -\int_{\Omega_T} [\gamma\Delta u \nabla \cdot (B(u)\boldsymbol{\eta}) + (B\Psi'')(u)\nabla u \cdot \boldsymbol{\eta}]$$

for all $\boldsymbol{\eta} \in L^2(0,T; H^1(\Omega, \mathbb{R}^n)) \cap L^\infty(\Omega_T, \mathbb{R}^n)$ which fulfill $\boldsymbol{\eta} \cdot \mathbf{n} = 0$ on $\partial\Omega \times (0,T)$.

We point out that the nonlinearity $(B\Psi'')(u) = \bar{B}(u)F(u) + B(u)\Psi_2''(u)$ is bounded and therefore the last integral in the formulation of the theorem is well defined.

An existence result for the Cahn–Hilliard equation with a degenerate mobility in a one-dimensional situation has been established by Yin Jingxue [23]. The existence result we present is for arbitrary space dimensions and uses a weak formulation which is different from the formulation of Yin Jingxue. Furthermore, we allow the bulk energy Ψ to have singularities when B degenerates. We also refer to the work of Bernis and Friedman [3] for results on fourth-order degenerate parabolic equations in one space dimension.

In §4, we prove a similar existence result for a viscous Cahn-Hilliard-type equation of the form

$$\begin{aligned} u_t &= -\nabla \cdot \mathbf{J}, \\ \mathbf{J} &= -B(u)\nabla w, \\ w &= -\gamma\Delta u + \Psi'(u) + \alpha u_t, \qquad \alpha \in \mathbb{R}^+, \end{aligned}$$

where we assume the mobility B and the homogeneous free energy Ψ to be as above.

We want to point out that our result includes the case $B(u) = 1 - u^2$ and

$$\Psi(u) = \frac{\theta}{2}\left((1+u)\ln(1+u) + (1-u)\ln(1-u)\right) + \frac{1}{2}(1-u^2). \tag{1.3}$$

In a recent work by Cahn, Elliott, and Novick-Cohen [7], a formal asymptotic result for the deep quench limit ($\theta \searrow 0$) of the Cahn–Hilliard equation with $B(u) = 1 - u^2$

and Ψ as in (1.3) has been established (they used the scaling $\gamma = \varepsilon^2, t \to \varepsilon^2 t$). They show that one gets in the limit $\varepsilon \searrow 0$ the following geometric motion for hypersurfaces:

$$V = -D\Delta_S \kappa, \qquad D \in \mathbb{R}^+, \tag{1.4}$$

where V is the normal velocity, κ denotes the mean curvature, and Δ_S is the surface Laplacian. Material scientists refer to this evolution law as motion by surface diffusion (see Cahn and Taylor [10], Davi and Gurtin [12], and Mullins [21]). The two components of the alloy are separated by a sharp free boundary which is evolving according to the law (1.4).

Cahn and Taylor [10, 11] also propose the motions

$$V = \Delta_S \left(\frac{1}{M}\Delta_S - \frac{1}{D} \right)^{-1} \kappa \qquad (M, D \in \mathbb{R}^+), \tag{1.5}$$

which formally give in the limit as $M \to \infty$ ($D \to \infty$, respectively) the laws

$$V = -D\Delta_S \kappa \qquad \text{if} \quad M \to \infty \tag{1.6}$$

and

$$V = M(\kappa - \kappa_\alpha) \qquad \text{if} \quad D \to \infty, \tag{1.7}$$

where κ_α is the average mean curvature on the surface.

Formal asymptotic results suggest that the intermediate motion (1.5) is the asymptotic limit of the viscous Cahn–Hilliard equation with a mobility $B(u) = 1 - u^2$ (as before, with a logarithmic free energy and in the deep quench limit with a scaling $\gamma = \varepsilon^2$ and $t \to \varepsilon^2 t$).

For the geometric motions (1.5)–(1.7), just a few results exist so far. We can prove local existence for the two-dimensional case, i.e., for the evolution of curves in the plane and results for the global behaviour if the initial data are close to a circle (see [13]).

This paper is organized as follows: In §2, we prove the existence of a solution to the Cahn–Hilliard equation with a mobility which is bounded away from zero. This result is used in §3 to establish the existence of approximate solutions to the degenerate problem. We derive energy estimates for the approximate solutions which enable us to pass to the limit in the approximate equation to get the existence of a weak solution as stated in Theorem 1. Section 4 is devoted to other applications of our method. In particular, the viscous Cahn–Hilliard equation and the deep quench limit are studied. Furthermore, our method can be used to establish an existence result for degenerate parabolic equations of fourth order in arbitrary space dimensions. Finally, we discuss some open questions and give suggestions for further research.

2. Existence theorems for positive mobilities. In this section, we study the Cahn–Hilliard equation with a mobility which is bounded away from zero. We prove existence of solutions under various conditions on the bulk energy ψ. In §3, we will use these solutions as approximate solutions for the degenerate case.

We consider the Cahn–Hilliard equation in the form

$$u_t = \nabla \cdot b(u)\nabla w, \tag{2.1}$$

$$w = -\gamma\Delta u + \psi'(u) \tag{2.2}$$

with Neumann and no-flux boundary conditions

$$\nabla u \cdot \mathbf{n} = 0 \quad \text{and} \quad \nabla w \cdot \mathbf{n} = 0 \quad \text{on} \quad \partial\Omega \times (0,T).$$

Here $\Omega \subset \mathbb{R}^n$ ($n \in \mathbb{N}$) is a bounded domain with Lipschitz boundary. We assume b and ψ to be such that

(i) $b \in C(\mathbb{R}, \mathbb{R}^+)$ and there exist $b_1, B_1 > 0$ such that $b_1 \leq |b(u)| \leq B_1$ for all $u \in \mathbb{R}$,

(ii) $\psi \in C^1(\mathbb{R}, \mathbb{R})$ and there exist constants $C_1, C_2, C_3 > 0$ such that

$$|\psi'(u)| \leq C_1 |u|^q + C_2 \quad \text{and} \quad \psi(u) \geq -C_3,$$

where $q = \frac{n}{n-2}$ if $n > 3$ and $q \in \mathbb{R}^+$ arbitrary if $n = 1, 2$.

Under these assumptions we can state the following theorem.

THEOREM 2. *Suppose $u_0 \in H^1(\Omega)$. Then there exists a pair of functions (u, w) such that*

(1) $u \in L^\infty(0,T; H^1(\Omega)) \cap C([0,T]; L^2(\Omega))$,
(2) $u_t \in L^2(0,T; (H^1(\Omega))')$,
(3) $u(0) = u_0$,
(4) $w \in L^2(0,T; H^1(\Omega))$

which satisfies equations (2.1) *and* (2.2) *in the following weak sense:*

$$\int_0^T \langle \zeta(t), u_t(t) \rangle_{H^1,(H^1)'} = -\int_{\Omega_T} b(u) \nabla w \nabla \zeta \tag{2.3}$$

for all $\zeta \in L^2(0,T; H^1(\Omega))$ and

$$\int_\Omega w\phi = \gamma \int_\Omega \nabla u \nabla \phi + \int_\Omega \psi'(u)\phi \tag{2.4}$$

for all $\phi \in H^1(\Omega)$ and almost all $t \in [0,T]$.

Proof. To prove the theorem, we apply a Galerkin approximation. Let $\{\phi_i\}_{i \in \mathbb{N}}$ be the eigenfunctions of the Laplace operator with Neumann boundary conditions, i.e.,

$$-\Delta \phi_i = \lambda_i \phi_i \quad \text{in } \Omega \quad \text{and} \quad \nabla \phi_i \cdot \mathbf{n} = 0 \quad \text{on } \partial\Omega.$$

The eigenfunctions ϕ_i are orthogonal in the $H^1(\Omega)$ and the $L^2(\Omega)$ scalar product. We normalize the ϕ_i such that $(\phi_i, \phi_j)_{L^2(\Omega)} = \delta_{ij}$. Furthermore, we assume without loss of generality that $\lambda_1 = 0$.

Now we consider the following Galerkin ansatz for (2.1) and (2.2):

$$u^N(t,x) = \sum_{i=1}^N c_i^N(t)\phi_i(x)\,, \; w^N(t,x) = \sum_{i=1}^N d_i^N(t)\phi_i(x), \tag{2.5}$$

$$\int_\Omega \partial_t u^N \phi_j = -\int_\Omega b(u^N) \nabla w^N \nabla \phi_j \quad \text{for} \quad j = 1, \ldots, N, \tag{2.6}$$

$$\int_\Omega w^N \phi_j = \gamma \int_\Omega \nabla u^N \nabla \phi_j + \int_\Omega \psi'(u^N)\phi_j \quad \text{for} \quad j = 1, \ldots, N, \quad \text{and} \tag{2.7}$$

$$u^N(0) = \sum_{i=1}^N (u_0, \phi_i)_{L^2(\Omega)} \phi_i. \tag{2.8}$$

This gives an initial value problem for a system of ordinary differential equations for $(c_1, \ldots, c_N)$:

$$\partial_t c_j^N = -\sum_{k=1}^N d_k^N \int_\Omega b\left(\sum_{i=1}^N c_i^N \phi_i\right) \nabla\phi_k \nabla\phi_j, \tag{2.9}$$

$$d_j^N = \gamma\lambda_j c_j^N + \int_\Omega \psi'\left(\sum_{k=1}^N c_k^N \phi_k\right) \phi_j, \quad \text{and} \tag{2.10}$$

$$c_j^N(0) = (u_0, \phi_j)_{L^2(\Omega)}, \tag{2.11}$$

which has to hold for $j = 1, \ldots, N$. Since the right-hand side in (2.9) depends continuously on $c_1, \ldots, c_N$, the initial value problem has a local solution.

In order to derive a priori estimates, we differentiate the energy $\mathcal{E}$ and get

$$\begin{aligned} \frac{d}{dt}\mathcal{E}(t) &= \frac{d}{dt}\int_\Omega \left(\frac{\gamma}{2}|\nabla u^N|^2 + \psi(u^N)\right), \\ &= \int_\Omega \left(\gamma \nabla u^N \nabla u_t^N + \psi'(u^N) u_t^N\right), \\ &= \int_\Omega w^N u_t^N = -\int_\Omega b(u^N)|\nabla w^N|^2. \end{aligned}$$

This implies

$$\begin{aligned} &\int_\Omega \frac{\gamma}{2}|\nabla u^N(t)|^2 + \int_\Omega \psi(u^N(t)) + \int_{\Omega_T} b(u^N)|\nabla w^N|^2 \\ &= \int_\Omega \frac{\gamma}{2}|\nabla u^N(0)|^2 + \int_\Omega \psi(u^N(0)) \leq C. \end{aligned} \tag{2.12}$$

The last inequality follows from (2.8), assumption (ii), and the fact that $u_0 \in H^1(\Omega)$. Since $\frac{d}{dt}\int_\Omega u^N = 0$ (which follows from (2.6) with $j = 1$), Poincaré's inequality yields

$$\operatorname{ess\,sup}_{0<t<T} \|u(t)\|_{H^1(\Omega)} \leq C.$$

This estimate implies that the $(c_1^N, \ldots, c_N^N)$ are bounded and therefore a global solution to the initial value problem (2.9)–(2.11) exists.

If we denote by Π_N the projection of $L^2(\Omega)$ onto $span\{\phi_1, \ldots, \phi_N\}$, we get

$$\begin{aligned} \left|\int_{\Omega_T} \partial_t u^N \phi\right| &= \left|\int_{\Omega_T} \partial_t u^N \Pi_N \phi\right| \\ &= \left|\int_{\Omega_T} b(u^N)\nabla w^N \nabla \Pi_N \phi\right| \\ &\leq \left(\int_{\Omega_T} |b(u^N)\nabla w^N|^2\right)^{\frac{1}{2}} \left(\int_{\Omega_T} |\nabla \Pi_N \phi|^2\right)^{\frac{1}{2}} \\ &\leq B_1 \left(\int_{\Omega_T} b(u^N)|\nabla w^N|^2\right)^{\frac{1}{2}} \|\nabla\phi\|_{L^2(\Omega_T)} \\ &\leq C\|\nabla\phi\|_{L^2(\Omega_T)} \end{aligned}$$

for all $\phi \in L^2(0,T;H^1(\Omega))$. This implies

$$\|\partial_t u^N\|_{L^2(0,T;(H^1(\Omega))')} \leq C.$$

Using compactness results (see Lions [19] and Remark 1 below), we obtain for a subsequence (which we still denote by u^N)

$$\begin{array}{rclllll} u^N & \longrightarrow & u & \text{weak} - * & \text{in} & L^\infty(0,T;H^1(\Omega)), \\ u^N & \longrightarrow & u & \text{strongly} & \text{in} & C([0,T];L^2(\Omega)), \\ \partial_t u^N & \longrightarrow & \partial_t u & \text{weakly} & \text{in} & L^2(0,T;(H^1(\Omega))'), & \text{and} \\ u^N & \longrightarrow & u & \text{strongly} & \text{in} & L^2(0,T;L^p(\Omega)) \text{ and a.e. in } \Omega_T, \end{array}$$

where $p < \frac{2n}{n-2}$. It remains to show the convegence of w^N. Choosing $j = 1$ in (2.7) gives $\int_\Omega w^N(t) = \int_\Omega \psi'(u^N(t))$, which together with (2.12), assumption (ii), and Poincaré's inequality gives

$$\|w^N\|_{L^2(0,T;H^1(\Omega))} \leq C.$$

This implies (again for a subsequence)

$$w^N \longrightarrow w \quad \text{weakly} \quad \text{in} \quad L^2(0,T;H^1(\Omega)).$$

With the convergence properties proved so far and using the assumptions on b and ψ, we can pass to the limit in (2.6) and (2.7) in a standard fashion (see Lions [19] for details) to get that (2.3) and (2.4) hold for (u,w).

The strong convergence of u^N in $C([0,T];L^2(\Omega))$ and the fact that $u^N(0) \to u_0$ in $L^2(\Omega)$ gives $u(0) = u_0$. This proves the theorem. □

Remark 1. (a) Let X, Y and Z be Banach spaces with a compact embedding $X \hookrightarrow Y$ and a continuous embedding $Y \hookrightarrow Z$. Then the embeddings

$$\{u \in L^2(0,T;X) \mid \partial_t u \in L^2(0,T;Z)\} \hookrightarrow L^2(0,T;Y) \tag{2.13}$$

and

$$\{u \in L^\infty(0,T;X) \mid \partial_t u \in L^2(0,T;Z)\} \hookrightarrow C([0,T];Y) \tag{2.14}$$

are compact (for a proof, see Simon [22]).

(b) In the proof of Theorem 2, we applied the above result for the case $X = H^1(\Omega), Y = L^2(\Omega).(Y = L^p(\Omega)$ with $p < \frac{2n}{n-2}$, respectively), and $Z = (H^1(\Omega))'$.

(c) The solution in Theorem 2 lies in $C([0,T];H^\beta(\Omega))$ (where $\beta < 1$). We get this by choosing $X = H^1(\Omega), Y = H^\beta(\Omega)$, and $Z = (H^1(\Omega))'$ in (a). □

The existence result in Theorem 2 requires a bulk energy which is bounded from below. It is possible to generalize this result if we assume further assumptions on $\partial\Omega$ and the growth of ψ.

We assume now either $\partial\Omega \in C^{1,1}$ or Ω is convex. Furthermore, we replace assumption (ii) by the following:

(iii) $\psi \in C^2(\mathbb{R},\mathbb{R})$ and there exists a constant $D > 0$ such that $|\psi''(u)| \leq D$ for all $u \in \mathbb{R}$.

THEOREM 3. *Assume* (i), (iii), *and* $u_0 \in H^1(\Omega)$. *Then there exists a function* u *such that*

(1) $u \in L^2(0,T;H^1(\Omega)) \cap C([0,T];L^2(\Omega))$,
(2) $u_t \in L^2(0,T;(H^1(\Omega))')$,
(3) $u(0) = u_0$ *and* $\nabla u \cdot \mathbf{n} = 0$ *on* $\partial\Omega \times (0,T)$,
(4) $\nabla\Delta u \in L^2(\Omega_T)$

which satisfies the Cahn–Hilliard equation in the following sense:

$$\int_0^T \langle \zeta(t), u_t(t) \rangle_{H^1,(H^1)'} = \int_{\Omega_T} b(u)\nabla(-\gamma\Delta u + \psi'(u))\nabla\zeta$$

for all $\zeta \in L^2(0,T;H^1(\Omega))$.

Proof. As in the proof of Theorem 2, we apply a Galerkin approximation, but now we make an ansatz just in u:

$$u^N(t,x) = \sum_{i=1}^{N} c_i^N(t)\phi_i(x),$$

$$(2.15)\int_\Omega \partial_t u^N \phi_j = -\int_\Omega b(u_N)(-\gamma\nabla\Delta u^N + \psi''(u^N)\nabla u^N)\nabla\phi_j \text{ for } j=1,\dots,N,$$

$$u^N(0) = \sum_{i=1}^{N}(u_0,\phi_i)_{L^2(\Omega)}\phi_i.$$

Instead of differentiating the energy, we use Δu^N as a test function to get

$$\frac{1}{2}\partial_t\int_\Omega |\nabla u^N|^2 + \int_\Omega b(u^N)\gamma|\nabla\Delta u^N|^2 = \int_\Omega b(u^N)\psi''(u^N)\nabla u^N\nabla\Delta u^N.$$

Using Young's inequality and assumptions (i) and (iii), we derive

$$\partial_t\int_\Omega |\nabla u^N|^2 + b_1\gamma\int_\Omega |\nabla\Delta u^N|^2 \le C\int_\Omega |\nabla u^N|^2.$$

A Gronwall argument now gives

$$\int_\Omega |\nabla u^N(t)|^2 + \int_{\Omega_T} |\nabla\Delta u^N|^2 \le C(T).$$

With this estimate, the rest of the proof is straightforward using compactness results (see Lions [19] and Remark 1) and passing to the limit in equation (2.15). □

3. Existence proof for the degenerate case. In this section, we prove Theorem 1. Our approach is to approximate the degenerate problem by nondegenerate equations, i.e., by equations with a positive mobility. Furthermore, we modify the bulk energy Ψ so that it is defined on all $\mathbb{R}$.

We introduce a positive mobility B_ε as

$$B_\varepsilon(u) := \begin{cases} B(-1+\varepsilon) & \text{for} \quad u \le -1+\varepsilon, \\ B(u) & \text{for} \quad |u| < 1-\varepsilon, \\ B(1-\varepsilon) & \text{for} \quad u \ge 1-\varepsilon \end{cases}$$

and we define Φ_ε such that $\Phi_\varepsilon''(u) = \frac{1}{B_\varepsilon(u)}$ and $\Phi_\varepsilon'(0) = \Phi_\varepsilon(0) = 0$. We point out that $\Phi_\varepsilon(u) = \Phi(u)$ when $|u| \le 1-\varepsilon$.

The modified bulk energy $\Psi_\varepsilon : \mathbb{R} \longrightarrow \mathbb{R}$ is taken to be $\Psi_\varepsilon := \Psi_\varepsilon^1 + \Psi^2$, where

$$\left(\Psi_\varepsilon^1\right)''(u) := \begin{cases} \left(\Psi^1\right)''(-1+\varepsilon) & \text{for} \quad u \le -1+\varepsilon, \\ \left(\Psi^1\right)''(u) & \text{for} \quad |u| < 1-\varepsilon, \\ \left(\Psi^1\right)''(1-\varepsilon) & \text{for} \quad u \ge 1-\varepsilon \end{cases}$$

and $\Psi_\varepsilon^1(0) = \Psi^1(0)$, $\left(\Psi_\varepsilon^1\right)'(0) = \left(\Psi^1\right)'(0)$. As for Φ, we get $\Psi_\varepsilon(u) = \Psi(u)$ if $|u| \le 1-\varepsilon$. Furthermore, Ψ^2 is extended to be a function on all $\mathbb{R}$ such that $\|\Psi^2\|_{C^2(\mathbb{R})} \le C$.

With this choice of B_ε and Ψ_ε, Theorem 2 give the existence of a weak solution to the equation

$$\begin{aligned} u_t &= \nabla \cdot B_\varepsilon(u)\nabla w && \text{in } \Omega_T, \\ w &= -\gamma\Delta u + (\Psi_\varepsilon)'(u) && \text{in } \Omega_T, \\ \nabla u \cdot \mathbf{n} &= 0 \quad \text{and} \quad \nabla w \cdot \mathbf{n} = 0 && \text{on } \partial\Omega \times (0,T). \end{aligned}$$

We denote the solution by $(u_\varepsilon, w_\varepsilon)$. From now on, we assume either $\partial\Omega \in C^{1,1}$ or Ω is convex. With this assumption, we can state the following lemma.

LEMMA 1. *The solution u_ε belongs to the space $L^2(0,T;H^2(\Omega))$ and $\nabla\Delta u_\varepsilon \in L^2(\Omega_T)$.*

Proof. Since

$$\int_\Omega w_\varepsilon \phi = \int_\Omega \gamma \nabla u_\varepsilon \nabla \phi + \int_\Omega \Psi'_\varepsilon(u_\varepsilon)\phi$$

for all $\phi \in H^1(\Omega)$ and almost all $t \in (0,T)$, the first assertion follows from elliptic regularity theory. Because $\nabla w_\varepsilon \in L^2(\Omega_T)$ and $\nabla \Psi'_\varepsilon(u_\varepsilon) = \Psi''_\varepsilon(u_\varepsilon)\nabla u_\varepsilon \in L^2(\Omega_T)$, the identity $w_\varepsilon = -\gamma\Delta u_\varepsilon + \Psi'_\varepsilon(u_\varepsilon)$ gives $\nabla\Delta u_\varepsilon \in L^2(\Omega_T)$. □

Therefore, we get

$$(3.1) \qquad \int_0^T \langle \zeta, \partial_t u_\varepsilon \rangle_{H^1,(H^1)'} = -\int_{\Omega_T} B_\varepsilon(u_\varepsilon)\nabla(-\gamma\Delta u_\varepsilon + \Psi'_\varepsilon(u_\varepsilon))\nabla\zeta$$

for all $\zeta \in L^2(0,T;H^1(\Omega))$.

In the next step, we prove the following energy estimates.

LEMMA 2. *There exists an ε_0 such that for all $0 < \varepsilon \le \varepsilon_0$ the following estimates hold with a constant C independent of ε:*

$$\text{(a)} \qquad \operatorname{ess\,sup}_{0\le t\le T} \int_\Omega \left(\frac{\gamma}{2}|\nabla u_\varepsilon(t)|^2 + \Psi_\varepsilon(u_\varepsilon(t))\right) + \int_{\Omega_T} B_\varepsilon(u_\varepsilon)|\nabla w_\varepsilon|^2 \le C,$$

$$\text{(b)} \qquad \operatorname{ess\,sup}_{0\le t\le T} \int_\Omega \Phi_\varepsilon(u_\varepsilon(t)) + \int_{\Omega_T} \left(\gamma|\Delta u_\varepsilon|^2 + \left(\Psi^1_\varepsilon\right)''(u_\varepsilon)|\nabla u_\varepsilon|^2\right) \le C,$$

$$\text{(c)} \qquad \operatorname{ess\,sup}_{0\le t\le T} \int_\Omega (|u_\varepsilon| - 1)_+^2 \le C\varepsilon^m,$$

$$\text{(d)} \qquad \int_{\Omega_T} |\mathbf{J}_\varepsilon|^2 \le C, \quad \textit{where} \quad \mathbf{J}_\varepsilon := B_\varepsilon(u_\varepsilon)\nabla w_\varepsilon\,.$$

Proof. The function $w_\varepsilon = -\gamma\Delta u_\varepsilon + \Psi'_\varepsilon(u_\varepsilon) \in L^2(0,T;H^1(\Omega))$ is a valid test function in (3.1). Therefore, we obtain

$$(3.2) \qquad \int_0^t \langle -\gamma\Delta u_\varepsilon + \Psi'_\varepsilon(u_\varepsilon), \partial_t u_\varepsilon \rangle_{H^1,(H^1)'} = -\int_{\Omega_t} B_\varepsilon(u_\varepsilon)|\nabla w_\varepsilon|^2$$

for all $t \in [0,T]$.

CLAIM. *For almost all $t \in [0,T]$,*

$$(3.3) \qquad \begin{aligned} &\int_0^t \langle -\gamma\Delta u_\varepsilon + \Psi'_\varepsilon(u_\varepsilon), \partial_t u_\varepsilon \rangle_{H^1,(H^1)'} \\ &\quad = \int_\Omega |\nabla u_\varepsilon(t)|^2 + \int \Psi_\varepsilon(u_\varepsilon(t)) - \int_\Omega |\nabla u_0|^2 - \int_\Omega \Psi_\varepsilon(u_0) \end{aligned}$$

holds.

To prove this, we define functions

$$u_{\varepsilon h}(t,x) := \frac{1}{h}\int_{t-h}^{t} u_\varepsilon(\tau,x)d\tau, \tag{3.4}$$

where we set $u_\varepsilon(t,x) = u_0(x)$ when $t \leq 0$. It is easily proved that

$$\begin{aligned} \Delta u_{\varepsilon h} &\longrightarrow \Delta u_\varepsilon \quad \text{strongly in} \quad L^2(0,T;H^1(\Omega)) \text{ and}\\ \Psi'(u_{\varepsilon h}) &\longrightarrow \Psi'(u_\varepsilon) \quad \text{strongly in} \quad L^2(0,T;H^1(\Omega)) \end{aligned}$$

for at least a subsequence (as $h \searrow 0$). Furthermore, we can show $\partial_t u_{\varepsilon h} \longrightarrow \partial_t u_\varepsilon$ strongly in $L^2(0,T;(H^1(\Omega))')$. For any $\zeta \in L^2(0,T;H^1(\Omega))$ we have

$$\begin{aligned} |\langle \zeta, \partial_t u_{\varepsilon h} - \partial_t u_\varepsilon \rangle_{L^2(H^1),L^2((H^1)')}| &= \frac{1}{h}\left| \int_0^T \left\langle \zeta, \int_{t-h}^{t} (\partial_t u_\varepsilon(\tau) - \partial_t u_\varepsilon(t))d\tau \right\rangle_{H^1,(H^1)'} dt \right|\\ &= \frac{1}{h}\left| \int_0^T \left\langle \zeta, \int_{-h}^{0} (\partial_t u_\varepsilon(t+s) - \partial_t u_\varepsilon(t)ds \right\rangle dt \right|\\ &\leq \frac{1}{h}\int_{-h}^{0} \left| \int_0^T \int_\Omega \nabla\zeta \cdot (\mathbf{J}_\varepsilon(t+s) - \mathbf{J}_\varepsilon(t))dx\,dt \right| ds\\ &\leq \|\nabla\zeta\|_{L^2(\Omega_T)} \sup_{-h\leq s\leq 0} \|\mathbf{J}_\varepsilon(.+s) - \mathbf{J}_\varepsilon(.)\|_{L^2(\Omega_T)}. \end{aligned}$$

Since

$$\sup_{-h\leq s\leq 0} \|\mathbf{J}_\varepsilon(.+s) - \mathbf{J}_\varepsilon(.)\|_{L^2(\Omega_T)} \longrightarrow 0 \quad \text{as} \quad h \longrightarrow 0,$$

it follows that

$$\partial_t u_{\varepsilon h} \longrightarrow \partial_t u_\varepsilon \quad \text{strongly in} \quad L^2(0,T;(H^1(\Omega))').$$

Using $\partial_t u_{\varepsilon h} \in L^2(\Omega_T)$, we have for almost all $t \in [0,T]$

$$\begin{aligned} &\int_0^t \langle -\gamma\Delta u_{\varepsilon h} + \Psi_\varepsilon'(u_{\varepsilon h}), \partial_t u_{\varepsilon h} \rangle_{H^1,(H^1)'}\\ &= \int_{\Omega_t} \left(-\gamma\Delta u_{\varepsilon h} + \Psi_\varepsilon'(u_{\varepsilon h})\right) \partial_t u_{\varepsilon h}\\ &= \partial_t \int_0^t \int_\Omega \left(\frac{\gamma}{2}|\nabla u_{\varepsilon h}|^2 + \Psi(u_{\varepsilon h})\right)\\ &= \int_\Omega \left(\frac{\gamma}{2}|\nabla u_{\varepsilon h}(t)|^2 + \Psi(u_{\varepsilon h}(t))\right) - \int_\Omega \left(\frac{\gamma}{2}|\nabla u_0|^2 + \Psi_\varepsilon(u_0)\right). \end{aligned}$$

Passing to the limit ($h \searrow 0$) in this equation, where we apply the convergence properties of $u_{\varepsilon h}$ proved above, and using (3.2) gives for almost all t

$$\int_\Omega \left(\frac{\gamma}{2}|\nabla u_\varepsilon(t)|^2 + \Psi_\varepsilon(u_\varepsilon(t))\right) + \int_{\Omega_t} B_\varepsilon(u_\varepsilon)|\nabla w_\varepsilon|^2 = \int_\Omega \left(\frac{\gamma}{2}|\nabla u_0|^2 + \Psi_\varepsilon(u_0)\right).$$

Noting that $\Psi_\varepsilon(u) \leq \Psi(u)$ for ε sufficiently small proves (a).

To prove (b), we want to use $\Phi'_\varepsilon(u_\varepsilon)$ as a test function in (3.1). Since Φ''_ε is bounded, we have $\Phi'_\varepsilon(u_\varepsilon) \in L^2(0,T;H^1(\Omega))$ and therefore is an admissible test function. With a similar argument as in the proof of (a), we can prove that

$$\int_0^t \langle \Phi'_\varepsilon(u_\varepsilon), \partial_t u_\varepsilon \rangle_{H^1,(H^1)'} = \int_\Omega \Phi_\varepsilon(u_\varepsilon(t)) - \int_\Omega \Phi_\varepsilon(u_0)$$

is true for almost all $t \in [0,T]$. On the other hand, we derive

$$\begin{aligned}
&\int_{\Omega_t} B_\varepsilon(u_\varepsilon)\nabla(-\gamma\Delta u_\varepsilon + \Psi'_\varepsilon(u_\varepsilon))\nabla\Phi'_\varepsilon(u_\varepsilon) \\
&= \int_{\Omega_t} (-\gamma\nabla\Delta u_\varepsilon + \Psi''_\varepsilon(u_\varepsilon)\nabla u_\varepsilon)\, B_\varepsilon(u_\varepsilon)\Phi''_\varepsilon(u_\varepsilon)\nabla u_\varepsilon \\
&= \int_{\Omega_t} (\gamma|\Delta u_\varepsilon|^2 + \Psi''_\varepsilon(u_\varepsilon)|\nabla u_\varepsilon|^2)\,.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\int_\Omega \Phi_\varepsilon(u_\varepsilon(t)) + \int_{\Omega_t} \gamma|\Delta u_\varepsilon|^2 + \int_{\Omega_t} \left(\Psi^1_\varepsilon\right)''(u_\varepsilon)|\nabla u_\varepsilon|^2 \\
&\le \int_\Omega \Phi_\varepsilon(u_0) + \int_{\Omega_t} \left|\left(\Psi^2\right)''(u_\varepsilon)\right| |\nabla u_\varepsilon|^2.
\end{aligned}$$

Since $\Phi_\varepsilon(u) \le \Phi(u)$ for ε sufficiently small and $\left(\Psi^2\right)''$ is bounded, we have proved (b) (note that we have estimated $\int_\Omega |\nabla u_\varepsilon|^2$ in (a)).

Now we can use the bound for $\int_\Omega \Phi_\varepsilon(u_\varepsilon)$ to derive a bound for $\int_\Omega (|u_\varepsilon| - 1)_+^2$. If $z > 1$ and $\varepsilon < 1$, then we have

$$\begin{aligned}
\Phi_\varepsilon(z) &= \underbrace{\Phi(1-\varepsilon)}_{\ge 0} + \underbrace{\Phi'(1-\varepsilon)}_{\ge 0}\underbrace{(z-(1-\varepsilon))}_{\ge 0} + \frac{1}{2}\Phi''(1-\varepsilon)\,(z-(1-\varepsilon))^2 \\
&\ge \frac{1}{2}\Phi''(1-\varepsilon)(z-1)^2 = \frac{1}{2}\frac{1}{B(1-\varepsilon)}(z-1)^2 \\
&= \frac{1}{2}\frac{1}{(1-(1-\varepsilon)^2)^m\,\bar{B}(1-\varepsilon)}(z-1)^2 \ge C^{-1}\varepsilon^{-m}(z-1)^2.
\end{aligned}$$

It follows now that $(z-1)^2 \le C\varepsilon^m\Phi_\varepsilon(z)$. Similarly, we obtain $(|z|-1)^2 \le C\varepsilon^m\Phi_\varepsilon(z)$ for $z < -1$. This implies

$$\int_\Omega (|u_\varepsilon| - 1)_+^2 \le C\varepsilon^m \int_\Omega \Phi_\varepsilon(u_\varepsilon) \le C\varepsilon^m,$$

which proves (c).

Assertion (d) follows easily from (a), and this finishes the proof of Lemma 2. □

Since Δu_ε is uniformly bounded in $L^2(\Omega_T)$, $\nabla u_\varepsilon \cdot \mathbf{n} = 0$, and $\int_\Omega u_\varepsilon = \int_\Omega u_0$, elliptic regularity theory yields

$$\|u_\varepsilon\|_{L^2(0,T;H^2(\Omega))} \le C.$$

Now we apply the compactness result mentioned in Remark 1 (2.13) with $X = H^2(\Omega)$, $Y = H^1(\Omega)$, and $Z = (H^1(\Omega))'$ to conclude the existence of a subsequence of $(u_\varepsilon)_{\varepsilon>0}$ (which we still denote by $(u_\varepsilon)_{\varepsilon>0}$) such that

$$u_\varepsilon, \nabla u_\varepsilon \longrightarrow u, \nabla u \quad \text{strongly in} \quad L^2(\Omega_T) \quad \text{and a.e. in } \Omega_T.$$

Furthermore, using standard compactness properties, we obtain the convergence

$$\begin{aligned} \partial_t u_\varepsilon &\longrightarrow \partial_t u \quad \text{weakly in} \quad L^2(0,T;(H^1(\Omega))'), \\ \Delta u_\varepsilon &\longrightarrow \Delta u \quad \text{weakly in} \quad L^2(\Omega_T), \text{ and} \\ \mathbf{J}_\varepsilon &\longrightarrow \mathbf{J} \quad\ \ \text{weakly in} \quad L^2(\Omega_T). \end{aligned}$$

Passing to the limit in

$$\int_\Omega (|u_\varepsilon| - 1)_+^2 \le C\varepsilon^m$$

yields $|u| \le 1$ a.e. in Ω_T.

It remains to show that u fulfills the limit equation. The weak convergence of $\partial_t u_\varepsilon$ and $\mathbf{J}_\varepsilon$ gives in the limit

$$\int_0^T \langle \zeta, \partial_t u \rangle_{H^1,(H^1)'} = \int_{\Omega_T} \mathbf{J} \cdot \nabla\zeta$$

for all $\zeta \in L^2(0,T;H^1(\Omega))$. Now we have to identify $\mathbf{J}$. Therefore, we want to pass to the limit in the equation

$$\int_{\Omega_T} \mathbf{J}_\varepsilon \cdot \boldsymbol{\eta} = \int_{\Omega_T} B_\varepsilon(u_\varepsilon)\nabla\left(-\gamma\Delta u_\varepsilon + \Psi'_\varepsilon(u_\varepsilon)\right)\boldsymbol{\eta}, \tag{3.5}$$

where $\boldsymbol{\eta} \in L^2(0,T;H^1(\Omega,\mathbb{R}^n)) \cap L^\infty(\Omega_T,\mathbb{R}^n)$ with $\boldsymbol{\eta}\cdot\mathbf{n} = 0$ on $\partial\Omega \times (0,T)$. The left-hand side converges to $\int_{\Omega_T} \mathbf{J}\cdot\boldsymbol{\eta}$. Since $\nabla\Delta u_\varepsilon$ may not have a limit in $L^2(\Omega_T)$, we integrate the first term on the right-hand side of (3.5) by parts to get

$$\int_{\Omega_T} B_\varepsilon(u_\varepsilon)\nabla(-\gamma\Delta u_\varepsilon)\boldsymbol{\eta} = \int_{\Omega_T} \gamma\Delta u_\varepsilon B_\varepsilon(u_\varepsilon)\nabla\cdot\boldsymbol{\eta} + \int \gamma\Delta u_\varepsilon B'_\varepsilon(u_\varepsilon)\nabla u_\varepsilon\cdot\boldsymbol{\eta} =: \mathrm{I} + \mathrm{II}.$$

Using the fact that for all $z \in \mathbb{R}$

$$|B_\varepsilon(z) - B(z)| \le \sup_{1-\varepsilon\le|y|\le1} |B(y)| \longrightarrow 0 \quad \text{as} \quad \varepsilon \longrightarrow 0,$$

it follows that $B_\varepsilon \longrightarrow B$ uniformly.

Hence we have

$$B_\varepsilon(u_\varepsilon) \longrightarrow B(u) \quad \text{a.e. in } \Omega_T.$$

Since $\Delta u_\varepsilon \longrightarrow \Delta u$ weakly in $L^2(\Omega_T)$ and B_ε is uniformly bounded, we conclude

$$\int_{\Omega_T} \gamma\Delta u_\varepsilon B_\varepsilon(u_\varepsilon)\nabla\cdot\boldsymbol{\eta} \longrightarrow \int_{\Omega_T} \gamma\Delta u B(u)\nabla\cdot\boldsymbol{\eta} \quad \text{as} \quad \varepsilon \longrightarrow 0.$$

Now we pass to the limit in II. First of all, we consider the case $m > 1$. As for B, we have $B'_\varepsilon \longrightarrow B'$ uniformly, which gives

$$B'_\varepsilon(u_\varepsilon) \longrightarrow B'(u) \quad \text{a.e. in } \Omega_T.$$

By using

$$\nabla u_\varepsilon \longrightarrow \nabla u \quad \text{in} \quad L^2(\Omega_T) \quad \text{and a.e. in } \Omega_T,$$

and the fact that B'_ε is uniformly bounded a generalized version of the Lebesgue convergence theorem yields

$$B'_\varepsilon(u_\varepsilon)\nabla u_\varepsilon \longrightarrow B'(u)\nabla u \quad \text{in} \quad L^2(\Omega_T).$$

Hence

$$\int_{\Omega_T} \gamma \Delta u_\varepsilon B'_\varepsilon(u_\varepsilon)\nabla u_\varepsilon \cdot \boldsymbol{\eta} \longrightarrow \int_{\Omega_T} \gamma \Delta u B'(u)\nabla u \cdot \boldsymbol{\eta},$$

where we use the fact that $\boldsymbol{\eta} \in L^\infty(\Omega_T)$.

In the case $m = 1$, the function B' is discontinuous and we have to use a more subtle argument.

Claim. $B'_\varepsilon(u_\varepsilon)\nabla u_\varepsilon \longrightarrow B'(u)\nabla u$ *in* $L^2(\Omega_T)$.

We analyze the following integrals:

$$\begin{aligned}&\int_{\Omega_T} |B'_\varepsilon(u_\varepsilon)\nabla u_\varepsilon - B'(u)\nabla u|^2 \\ &\quad = \int_{\Omega_T \cap \{|u|<1\}} |B'_\varepsilon(u_\varepsilon)\nabla u_\varepsilon - B'(u)\nabla u|^2 + \int_{\Omega_T \cap \{|u|=1\}} |B'_\varepsilon(u_\varepsilon)\nabla u_\varepsilon - B'(u)\nabla u|^2.\end{aligned}$$

Since $\nabla u = 0$ on the set $\{|u| = 1\}$ (see [16, Lem. 7.7]), we obtain

$$\begin{aligned}\int_{\Omega_T \cap \{|u|=1\}} |B'_\varepsilon(u_\varepsilon)\nabla u_\varepsilon - B'(u)\nabla u|^2 &= \int_{\Omega_T \cap \{|u|=1\}} |B'_\varepsilon(u_\varepsilon)\nabla u_\varepsilon|^2 \\ &\leq C \int_{\Omega_T \cap \{|u|=1\}} |\nabla u_\varepsilon|^2 \longrightarrow C \int_{\Omega_T \cap \{|u|=1\}} |\nabla u|^2 = 0.\end{aligned}$$

On the set $\{|u| < 1\}$, we know $B'_\varepsilon(u_\varepsilon) \longrightarrow B'(u)$ a.e.

Hence we have

$$B'_\varepsilon(u_\varepsilon)\nabla u_\varepsilon \longrightarrow B'(u)\nabla u \quad \text{a.e. in } \Omega_T.$$

The generalized Lebesgue convergence theorem now gives

$$\int_{\Omega_T \cap \{|u|<1\}} |B'_\varepsilon(u_\varepsilon)\nabla u_\varepsilon - B'(u)\nabla u|^2 \longrightarrow 0,$$

which proves our claim. Furthermore, this proves that we can pass to the limit in II.

To complete the proof of Theorem 1, we have to show

$$\int_{\Omega_T} B_\varepsilon(u_\varepsilon)\Psi''_\varepsilon(u_\varepsilon)\nabla u_\varepsilon \cdot \boldsymbol{\eta} \longrightarrow \int_{\Omega_T} (B\Psi'')(u)\nabla u \cdot \boldsymbol{\eta}. \tag{3.6}$$

First of all, we want to point out that $B_\varepsilon \Psi''_\varepsilon$ is uniformly bounded. Therefore, it is sufficient to show

$$B_\varepsilon(u_\varepsilon)\Psi''_\varepsilon(u_\varepsilon) \longrightarrow (B\Psi'')(u) = \bar{B}(u)F(u) + B(u)\Psi''_2(u) \quad \text{a.e. in } \Omega_T. \tag{3.7}$$

If $|u(t,x))| < 1$, the convergence in (3.7) follows from the definition of B_ε and Ψ_ε ($B_\varepsilon(z) = B(z)$ and $\Psi_\varepsilon(z) = \Psi(z)$ if $|z| < 1 - \varepsilon$). Now let us consider points (t,x), where $|u(t,x)| = 1$. Without loss of generality, we assume $u_\varepsilon(t,x) \longrightarrow 1 = u(t,x)$.

For ε with $u_\varepsilon(t,x) \geq 1-\varepsilon$, we have

$$\begin{aligned} B_\varepsilon(u_\varepsilon(t,x))\Psi''_\varepsilon(u_\varepsilon(t,x)) &= \bar{B}(1-\varepsilon)F(1-\varepsilon) + B(1-\varepsilon)\Psi''_2(u_\varepsilon(t,x)) \\ &\longrightarrow \bar{B}(1)F(1) + B(1)\Psi''_2(1) = (B\Psi'')(u(t,x)). \end{aligned}$$

On the other side, if $u_\varepsilon(t,x) \leq 1-\varepsilon$ and $u_\varepsilon(t,x) \longrightarrow 1$, we have

$$\begin{aligned} B_\varepsilon(u_\varepsilon(t,x))\Psi''_\varepsilon(u_\varepsilon(t,x)) &= B(u_\varepsilon(t,x))\Psi''(u_\varepsilon(t,x)) \\ &= \bar{B}(u_\varepsilon(t,x))F(u_\varepsilon(t,x)) + B(u_\varepsilon(t,x))\Psi''_2(u_\varepsilon(t,x)) \\ &\longrightarrow (B\Psi'')(u(t,x)). \end{aligned}$$

We proved $B_\varepsilon(u_\varepsilon)\Psi''_\varepsilon(u_\varepsilon) \longrightarrow (B\Psi'')(u)$ a.e. in Ω_T, which together with the strong convergence of ∇u_ε in $L^2(\Omega_T)$ gives (3.6). This shows that u solves the Cahn–Hilliard equation in the sense of Theorem 1. The facts that $u \in C([0,T]; L^2(\Omega))$ and $u_\varepsilon(0) = u_0$ follow as in the proof of Theorem 2 from an application of the compactness result mentioned in Remark 1. In fact, it holds that $u \in C([0,T]; H^\beta(\Omega))$ with $\beta < 1$. □

Remark 2 (generalized Lebesgue convergence theorem). Assume $E \subset \mathbb{R}^n$ is measurable, $g_n \longrightarrow g$ in $L^q(E)$ with $1 \leq q < \infty$ and $f_n, f : E \longrightarrow \mathbb{R}^n$ are measurable functions such that

$$\begin{aligned} f_n &\longrightarrow f \quad \text{a.e. in } E, \\ |f_n|^p &\leq |g_n|^q \quad \text{a.e. in } E \end{aligned}$$

with $1 \leq p < \infty$. Then $f_n \longrightarrow f$ in $L^p(E)$.

For a proof see [1, A 1.26]. □

Remark 3. For $m \in [1,2)$, the functions Φ and Ψ are bounded on the interval $[-1,1]$ and therefore the assumption

$$\int_\Omega (\Phi(u_0) + \Psi(u_0)) \leq C$$

imposes no restrictions on the initial data. This is in particular true for the case $B(u) = 1-u^2$ and Ψ of the logarithmic form (1.2). □

The following corollary gives an additional result in the case $m \geq 2$.

COROLLARY. *Assume $m \geq 2$ and u is the solution constructed in Theorem* 1. *Then*

(a) $$\operatorname{ess\,sup}_{0\leq t\leq T} \int_\Omega (\Phi(u(t)) + \Psi(u(t))) \leq C,$$

(b) *the set* $\{x \mid |u(t,x)| = 1\}$ *has zero measure for almost all* $t \in [0,T]$.

Proof. We have proved

$$\int_\Omega \Phi_\varepsilon(u_\varepsilon(t)) \leq C$$

for almost all $t \in [0,T]$. Since $\Phi_\varepsilon(u_\varepsilon) \geq 0$, the Lemma of Fatou gives

$$\int_\Omega \liminf_{\varepsilon\searrow 0} \Phi_\varepsilon(u_\varepsilon(t)) \leq \liminf_{\varepsilon\searrow 0} \int_\Omega \Phi_\varepsilon(u_\varepsilon(t)) \leq C.$$

CLAIM.

$$\liminf_{\varepsilon\searrow 0} \Phi_\varepsilon(u_\varepsilon) = \begin{cases} \Phi(u) & \textit{if } |u| < 1, \\ \infty & \textit{elsewhere.} \end{cases}$$

If $|u| < 1$, it is clear that $\lim_{\varepsilon\searrow 0}\Phi_\varepsilon(u_\varepsilon) = \Phi(u)$. Now we consider points (t,x), where $\lim_{\varepsilon\searrow 0} u_\varepsilon(t,x) = 1$. In this case, we have

$$\Phi_\varepsilon(u_\varepsilon(t,x)) \geq \min\left(\Phi(1-\varepsilon), \Phi(u_\varepsilon(t,x))\right) \longrightarrow \infty$$

as $\varepsilon \longrightarrow 0$. The same argument can be applied for $\lim_{\varepsilon\searrow 0} u_\varepsilon(t,x) = -1$, which proves the claim. Therefore, the set $\{x \mid |u(t,x)| = 1\}$ has zero measure and

$$\liminf_{\varepsilon\searrow 0} \Phi_\varepsilon(u_\varepsilon) = \lim_{\varepsilon\searrow 0}\Phi_\varepsilon(u_\varepsilon) = \Phi(u) \quad \text{a.e. in } \Omega_T.$$

The estimate $\int_\Omega \Psi(u(t)) \leq C$ is proved similarly. □

Remark 4. Since F can vanish at ±1, Ψ'' can be less singular than of order m. In particular, the smooth double-well potential $\Psi(u) = (1-u^2)^2$ and the double-obstacle potential are possible choices for all $m \geq 1$.

4. Some generalizations.

4.1. The viscous Cahn–Hilliard equation. In this subsection, we consider the viscous Cahn–Hilliard equation with a nonconstant mobility

$$\begin{aligned} u_t &= -\nabla\cdot\mathbf{J},\\ \mathbf{J} &= -B(u)\nabla w,\\ w &= -\gamma\Delta u + \Psi'(u) + \alpha u_t, \qquad \alpha\in\mathbb{R}^+, \end{aligned}$$

supplemented with the boundary conditions $\mathbf{J}\cdot\mathbf{n} = 0$ and $\nabla u\cdot\mathbf{n} = 0$ on $\partial\Omega\times(0,T)$. For a mobility $B \equiv 1$, this is the usual viscous Cahn–Hilliard equation as studied by Novick-Cohen, Elliott, Stuart, and others [2, 15, 20].

In a first step, we state a theorem for the nondegenerate case. Therefore, we assume b and ψ to fulfill assumptions (i) and (ii) in §2.

THEOREM 4. *Suppose $u_0 \in H^1(\Omega)$ and $\partial\Omega$ Lipschitz. Then there exists a pair (u,w) such that*

(1) $u \in L^\infty(0,T;H^1(\Omega)) \cap C([0,T];L^2(\Omega))$,
(2) $u_t \in L^2(\Omega_T)$,
(3) $u(0) = u_0$,
(4) $w \in L^2(0,T;H^1(\Omega))$

which satisfies

$$\int_{\Omega_T} \zeta u_t = -\int_{\Omega_T} b(u)\nabla w\nabla\zeta$$

for all $\zeta \in L^2(0,T;H^1(\Omega))$ and

$$\int_\Omega w\phi = \gamma\int_\Omega \nabla u\nabla\phi + \int_\Omega \psi'(u)\phi + \int_\Omega \alpha u_t\phi$$

for all $\phi \in H^1(\Omega)$ and almost all $t \in [0,T]$.

Proof. As in the proof of Theorem 2, we apply a Galerkin approximation

$$u^N(t,x) = \sum_{i=1}^N c_i^N(t)\phi_i(x)\ ,\ w^N(t,x) = \sum_{i=1}^N d_i^N(t)\phi_i(x),$$

$$\int_\Omega \partial_t u^N\phi_j = -\int_\Omega b(u_N)\nabla w^N\nabla\phi_j \quad \text{for} \quad j = 1,\ldots,N,$$

$$\int_\Omega w^N\phi_j = \gamma\int_\Omega \nabla u^N\nabla\phi_j + \int_\Omega \psi'(u^N)\phi_j + \alpha\int_\Omega \partial_t u^N\phi_j \quad \text{for} \quad j=1,\dots,N,$$
$$u^N(0) = \sum_{i=1}^N (u_0,\phi_i)_{L^2(\Omega)}\phi_i,$$

which gives

$$\partial_t c_j^N = -\sum_{k=1}^N d_k^N \int_\Omega b\left(\sum_{i=1}^N c_i^N\phi_i\right)\nabla\phi_k\nabla\phi_j, \tag{4.1}$$

$$d_j^N = \gamma\lambda_j c_j^N + \int_\Omega \psi'\left(\sum_{l=1}^N c_l^N\phi_l\right)\phi_j + \alpha\partial_t c_j, \quad \text{and} \tag{4.2}$$

$$c_j^N(0) = (u_0,\phi_j)_{L^2(\Omega)}. \tag{4.3}$$

These equations have to hold for $j=1,\dots,N$. This yields to the following initial value problem for $(c_1^N,\dots,c_N^N)$:

$$\begin{aligned}\partial_t c_j^N &+ \alpha\sum_{k=1}^N \partial_t c_k\int_\Omega b\left(\sum_{i=1}^N c_i^N\phi_i\right)\nabla\phi_k\nabla\phi_j \\ &= -\sum_{k=1}^N\left(\gamma\lambda_k c_k^N + \int_\Omega\psi'\left(\sum_{l=1}^N c_l^N\phi_l\right)\phi_k\right)\int_\Omega b\left(\sum_{i=1}^N c_i^N\phi_i\right)\nabla\phi_k\nabla\phi_j,\end{aligned} \tag{4.4}$$

$$c_j^N(0) = (u_0,\phi_j)_{L^2(\Omega)}. \tag{4.5}$$

Since the matrix $(g_{jk})_{j,k=1,N}$ with

$$g_{jk} = \int_\Omega b\left(\sum_i c_i^N\phi_i\right)\nabla\phi_k\nabla\phi_j$$

is positive definite, the initial value problem (4.4), (4.5) has a local solution.

Now we use

$$\begin{aligned}\frac{d}{dt}\int_{\Omega_T}\left(|\nabla u^N|^2 + \psi(u^N)\right) &= \int_{\Omega_T}\left(-\gamma\Delta u^N + \psi'(u^N)\right)u_t^N \\ &= \int_\Omega w^N u_t^N - \int_\Omega \alpha\left(u_t^N\right)^2 \\ &= -\int_\Omega b(u^N)|\nabla w^N|^2 - \int_{\Omega_T}\alpha\left(u_t^N\right)^2\end{aligned}$$

to establish a priori estimates. The rest is proved in a similar way as in the proof of Theorem 2. □

Having proved an existence theorem for a positive mobility, we are now in a position to prove existence for the degenerate case. We assume Ψ and B to be as in the introduction. Furthermore, we assume either $\partial\Omega \in C^{1,1}$ or Ω convex.

THEOREM 5. *Let $u_0 \in H^1(\Omega)$ satisfy $|u_0| \le 1$ a.e. in Ω_T and*

$$\int_\Omega (\Psi(u_0) + \Phi(u_0)) \le C, \quad C \in \mathbb{R}^+.$$

Then there exists a pair $(u, \mathbf{J})$ *such that*

(a) $u \in L^2(0,T;H^2(\Omega)) \cap L^\infty(0,T;H^1(\Omega)) \cap C([0,T];L^2(\Omega))$,
(b) $u_t \in L^2(\Omega_T)$,
(c) $u(0) = u_0$ *and* $\nabla u \cdot \mathbf{n} = 0$ *on* $\partial\Omega \times (0,T)$,
(d) $|u| \le 1$ *a.e in* $\Omega_T := \Omega \times (0,T)$,
(e) $\mathbf{J} \in L^2(\Omega_T, \mathbb{R}^n)$

which satisfies $u_t = -\nabla \cdot \mathbf{J}$ *in* $L^2(0,T;(H^1(\Omega))')$, *i.e.,*

$$\int_{\Omega_T} \zeta u_t = \int_{\Omega_T} \mathbf{J} \cdot \nabla \zeta$$

for all $\zeta \in L^2(0,T;H^1(\Omega))$ *and*

$$\mathbf{J} = -B(u)\nabla\left(-\gamma\Delta u + \Psi'(u) + \alpha u_t\right)$$

in the following weak sense:

$$\int_{\Omega_T} \mathbf{J} \cdot \boldsymbol{\eta} = -\int_{\Omega_T} \left[(\gamma\Delta u - \alpha u_t)\,\nabla \cdot (B(u)\boldsymbol{\eta}) + (B\Psi'')(u)\nabla u \cdot \boldsymbol{\eta}\right]$$

for all $\boldsymbol{\eta} \in L^2(0,T;H^1(\Omega,\mathbb{R}^n)) \cap L^\infty(\Omega_T,\mathbb{R}^n)$ *which fulfill* $\boldsymbol{\eta} \cdot \mathbf{n} = 0$ *on* $\partial\Omega \times (0,T)$.

Proof. We modify B and Ψ in the same manner as in the proof of Theorem 1 to get functions B_ε and Ψ_ε. For the modified equation, we proved existence in Theorem 4. In a similar fashion as in the proof of Theorem 1, we can derive the following identities for the approximating solutions $(u_\varepsilon, w_\varepsilon)$:

$$\operatorname{ess\,sup}_{0\le t\le T} \int_\Omega \left(\frac{\gamma}{2}|\nabla u_\varepsilon(t)|^2 + \Psi(u_\varepsilon(t))\right) + \int_{\Omega_T} B_\varepsilon(u_\varepsilon)|\nabla w_\varepsilon|^2 + \\ + \alpha\int_{\Omega_T} (\partial_t u_\varepsilon)^2 = \int_\Omega \left(\frac{\gamma}{2}|\nabla u_0|^2 + \Psi_\varepsilon(u_0)\right)$$

and

$$\operatorname{ess\,sup}_{0\le t\le T} \int_\Omega \left(\Phi_\varepsilon(u_\varepsilon(t)) + \frac{\alpha}{2}|\nabla u_\varepsilon(t)|^2\right) + \int_{\Omega_T} \gamma|\Delta u_\varepsilon|^2 + \\ \int_{\Omega_T} \psi_\varepsilon''(u_\varepsilon)|\nabla u_\varepsilon|^2 = \int_\Omega \left(\Phi_\varepsilon(u_0) + \frac{\alpha}{2}|\nabla u_0|^2\right).$$

With these estimates, the remaining part of the proof follows the outline of the proof of Theorem 1. One uses compactness results to conclude the existence of a converging subsequence and passes to the limit in the approximating equation. □

4.2. The deep quench limit. Now we consider the case $B(u) = 1 - u^2$ and

$$\Psi_\theta(u) = \frac{\theta}{2}\left((1+u)\ln(1+u) + (1-u)\ln(1-u)\right) + \frac{1}{2}(1-u^2), \tag{4.6}$$

where $\theta > 0$. Let us denote the solution we constructed in the proof of Theorem 1 by u_θ. Cahn, Elliott, and Novick-Cohen [7] studied the deep quench limit ($\theta \searrow 0$) of these solutions. The purpose of this subsection is to show that the solutions u_θ converge to weak solutions of the Cahn–Hilliard equation with a mobility $B(u) = 1 - u^2$ and a bulk energy $\Psi(u) = 1 - u^2$, which is the case where we set $\theta = 0$ in (4.6).

For u_θ, we have the following a priori estimates, which follow from the estimates derived in Lemma 2 and the weak lower semicontinuity of the L^2-norm:

$$\operatorname{ess\,sup}_{0\le t\le T}\int_\Omega |\nabla u_\theta(t)|^2 + \int_{\Omega_T} |\mathbf{J}_\theta|^2 \le \int_\Omega \left(\frac{\gamma}{2}|\nabla u_0|^2 + \Psi_\theta(u_0)\right) \le C$$

and

$$\int_{\Omega_T} \gamma|\Delta u_\theta|^2 \le \int_\Omega \Phi(u_0) + \int_{\Omega_T} |\nabla u_\theta|^2 \le C$$

with a constant C independently of θ. From these estimates, we obtain

$$\|\partial_t u_\theta\|_{L^2(0,T;(H^1(\Omega))')} + \|u_\theta\|_{L^2(0,T;H^2(\Omega))} \le C.$$

Using the same compactness results as before, we get (for a subsequence $\theta \searrow 0$)

$$\begin{array}{rcllll} u_\theta, \nabla u_\theta & \longrightarrow & u, \nabla u & \text{strongly} & \text{in} & L^2(\Omega_T) \text{ and a.e. in } \Omega_T, \\ \Delta u_\theta & \longrightarrow & \Delta u & \text{weakly} & \text{in} & L^2(\Omega_T), \\ \partial_t u_\theta & \longrightarrow & \partial_t u & \text{weakly} & \text{in} & L^2(0,T;(H^1(\Omega))'), \quad \text{and} \\ \mathbf{J}_\theta & \longrightarrow & \mathbf{J} & \text{weakly} & \text{in} & L^2(\Omega_T). \end{array}$$

Since $|u_\varepsilon| \le 1$ a.e. in Ω_T, the same is true for u. Furthermore, we get $\partial_t u = \nabla\cdot\mathbf{J}$ in $L^2(0,T;(H^1(\Omega))')$. It remains to pass to the limit in

$$\int_{\Omega_T} \mathbf{J}_\theta\cdot\boldsymbol{\eta} = -\int_{\Omega_T} \left[\gamma\Delta u_\theta\nabla\cdot(B(u_\theta)\boldsymbol{\eta}) + \theta\nabla u_\theta\boldsymbol{\eta} - B(u_\theta)\nabla u_\theta\boldsymbol{\eta}\right].$$

The fact that ∇u_θ is uniformly bounded in $L^2(\Omega_T)$ yields

$$\int_{\Omega_T} \theta\nabla u_\theta\boldsymbol{\eta} \longrightarrow 0.$$

All other terms can be handled as in the proof of Theorem 1 to get

$$\int_{\Omega_T} \mathbf{J}\cdot\boldsymbol{\eta} = -\int_{\Omega_T} \left[\gamma\Delta u\nabla\cdot(B(u)\boldsymbol{\eta}) - B(u)\nabla u\cdot\boldsymbol{\eta}\right].$$

This proves that u is a weak solution in the case $B(u) = 1-u^2$ and $\Psi(u) = \frac{1}{2}(1-u^2)$.

We note that we have not proved the convergence of the whole sequence. This is due to the fact that so far there is no uniqueness result for the Cahn–Hilliard equation with a degenerate mobility.

4.3. Other applications. In a paper by Bernis and Friedman [3], the equation

$$u_t = -(f(u)u_{xxx})_x, \tag{4.7}$$

where

$$f(u) = |u|^m f_0(u), \qquad f_0 \in C^{1+\alpha}(\mathbb{R}),\ f_0 > 0, \quad \text{and} \quad m \ge 1, \tag{4.8}$$

was studied. They proved the existence of a nonnegative continuous solution and properties of the support of the solution. For example, they proved that the support increases when $m \ge 2$. We also refer to [3] for other applications of degenerate parabolic equations of higher order.

With straightforward modifications, we can apply our techniques to the following generalization of (4.7) in several space dimensions

$$u_t = \nabla \cdot (f(u)\nabla(-\Delta u + \Psi'(u)))$$

supplemented with Neumann- and no-flux boundary conditions. Under appropriate conditions on the nonlinearities f and Ψ, our method gives the existence of a nonnegative solution in the sense of Theorem 1. In particular, we have to assume that f, f', and $\Psi'' f$ are bounded.

Degenerate parabolic equations of the form

$$u_t = -\nabla \cdot (f(u)(\nabla\Delta u + \nabla u)) + g(t, x, u) \tag{4.9}$$

arising in the theory of plasticity have been independently studied by Grün [17].

5. Conclusion. We proved the existence of a weak solution to the Cahn–Hilliard equation with a degenerate mobility. As was pointed out, our method is also applicable for other fourth-order degenerate parabolic equations. So far, a uniqueness result for fourth-order degenerate parabolic equations has not been established. Methods for proving uniqueness in the case of second-order degenerate parabolic equations seem not to be applicable directly.

Besides studying the question of uniqueness, it is important to get a better understanding of the qualitative behaviour of solutions. Questions are, for example, the following: What kind of singularities occur when $|u| \longrightarrow 1$? What is the evolutionary behaviour of the set $\{|u| = 1\}$? In the case of the deep quench limit, for example, one would expect that the set $\{|u| = 1\}$ develops an interior. If this is the case, one would get a free boundary problem for $\partial\{|u| = 1\}$.

Furthermore, we are interested in the asymptotic behaviour of solutions as $t \longrightarrow \infty$. For second-order degenrate parabolic equations, similarity solutions were important for the understanding of the asymptotic behaviour of solutions. There are results by Bernis, Peletier, and Williams [4] on similarity solutions in one space dimension. It would be interesting to study if similarity solutions in higher space dimensions exist.

REFERENCES

[1] H. W. Alt, *Lineare Funktionalanalysis,* Springer-Verlag, Berlin, 1985.

[2] F. Bai, C. M. Elliott, A. Gardiner, A. Spence, and A. M. Stuart, *The viscous Cahn–Hilliard equation Part* I: *Computations,* Nonlinearity, 8 (1995), pp. 131–160.

[3] F. Bernis and A. Friedman, *Higher order nonlinear degenerate parabolic equations,* J. Differential Equations, 83 (1990), pp. 179–206.

[4] F. Bernis, L. A. Peletier, and S. M. Williams, *Source type solutions of a fourth order nonlinear degenerate parabolic equation,* Nonlinear Anal., 18 (1992), pp. 217–234.

[5] J. F. Blowey and C. M. Elliott, *The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy, part* I: *Mathematical analysis*, European J. Appl. Math., 2 (1991), pp. 233–279.

[6] ———, *The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy, part* II: *Numerical analysis*, European J. Appl. Math., 3 (1992), pp. 147–179.

[7] J. W. Cahn, C. M. Elliott, and A. Novick-Cohen, *The Cahn-Hilliard equation with a concentration dependent mobility: Motion by minus the Laplacian of the mean curvature*, European J. Appl. Math., to appear.

[8] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system* I: *Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.

[9] ———, *Spinodal decomposition: A reprise*, Acta Metallurgica, 19 (1971), pp. 151–161.

[10] J. W. Cahn and J. E. Taylor, *Surface motion by surface diffusion*, Acta Metallurgica, 42 (1994), pp. 1045–1063.

[11] ———, *Linking anisotropic and diffusive surface motion laws via gradient flows,* J. Statist. Phys., 77 (1994), pp. 183–197.

[12] F. Davi and M. Gurtin, *On the motion of a phase interface by surface diffussion,* J. Appl. Math. Phys. (ZAMP), 41 (1990), pp. 782–811.

[13] C. M. Elliott and H. Garcke, *Existence results for diffusive surface motion laws*, Adv. Math. Sci. Appl., to appear.

[14] C. M. Elliott and S. Luckhaus, *A generalized diffusion equation for phase separation of a multi-component mixture with interfacial free energy*, SFB256 preprint 195, Universität Bonn, Bonn, Germany, 1991.

[15] C. M. Elliott and A. M. Stuart, *The viscous Cahn–Hilliard equation, part* II: *Analysis,* J. Differential Equations, to appear.

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

[17] G. Grün, *Degenerate parabolic differential equations of fourth order and a plasticity model with nonlocal hardening*, Z. Anal. Anwendungen, 14 (1995), pp. 541–574.

[18] J. E. Hilliard, *Spinodal decomposition*, in Phase Transformations, American Society for Metals, Cleveland, 1970, pp. 497–560.

[19] J. L. Lions, *Quelque méthodes de résolution de problèmes aux limites non linéaires*, Dunod, Paris, 1969.

[20] A. Novick-Cohen, *On the viscous Cahn–Hilliard equation*, in Material Instabilities in Continuum and Related Mathematical Problems, J. M. Ball, ed., Oxford University Press, Oxford, 1988, pp. 329–342.

[21] W. W. Mullins, *Theory of thermal grooving,* J. Appl. Phys., 28 (1957), pp. 333–339.

[22] J. Simon, *Compact sets in the space* $L^p(0,T;B)$, Ann. Math. Pura Appl., 146 (1987), pp. 65–96.

[23] Y. Jingxue, *On the existence of nonnegative continuous solutions of the Cahn–Hilliard equation*, J. Differential Equations, 97 (1992), pp. 310–327.

SIAM J. MATH. ANAL.
Vol. 27, No. 2, pp. 424–448, March 1996

006

ANALYTICITY OF ESSENTIALLY BOUNDED SOLUTIONS TO SEMILINEAR PARABOLIC SYSTEMS AND VALIDITY OF THE GINZBURG–LANDAU EQUATION[*]

P. TAKÁČ[†], P. BOLLERMAN[‡], A. DOELMAN[§], A. VAN HARTEN[¶], AND E. S. TITI[‖]

Abstract. Some analytic smoothing properties of a general strongly coupled, strongly parabolic semilinear system of order $2m$ in $\mathbb{R}^D \times (0,T)$ with analytic entries are investigated. These properties are expressed in terms of holomorphic continuation in space and time of essentially bounded global solutions to the system. Given $0 < T' < T \leq \infty$, it is proved that any weak, essentially bounded solution $\mathbf{u} = (u_1, \dots, u_N)$ in $\mathbb{R}^D \times (0,T)$ possesses a bounded holomorphic continuation $\mathbf{u}(x+iy, \sigma + i\tau)$ into a region in $\mathbb{C}^D \times \mathbb{C}$ defined by $(x,\sigma) \in \mathbb{R}^D \times (T',T)$, $|y| < A'$ and $|\tau| < B'$, where A' and B' are some positive constants depending upon T'. The proof is based on analytic smoothing properties of a parabolic Green function combined with a contraction mapping argument in a Hardy space H^∞. Applications include weakly coupled semilinear systems of complex reaction-diffusion equations such as the complex Ginzburg–Landau equations. Special attention is given to the problem concerning the validity of the derivation of amplitude equations which describe various instability phenomena in hydrodynamics.

AMS subject classifications. 35K45, 32A35, 35Q55, 76E15

Key words. weak L^∞-solution, analytic Green function, holomorphic continuation, Hardy space, Ginzburg–Landau equation

1. Introduction.

In this article we investigate analyticity of weak L^∞-solutions $\mathbf{u}(x,t) \in \mathbb{R}^N$ (or $\mathbb{C}^N$), for $(x,t) \in \mathbb{R}^D \times (0,T)$, of a general strongly coupled, strongly parabolic system of N semilinear partial differential equations of order $2m$ (where $m \geq 1$ is an integer) having the following form:

$$(1.1)\qquad \begin{aligned} &\frac{\partial \mathbf{u}}{\partial t} + \mathbf{P}\Big(\frac{1}{i}\frac{\partial}{\partial x}\Big)\mathbf{u} = \mathbf{F}\Big(x,t,\Big(\frac{\partial^{|\alpha|}\mathbf{u}}{\partial x^\alpha}\Big)_{|\alpha|\leq \ell}\Big) \quad \text{for } (x,t) \in \mathbb{R}^D \times (0,T),\\ &\mathbf{u}(x,0) = \mathbf{u}_0(x) \quad \text{for } x \in \mathbb{R}^D. \end{aligned}$$

Here, $\partial/\partial x = (\partial/\partial x_1, \dots, \partial/\partial x_D)$ is the gradient and $(\partial^{|\alpha|}\mathbf{u}/\partial x^\alpha)_{|\alpha|\leq\ell}$ denotes the collection of all partial derivatives $\partial^{|\alpha|}\mathbf{u}/\partial x^\alpha = \frac{\partial^{|\alpha|}\mathbf{u}}{\partial x_1^{\alpha_1}\dots\partial x_D^{\alpha_D}}$ of $\mathbf{u}$ up to order ℓ (where ℓ is an integer, $0 \leq \ell < 2m$), where $\alpha = (\alpha_1, \dots, \alpha_D) \in (\mathbb{Z}_+)^D$ is a multiindex of order $|\alpha| = \alpha_1 + \dots + \alpha_D$.

[*] Received by the editors January 31, 1994; accepted for publication (in revised form) July 27, 1994.

[†] Department of Mathematics, Washington State University, Pullman, WA 99164–3113 (takac@delta.math.wsu.edu). The research of this author was supported in part by National Science Foundation grant DMS-9401418.

[‡] Mathematisch Instituut, Rijksuniversiteit Utrecht, Budapestlaan 6, Postbus 80.010, NL-3508 TA Utrecht, The Netherlands (bollerma@math.ruu.nl). The research of this author was supported in part by Dutch National Research Foundation grant 611-306-528.

[§] Mathematisch Instituut, Rijksuniversiteit Utrecht, Budapestlaan 6, Postbus 80.010, NL-3508 TA Utrecht, The Netherlands (doelman@math.ruu.nl).

[¶] Mathematisch Instituut, Rijksuniversiteit Utrecht, Budapestlaan 6, Postbus 80.010, NL-3508 TA Utrecht, The Netherlands (harten@math.ruu.nl).

[‖] Department of Mathematics and Department of Mechanical and Aerospace Engineering, University of California at Irvine, Irvine, CA 92717–3875 (etiti@math.uci.edu) and Center for Applied Mathematics, Cornell University, Ithaca, NY 14853. The research of this author was supported in part by the National Science Foundation, Air Force Office of Scientific Research grant F49620-92-J-0287, and the Joint University of California—Los Alamos National Laboratory Institute for Cooperative Research.

As usual, $\mathbb{R}^N$ and $\mathbb{C}^N$, respectively, denote the N-dimensional real and complex Euclidean spaces, $i = \sqrt{-1}$ and $D, N \in \mathbb{N}$, where $\mathbb{N} = \mathbb{Z}_+ \setminus \{0\}$ and $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$.

Our main objective is to cover the following prototype problem which inspired our work reported here, namely, the complex Ginzburg–Landau equation (derived in Newell and Whitehead [29]):

$$\begin{aligned} (1.2) \qquad & \frac{\partial u}{\partial t} - (1 + i\nu)\Delta u = Ru - (1 + i\mu)|u|^2 u \quad \text{for } (x, t) \in \mathbb{R}^D \times (0, T), \\ & u(x, 0) = u_0(x) \quad \text{for } x \in \mathbb{R}^D, \end{aligned}$$

and weakly coupled systems of such equations as well. Here, $\Delta = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_D^2}$ denotes the Laplacian, $\mu, \nu, R \in \mathbb{R}$ are given constants, and $u(x, t) \in \mathbb{C}$ is the unknown function. We assume that $u \in L^\infty(\mathbb{R}^D \times (0, T))$ is a weak L^∞-solution of (1.2) which we define later in §2, Definition 2.1.

Although $|u|^2 = u_1^2 + u_2^2$ is not a holomorphic (i.e. complex analytic) function of $u \equiv u_1 + iu_2 \in \mathbb{C} \equiv \mathbb{R} \oplus i\mathbb{R}$, it is still real analytic in u_1 and u_2. The following example shows how (1.2) can be turned into a system of two real equations with real analytic nonlinearities satisfying all hypotheses we impose below on our general system (1.1).

Example 1.1. Let $u \equiv u_1 + iu_2$ be the unknown function in (1.2), where $u_1 = \Re\mathfrak{e}\, u$ and $u_2 = \Im\mathfrak{m}\, u$, and set $\mathbf{u} = \binom{u_1}{u_2}$. Then (1.2) is equivalent to the following system for the unknown $\mathbf{u}$:

$$\begin{aligned} (1.3) \quad & \frac{\partial \mathbf{u}}{\partial t} - \begin{pmatrix} 1 & -\nu \\ \nu & 1 \end{pmatrix} \Delta \mathbf{u} = R\mathbf{u} - |\mathbf{u}|^2 \begin{pmatrix} 1 & -\mu \\ \mu & 1 \end{pmatrix} \mathbf{u} \quad \text{for } (x, t) \in \mathbb{R}^D \times (0, T), \\ & \mathbf{u}(x, 0) = \mathbf{u}_0(x) \quad \text{for } x \in \mathbb{R}^D. \end{aligned}$$

As usual, $\Delta\mathbf{u} = \binom{\Delta u_1}{\Delta u_2}$ and $|\mathbf{u}|^2 = u_1^2 + u_2^2$. Both components of the nonlinearity

$$\mathbf{F}(\mathbf{u}) \equiv \begin{pmatrix} F_1(\mathbf{u}) \\ F_2(\mathbf{u}) \end{pmatrix} = (u_1^2 + u_2^2) \begin{pmatrix} u_1 - \mu u_2 \\ \mu u_1 + u_2 \end{pmatrix}$$

are now third-degree polynomials in u_1 and u_2. Hence, (1.3) is a real analytic system with an obvious canonical extension to a holomorphic one.

Rigorous investigation of the validity of (1.2) presented in van Harten [15] and Bollerman [3] exploits the kind of analyticity results we show in the present article. Typically, (1.2) is an amplitude equation which is formally derived by means of asymptotic expansion from other equations of mathematical physics, such as the Navier–Stokes equations or the Rayleigh–Bénard convection equations and many others, in the context of studying weakly nonlinear instabilities of particular solutions of those equations. This derivation involves an asymptotic expansion of the solution in terms of a small control parameter $\varepsilon > 0$, which in turn makes use of the analyticity and asymptotics (as $|x| \to \infty$) of the solution, cf. [3], [6], [15], [22], [29], [31], and [32]. To give a rigorous validation of the formal derivation one needs to prove suitable analyticity results for the solution of (1.2). This is the content of §5 in our article.

With the notation of Example 1.1, in [15] (and similarly in [3]) it is assumed that $D = 1$ and the initial distribution $\mathbf{u}(\cdot, 0) = \mathbf{u}_0 \colon \mathbb{R} \to \mathbb{C}^2$ has a bounded holomorphic continuation $\mathbf{u}_0(x + iy)$ into a strip $S_a = \{z = x + iy \in \mathbb{C} \colon |y| < a\}$ for some constant $a \in (0, \infty)$. This analyticity hypothesis on $\mathbf{u}_0$ is weakened in our present work. We assume only $\mathbf{u}_0 \in L^\infty(\mathbb{R} \to \mathbb{C}^2)$ and prove an analytic smoothing property for time $t \in (0, T)$.

In order to be able to treat our general system (1.1) in a simple but reasonably general way, we assume that

$$\mathbf{P}\Big(\frac{1}{i}\frac{\partial}{\partial x}\Big) = \sum_{|\alpha|\le 2m} i^{-|\alpha|}\mathbf{P}^{(\alpha)}\frac{\partial^{|\alpha|}}{\partial x^{\alpha}} \tag{1.4}$$

is a strongly elliptic, linear, partial differential operator of order $2m$ with constant coefficients $i^{-|\alpha|}\mathbf{P}^{(\alpha)}$, where each $i^{-|\alpha|}\mathbf{P}^{(\alpha)} = (i^{-|\alpha|}P_{jk}^{(\alpha)})_{j,k=1}^{N}$ is an $N \times N$ matrix with real (or complex) entries $i^{-|\alpha|}P_{jk}^{(\alpha)}$. The strong ellipticity hypothesis for $\mathbf{P}$ means that there exists a constant $c \in (0,\infty)$ such that the following inequality holds for all $\xi = (\xi_1, \dots, \xi_D) \in \mathbb{R}^D$ and $\boldsymbol{\eta} = (\eta_1, \dots, \eta_N) \in \mathbb{R}^N$ (or $\mathbb{C}^N$):

$$\Re e\Big(\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{|\alpha|=2m} P_{jk}^{(\alpha)}\xi^{\alpha}\eta_k\eta_j^{*}\Big) \ge c|\xi|^{2m}|\boldsymbol{\eta}|^2, \tag{1.5}$$

where $\xi^{\alpha} = \xi_1^{\alpha_1}\dots\xi_D^{\alpha_D}$ for $\alpha = (\alpha_1, \dots, \alpha_D) \in (\mathbb{Z}_+)^D$. We use an asterisk (*) to denote the complex conjugate of a number in the complex plane $\mathbb{C}$.

The nonlinear mapping $\mathbf{F}\colon \mathbb{R}^D \times (0,T) \times \mathbb{R}^{N(1+D+\cdots+D^{\ell})} \to \mathbb{R}^N$ (or $\mathbb{R}^D \times (0,T) \times \mathbb{C}^{N(1+D+\cdots+D^{\ell})} \to \mathbb{C}^N$) is assumed to be real (or complex) analytic in all its variables

$$(x,t) \in \mathbb{R}^D \times (0,T) \quad \text{and} \quad \boldsymbol{\zeta}_{\alpha} = \frac{\partial^{|\alpha|}}{\partial x^{\alpha}}\mathbf{u}(x,t) \in \mathbb{R}^N \text{ (or } \mathbb{C}^N) \quad \text{for } |\alpha| \le \ell.$$

Moreover, we assume that $\mathbf{F}$ possesses a holomorphic continuation $\mathbf{F}(x+iy, \sigma + i\tau, (\boldsymbol{\xi}_{\alpha} + i\boldsymbol{\eta}_{\alpha})_{|\alpha|\le\ell})$ into a region $\Omega_{\mathbf{F}} = \Omega'_{\mathbf{F}} \times \mathbb{C}^{N(1+D+\cdots+D^{\ell})}$, where $\Omega'_{\mathbf{F}} \subset \mathbb{C}^D \times \mathbb{C}$ is defined by

$$\Omega'_{\mathbf{F}} = \{(x+iy, \sigma+i\tau)\colon (x,\sigma) \in \mathbb{R}^D \times (0,T),\ |y| < A \ \text{ and } \ |\tau| < B\},$$

such that $\mathbf{F}$ and all partial Fréchet derivatives $\partial\mathbf{F}/\partial\boldsymbol{\zeta}_{\alpha}$, for $|\alpha| \le \ell$, are bounded in $\Omega'_{\mathbf{F}} \times \mathbf{B}$ for every bounded set $\mathbf{B} \subset \mathbb{C}^{N(1+D+\cdots+D^{\ell})}$. Here, $A, B \in (0,\infty)$ are some constants.

A brief, weaker version of our main result, Theorem 2.1 in §2, can be stated as follows.

THEOREM 1.1. *Let A and B be as above, and $0 < T' < T \le \infty$. Assume that $\mathbf{u}$ is a weak L^{∞}-solution of our system (1.1) in $\mathbb{R}^D \times (0,T)$ such that*

$$M'_{\alpha} \overset{\text{def}}{=} \operatorname*{ess\,sup}_{\mathbb{R}^D\times(0,T)}\left|\frac{\partial^{|\alpha|}\mathbf{u}}{\partial x^{\alpha}}\right| < \infty \quad \textit{for } |\alpha| \le \ell.$$

Then the function $\mathbf{u}$ has a bounded holomorphic continuation $\mathbf{u}(x+iy, \sigma+i\tau)$ into a region $\Omega'_{\mathbf{u}} \subset \mathbb{C}^D \times \mathbb{C}$ defined by

$$\Omega'_{\mathbf{u}} = \{(x+iy, \sigma+i\tau)\colon (x,\sigma) \in \mathbb{R}^D \times (T',T),\ |y| < A' \ \textit{ and } \ |\tau| < B'\},$$

such that all partial derivatives $\partial^{|\alpha|}\mathbf{u}/\partial x^{\alpha}$, for $|\alpha| \le \ell$, are bounded in $\Omega'_{\mathbf{u}}$. Here, $A' \in (0,A]$ and $B' \in (0,B]$ are some constants depending upon T' and $\mathbf{u}$ solely through the numbers $M'_{\alpha} \in [0,\infty)$ for $|\alpha| \le \ell$.

Remark 1.1. In our Theorem 2.1 below, we specify also the dependence of the constants $A' \equiv A'(T')$ and $B' \equiv B'(T')$ upon $T' \in (0,T)$. In general, we cannot take

$T' = 0$ unless $A' = 0$ or $B' = 0$. Namely, given any $A' \in (0, A]$ and $B' \in (0, B]$, the holomorphic continuation $\mathbf{u}(x+iy, \sigma+i\tau)$ of the solution $\mathbf{u}$ into the region $\Omega_{\mathbf{u}} \subset \mathbb{C}^D \times \mathbb{C}$ defined by

$$\Omega_{\mathbf{u}} = \{(x+iy, \sigma+i\tau)\colon (x,\sigma) \in \mathbb{R}^D \times (0,T),\ |y| < A' \text{ and } |\tau| < B'\}$$

may either fail to exist or else fail to be bounded, cf. Remark 2.6 below, unless $\mathbf{u}_0$ has a bounded holomorphic continuation $\mathbf{u}_0(x+iy)$ into the strip $S_{A'} = \{z = x+iy \in \mathbb{C}^D\colon |y| < A'\}$, in which case one can apply the results of [15].

Some results about the analyticity of solutions of nonlinear parabolic systems, which are related to ours, are stated in Friedman [13, Thms. 3 and 4] without proofs. For the Navier–Stokes equations, such analyticity results are proved in Masuda [26]. These results state local analyticity of infinitely differentiable solutions without any description of their domain of holomorphy (i.e., domain of complex analyticity). Our present article provides such a description in Theorem 2.1.

As we have already indicated, our main result, Theorem 2.1, is stated in §2 and proved in §3. Its proof is based on some well-known analytic smoothing properties of the Green function corresponding to the initial value problem for the linear part of System (1.1), which are stated and proved in the Appendix. These properties are combined with a standard contraction-mapping argument in a Hardy space H^∞ of bounded holomorphic functions (rather than a commonly used Hölder or Sobolev space). The contractivity is obtained via Cauchy's theorem (path independence of the integral). In §4 we suggest possible generalizations of our main result to a wider class of systems (1.1) involving a much more general pseudodifferential operator as their linear part. In §5 we discuss direct applications of our main result to the validity problem for the complex Ginzburg–Landau equation (1.2). Finally, §6 contains a discussion about further applications of our results to some current issues in mathematical physics and dynamical systems. We also compare previously known results to ours.

2. Statement of the main result. Recalling $D, N \in \mathbb{N}$ and $\ell, m \in \mathbb{Z}_+$ with $0 \le \ell < 2m$, we introduce the following notation: The complexifications of the space and time variables, respectively, are denoted by $z = x + iy \in \mathbb{C}^D \equiv \mathbb{R}^D \oplus i\mathbb{R}^D$ and $t = \sigma + i\tau \in \mathbb{C} \equiv \mathbb{R} \oplus i\mathbb{R}$. As usual, $|\cdot|$ denotes the Euclidean norm in a real or complex Euclidean space.

Given any Lebesgue-measurable set Ω in a Euclidean space, we denote by $\mathbf{L}^\infty(\Omega) = L^\infty(\Omega \to \mathbb{C}^N)$ the Lebesgue space of all (equivalence classes of) Lebesgue-measurable, essentially bounded functions $\mathbf{f} = (f_1, \dots, f_N)\colon \Omega \to \mathbb{C}^N$ endowed with the norm

$$\|\mathbf{f}\|_\infty = \operatorname*{ess\,sup}_{\omega \in \Omega} |\mathbf{f}(\omega)| < \infty.$$

Recall that $\mathbf{L}^\infty(\Omega)$ is the dual space of $\mathbf{L}^1(\Omega) = L^1(\Omega \to \mathbb{C}^N)$. Hence, any bounded closed ball in $\mathbf{L}^\infty(\Omega)$ endowed with the weak*-topology is a compact metrizable space. In particular, a sequence $\{\mathbf{f}^n\}_{n=1}^\infty \subset \mathbf{L}^\infty(\Omega)$ converges to $\mathbf{f}^0 \in \mathbf{L}^\infty(\Omega)$ $(n \to \infty)$ in the weak*-topology on $\mathbf{L}^\infty(\Omega)$ if and only if

$$\int_\Omega f_j^n(x)\phi(x)\,dx \to \int_\Omega f_j^0(x)\phi(x)\,dx \qquad \text{as } n \to \infty,$$

for every $\phi \in L^1(\Omega)$ and $j = 1, \dots, N$.

We define the notion of a weak L^∞-solution of our system (1.1) as follows.

Definition 2.1. *Let $T \in (0, \infty]$. A function $\mathbf{u}\colon \mathbb{R}^D \times (0,T) \to \mathbb{C}^N$ is called a weak L^∞-solution of system* (1.1) *if it satisfies the following four conditions:*

(i) $\mathbf{u}$ *and all distributional partial derivatives* $\partial^{|\alpha|}\mathbf{u}/\partial x^{\alpha}$, *for* $|\alpha| \le \ell$, *belong to* $\mathbf{L}^{\infty}(\mathbb{R}^D \times (0,T))$.

(ii) $t \in [0,T) \mapsto \mathbf{u}(\cdot,t) \in \mathbf{L}^{\infty}(\mathbb{R}^D)$ *is a weakly* continuous function of* t *with the weak*-limit* $\mathbf{u}_0 \in \mathbf{L}^{\infty}(\mathbb{R}^D)$ *as* $t \to 0+$. *The function* $\mathbf{u}(\cdot,0) = \mathbf{u}_0$ *is called the initial value of* $\mathbf{u}$.

(iii) $(x,t) \in \mathbb{R}^D \times (0,T) \mapsto \mathbf{F}\big(x,t,\big(\frac{\partial^{|\alpha|}}{\partial x^{\alpha}}\mathbf{u}(x,t)\big)_{|\alpha|\le\ell}\big) \in \mathbb{C}^N$ *is a locally integrable function.*

(iv) *Equation* (1.1) *is satisfied in the sense of distributions over* $\mathbb{R}^D \times (0,T)$.

Remark 2.1. If $0 < \ell < 2m$, Conditions (i) and (ii) from Definition 2.1 above imply also that each distributional partial derivative $t \in (0,T) \mapsto \partial^{|\alpha|}\mathbf{u}(\cdot,t)/\partial x^{\alpha} \in \mathbf{L}^{\infty}(\mathbb{R}^D)$, for $0 < |\alpha| \le \ell$, is a weakly* continuous function of t with the weak*-limit $\partial^{|\alpha|}\mathbf{u}_0/\partial x^{\alpha} \in \mathbf{L}^{\infty}(\mathbb{R}^D)$ as $t \to 0+$. This claim is obvious since each of these functions is valued in a bounded closed ball in $\mathbf{L}^{\infty}(\mathbb{R}^D)$, i.e., in a metrizable weak*-compact set. In particular, we have also $\partial^{|\alpha|}\mathbf{u}_0/\partial x^{\alpha} \in \mathbf{L}^{\infty}(\mathbb{R}^D)$ for $|\alpha| \le \ell$.

Remark 2.2. Condition (iii) from Definition 2.1 is automatically satisfied provided $\mathbf{F}$ satisfies Hypothesis (H2) below and $\mathbf{u}$ satisfies condition (i) from Definition 2.1.

The reader is referred to Edwards [11, Chap. 5] for the theory of distributions and weak topologies.

Let $\kappa_0, \nu_0 \in (0,\infty)$ be two constants to be specified later, and $s \in [0,\infty]$. We set

$$\Pi^{(s)}(\kappa_0) = \{z = x + iy \in \mathbb{C}^D \colon |y|^{2m} < \kappa_0 s\}, \tag{2.1}$$

$$\Sigma^{(s)}(\nu_0) = \{t = \sigma + i\tau \in \mathbb{C} \colon \nu_0|\tau| < \sigma = s\}, \tag{2.2}$$

and introduce the open parabolic region

$$\Lambda^{(s)}(\kappa_0,\nu_0) = \cup\{\Pi^{(r)}(\kappa_0) \times \Sigma^{(r)}(\nu_0) \colon r \in (0,s)\} \subset \mathbb{C}^D \times \mathbb{C}, \tag{2.3}$$

together with its time translation by r units, for $0 \le r < \infty$,

$$\Lambda_r^{(s)}(\kappa_0,\nu_0) = \{(z,t) \in \mathbb{C}^D \times \mathbb{C} \colon (z,t-r) \in \Lambda^{(s)}(\kappa_0,\nu_0)\}. \tag{2.4}$$

We now define our most important region in $\mathbb{C}^D \times \mathbb{C}$, for $0 \le s \le T \le \infty$,

$$\Gamma_T^{(s)}(\kappa_0,\nu_0) = \begin{cases} \cup\{\Lambda_r^{(s)}(\kappa_0,\nu_0) \colon r \in [0,T-s]\} & \text{if } s < T, \\ \Lambda^{(s)}(\kappa_0,\nu_0) & \text{if } s = T. \end{cases} \tag{2.5}$$

Given any $r \in [0,T)$, we observe that the time r section of $\Gamma_T^{(s)}(\kappa_0,\nu_0)$ is given by

$$\begin{aligned} \Theta^{(r,s)}(\kappa_0,\nu_0) &\overset{\text{def}}{=} \{(z,t) \in \Gamma_T^{(s)}(\kappa_0,\nu_0) \colon \Re\mathfrak{e}\, t = r\} \\ &= \Pi^{(r')}(\kappa_0) \times \Sigma^{(r')}(\nu_0), \qquad \text{where } r' = \min\{r,s\}. \end{aligned} \tag{2.6}$$

When discussing the validity of the derivation of (1.2) in §5, we will keep in mind that the regions defined above depend also on $2m$. In particular, we write $\Gamma_T^{(s)}(\kappa_0,\nu_0;2m) \equiv \Gamma_T^{(s)}(\kappa_0,\nu_0)$.

Finally, we denote by $\mathbf{B}(M)$ the open ball in $\mathbb{C}^N$ of radius M centered at the origin, and by $\overline{\mathbf{B}}(M)$ its closure.

Hypothesis. From now on we assume that $T \in (0,\infty]$ is fixed. We impose the following hypotheses on $\mathbf{P}$ and $\mathbf{F}$, respectively:

(H1) $\mathbf{P}$ satisfies (1.4) and (1.5).

(H2) There exist constants $\kappa_0, \nu_0 \in (0,\infty)$, $T_0 \in (0,T]$, and $M_\alpha \in (0,\infty)$, for $|\alpha| \leq \ell$, such that $\mathbf{F}\colon \mathbb{R}^D \times (0,T) \times \mathbb{R}^{N(1+D+\cdots+D^\ell)} \to \mathbb{R}^N$ (or $\mathbb{R}^D \times (0,T) \times \mathbb{C}^{N(1+D+\cdots+D^\ell)} \to \mathbb{C}^N$) possesses a holomorphic continuation $\mathbf{F}(x+iy, \sigma+i\tau, (\boldsymbol{\xi}_\alpha + i\boldsymbol{\eta}_\alpha)_{|\alpha|\leq\ell})$ into the region (cf. (2.5))

$$\Omega_{\mathbf{F}} = \Gamma_T^{(T_0)}(\kappa_0,\nu_0) \times \prod_{|\alpha|\leq\ell} \mathbf{B}(M_\alpha) \subset (\mathbb{C}^D \times \mathbb{C}) \times \mathbb{C}^{N(1+D+\cdots+D^\ell)}$$

such that $\mathbf{F}$ and all partial Fréchet derivatives $\partial\mathbf{F}/\partial\boldsymbol{\zeta}_\alpha$ with respect to $\boldsymbol{\zeta}_\alpha = \boldsymbol{\xi}_\alpha + i\boldsymbol{\eta}_\alpha \in \mathbb{C}^N$, for $|\alpha| \leq \ell$, are bounded in $\Omega_{\mathbf{F}}$.

We set

$$C_{\mathbf{F}} \overset{\text{def}}{=} \sup_{\Omega_{\mathbf{F}}} |\mathbf{F}| < \infty \quad \text{and} \quad C'_{\mathbf{F}} \overset{\text{def}}{=} \max_{|\alpha|\leq\ell} \sup_{\Omega_{\mathbf{F}}} \left|\frac{\partial\mathbf{F}}{\partial\boldsymbol{\zeta}_\alpha}\right| < \infty. \tag{2.7}$$

Of course, a holomorphic continuation into $\Omega_{\mathbf{F}}$ of a real- (or complex-) valued function that is analytic in $\mathbb{R}^D \times (0,T) \times \mathbb{R}^{N(1+D+\cdots+D^\ell)}$ is always unique, cf. John [21, Chap. 3, §3(c), pp. 70–72]. We refer to Hörmander [17, Chap. 2] for basic facts about complex analysis in several variables.

Remark 2.3. For instance, all constants $M_\alpha \in (0,\infty)$ in Hypothesis (H2) may be chosen arbitrarily large in case $\mathbf{F}(x+iy, \sigma+i\tau, (\boldsymbol{\xi}_\alpha + i\boldsymbol{\eta}_\alpha)_{|\alpha|\leq\ell})$ is an entire function (e.g., a polynomial) in all the variables $\boldsymbol{\zeta}_\alpha = \boldsymbol{\xi}_\alpha + i\boldsymbol{\eta}_\alpha \in \mathbb{C}^N$, $|\alpha| \leq \ell$ having a power series expansion all of whose coefficients are bounded holomorphic functions from $\Gamma_T^{(T_0)}(\kappa_0,\nu_0)$ to $\mathbb{C}^{N\times N}$. It is easy to see that (H2) is true in most practical applications.

The main result in this article can be stated as follows.

THEOREM 2.1. *Let* $\mathbf{P}$ *and* $\mathbf{F}$ *satisfy Hypotheses* (H1) *and* (H2) *above, where* $0 < T_0 \leq T \leq \infty$, $\kappa_0, \nu_0 \in (0,\infty)$, *and* $M_\alpha \in (0,\infty)$ *for* $|\alpha| \leq \ell$. *Then there exist three constants* $\nu'_0 \in [\nu_0,\infty)$, $C_{\mathbf{P}} \equiv C_{\mathbf{P}}(\kappa_0,\nu'_0) \in (1,\infty)$, *and* $T'_0 \in (0,T_0]$ *with the following property:*

Assume that $\mathbf{u}$ *is a weak* L^∞*-solution of our system* (1.1) *in* $\mathbb{R}^D \times (0,T)$ *such that*

$$\operatorname*{ess\,sup}_{\mathbb{R}^D\times(0,T)} \left|\frac{\partial^{|\alpha|}\mathbf{u}}{\partial x^\alpha}\right| \leq M'_\alpha \overset{\text{def}}{=} \frac{M_\alpha}{C_{\mathbf{P}}} \quad \textit{for } |\alpha| \leq \ell. \tag{2.8}$$

Then the function $\mathbf{u}$ *has a holomorphic continuation* $\mathbf{u}(x+iy, \sigma+i\tau)$ *into the region* $\Omega'_{\mathbf{u}} = \Gamma_T^{(T'_0)}(\kappa_0,\nu'_0) \subset \mathbb{C}^D \times \mathbb{C}$ *such that*

$$\sup_{\Omega'_{\mathbf{u}}} \left|\frac{\partial^{|\alpha|}\mathbf{u}}{\partial x^\alpha}\right| < M_\alpha \quad \textit{for } |\alpha| \leq \ell. \tag{2.9}$$

Furthermore, the continuation of $\mathbf{u}$ *satisfies* (1.1) *in* $\Omega'_{\mathbf{u}}$.

The constant $\nu'_0 \in [\nu_0,\infty)$ *is completely determined by* $\mathbf{P}$, *whereas* $C_{\mathbf{P}} \in (1,\infty)$ *is completely determined by* $\mathbf{P}$, κ_0, *and* ν'_0. *In particular, both* ν'_0 *and* $C_{\mathbf{P}}$ *are independent from the nonlinearity* $\mathbf{F}$ *and* M_α.

Remark 2.4. Strictly speaking, in Theorem 2.1 we do not need the existence of a weak L^∞-solution $\mathbf{u}$ of (1.1) in $\mathbb{R}^D \times (0,T)$. We need only the a priori bounds (2.8). Nevertheless, in our proof of Theorem 2.1 we will need the uniqueness of a weak L^∞-solution of (1.1), subject to a given initial condition. In case the a priori bounds

(2.8) can be established for a particular weak L^∞-solution $\mathbf{u}$ of (1.1) in $\mathbb{R}^D \times (0,T)$, subject to a given initial condition $\mathbf{u}(\cdot,0) = \mathbf{u}_0(\cdot) \in \mathbf{L}^\infty(\mathbb{R}^D)$, then the existence and uniqueness of such a solution is obtained by a well-known contraction mapping argument (identical with ours in the proof of Theorem 3.1, Step 4, except for the choice of space) in a closed ball in the Sobolev space $\mathcal{W}^{\ell,\infty} \equiv \mathcal{W}^{\ell,\infty}(\mathbb{R}^D \times (0,T))$. Here, $\mathcal{W}^{\ell,\infty}$ consists of all functions $\mathbf{f}\colon \mathbb{R}^D \times (0,T) \to \mathbb{C}^N$ of $(x,t) \in \mathbb{R}^D \times (0,T)$ such that $\mathbf{f}$ satisfies Conditions (i) and (ii) from Definition 2.1. Regarding the derivatives in Condition (i) as distributions, we observe that $\mathcal{W}^{\ell,\infty}$ is isomorphic to a closed linear subspace of the L^∞-product space $(\mathbf{L}^\infty(\mathbb{R}^D \times (0,T)))^{1+D+\cdots+D^\ell}$, the isomorphism being defined by $\mathbf{f} \mapsto (\partial^{|\alpha|}\mathbf{f}/\partial x^\alpha)_{|\alpha|\le\ell}$. In particular, $\mathcal{W}^{\ell,\infty}$ is a complex Banach space endowed with the norm

$$(2.10)\qquad \|\mathbf{f}\|_{\mathcal{W}^{\ell,\infty}} \overset{\text{def}}{=} \max_{|\alpha|\le\ell} \operatorname*{ess\,sup}_{\mathbb{R}^D\times(0,T)} \left|\frac{\partial^{|\alpha|}\mathbf{f}}{\partial x^\alpha}\right|.$$

Notice that $\mathcal{W}^{0,\infty}$ is a proper, closed linear subspace of $\mathbf{L}^\infty(\mathbb{R}^D \times (0,T))$.

We refer to Stein [33, Chap. V] for the basic theory of Sobolev spaces.

Remark 2.5. By Definition 2.1 and Remark 2.1, the inequalities (2.8) imply that the initial value $\mathbf{u}_0$ of a weak L^∞-solution of (1.1) satisfies $\partial^{|\alpha|}\mathbf{u}_0/\partial x^\alpha \in \mathbf{L}^\infty(\mathbb{R}^D)$ for all $|\alpha| \le \ell$, together with

$$(2.11)\qquad \operatorname*{ess\,sup}_{\mathbb{R}^D}\left|\frac{\partial^{|\alpha|}\mathbf{u}_0}{\partial x^\alpha}\right| \le M'_\alpha \overset{\text{def}}{=} \frac{M_\alpha}{C_{\mathbf{P}}} \qquad \text{for } |\alpha| \le \ell.$$

Only these bounds, not (2.8), will be used in our proof of the local (in time) version of Theorem 2.1 below, i.e., Theorem 3.1, for some $T = T_0 = T'_0 \in (0,\infty)$ small enough. The bounds (2.8) will be used subsequently to extend the local solution to the entire time interval $(0,T)$.

Remark 2.6. In Theorem 2.1 above we have specified also the dependence of the constants $A' \equiv A'(T')$ and $B' \equiv B'(T')$ upon $T' \in (0,T)$, used in its weaker version, Theorem 1.1. We have $A'(T') = (\kappa_0 T')^{1/2m}$ and $B'(T') = T'/\nu'_0$ for $T' \in (0,T'_0)$, and $A'(T') = (\kappa_0 T'_0)^{1/2m}$ and $B'(T') = T'_0/\nu'_0$ for $T' \in [T'_0,T)$. In general, we cannot take $T' = 0$ unless $A' = 0$ or $B' = 0$, even if the nonlinear system (1.1) takes the following simple form:

$$(2.12)\qquad \begin{aligned} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} &= 0 \quad \text{for } (x,t) \in \mathbb{R} \times (0,\infty),\\ u(x,0) &= u_0(x) \quad \text{for } x \in \mathbb{R}. \end{aligned}$$

The function (cf. John [21, Chap. 7, §1(a), p. 213])

$$u(x,t) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{x/\sqrt{4t}} e^{-\xi^2}\, d\xi, \qquad (x,t) \in \mathbb{R} \times (0,\infty)$$

is a real analytic solution to (2.12) satisfying $0 < u < 1$ in $\mathbb{R} \times (0,\infty)$ and $u(x,t) \to u_0(x)$ as $t \to 0+$, for every $x \in \mathbb{R}$, where

$$u_0(x) = \begin{cases} 0 & \text{if } x < 0,\\ 1/2 & \text{if } x = 0,\\ 1 & \text{if } x > 0. \end{cases}$$

We take $\ell = 0$ in Theorem 2.1 since there is no nonlinearity. By Cauchy's theorem (path independence of the integral), the holomorphic continuation of u into the region $\mathbb{C} \times ((0,\infty) + i\mathbb{R})$ is given by

$$u(x+iy,t) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{x/\sqrt{4t}} e^{-\xi^2}\, d\xi + \frac{1}{\sqrt{\pi}} \int_0^{y/\sqrt{4t}} e^{-((x/\sqrt{4t})+i\eta)^2}\, d\eta \tag{2.13}$$
$$\text{for } (x+iy,t) \in \mathbb{C} \times \mathbb{C} \text{ with } \mathfrak{Re}\, t > 0.$$

It is easy to see that the second integral is unbounded as $t \to 0+$, $t \in \mathbb{R}$, for any fixed $x+iy \in \mathbb{C}$ with $y \neq 0$. Furthermore, its upper bound $y/\sqrt{4(\sigma+i\tau)}$ clearly shows the optimality of our choice of a region in which $u(x+iy, \sigma+i\tau)$ is bounded.

We now return to Example 1.1 to illustrate a possible choice of the constant $\nu_0' \in [\nu_0, \infty)$ in Theorem 2.1 above.

Example 2.1. Let us consider system (1.3) again. The Green's function $\mathbf{G}(x,x';t)$ corresponding to the linear initial value problem

$$\frac{\partial \mathbf{u}}{\partial t} - \begin{pmatrix} 1 & -\nu \\ \nu & 1 \end{pmatrix} \Delta \mathbf{u} = \mathbf{0} \quad \text{for } (x,t) \in \mathbb{R}^D \times (0,\infty), \tag{2.14}$$
$$\mathbf{u}(x,0) = \mathbf{u}_0(x) \quad \text{for } x \in \mathbb{R}^D$$

is given by the formula (cf. John [21, Chap. 7, §1(a), p. 209])

$$\mathbf{G} = \begin{pmatrix} G_1 & -G_2 \\ G_2 & G_1 \end{pmatrix}, \quad \text{where } G_1 = \mathfrak{Re}\, G,\ G_2 = \mathfrak{Im}\, G \text{ and} \tag{2.15}$$
$$G(x,x';t) \equiv G(x-x';t) \overset{\text{def}}{=} (4\pi(1+i\nu)t)^{-D/2} \exp\Big(-\frac{|x-x'|^2}{4(1+i\nu)t}\Big)$$
$$\text{for } x, x' \in \mathbb{R}^D \text{ and } t \in (0,\infty).$$

Hence, the solution of the Cauchy problem (2.14) is given by the convolution $\mathbf{u}(t) = \mathbf{G}(\cdot;t) * \mathbf{u}_0$ for $t \geq 0$. Given any $s \in (0,\infty)$ fixed, the family of functions $|G(\cdot;t)|$, for $t = s+i\tau \in \Sigma^{(s)}(\nu_0')$, has a common integrable majorant over $\mathbb{R}^D$ if and only if there is a constant $\gamma > 0$ such that $\gamma\, \mathfrak{Re}((1+i\nu)t) = \gamma(s-\nu\tau) > s$ for every $t \in \Sigma^{(s)}(\nu_0')$. This is the case if and only if $|\nu| < \nu_0'$. Then such a majorant can be taken to be the function $Cs^{-D/2}\exp(-\gamma|\cdot|^2/s)$, where $C \in (0,\infty)$ is a constant. Moreover, if $|\nu| < \nu_0'$ and $\kappa_0 \in (0,\infty)$ is arbitrary, we can choose the constants $\gamma, C \in (0,\infty)$ so that the holomorphic continuation of G to $\Lambda^{(\infty)}(\kappa_0, \nu_0')$, cf. (2.3), satisfies

$$|G(z;t)| \leq C\sigma^{-D/2} e^{-\gamma|x|^2/\sigma}, \quad (z,t) = (x+iy, \sigma+i\tau) \in \Lambda^{(\infty)}(\kappa_0, \nu_0'). \tag{2.16}$$

This estimate plays a crucial role in our proof of Theorem 2.1. It provides the following bound on the holomorphic continuation to $\Lambda^{(\infty)}(\kappa_0, \nu_0')$ of the solution $\mathbf{u}$ of (2.14):

$$|\mathbf{u}(z,t)| \leq C_1 \operatorname*{ess\,sup}_{\mathbb{R}^D} |\mathbf{u}_0| \quad \text{for } (z,t) \in \Lambda^{(\infty)}(\kappa_0, \nu_0'), \tag{2.17}$$

where $C_1 \in (0,\infty)$ is a constant independent from $\mathbf{u}_0 \in \mathbf{L}^\infty(\mathbb{R}^D)$.

3. Proof of the main result. We implement our contraction-mapping argument (Banach's fixed-point theorem) in a closed ball in a Hardy space H^∞. Given any open set Ω in a complex Euclidean space, we write $\mathbf{H}^\infty(\Omega) = H^\infty(\Omega \to \mathbb{C}^N)$ to denote the Hardy space of all bounded holomorphic functions $\mathbf{f} = (f_1, \dots, f_N)\colon \Omega \to \mathbb{C}^N$

endowed with the $\mathbf{L}^\infty$-norm $\|\mathbf{f}\|_\infty$. It is well known that $\mathbf{H}^\infty(\Omega)$ is a closed linear subspace of $\mathbf{L}^\infty(\Omega)$, cf. Hörmander [17, Thm. 2.2.3].

Given $0 < T_0' \le T \le \infty$, $\kappa_0, \nu_0' \in (0,\infty)$, and an integer $\ell \in [0, 2m)$, we set $\mathcal{L}^\infty \overset{\text{def}}{=} \mathbf{L}^\infty(\Gamma_T^{(T_0')}(\kappa_0,\nu_0'))$, cf. (2.5), and define $\mathcal{H}^{\ell,\infty} \equiv \mathcal{H}^{\ell,\infty}(\Gamma_T^{(T_0')}(\kappa_0,\nu_0'))$ to be the space of all bounded holomorphic functions $\mathbf{f}\colon \Gamma_T^{(T_0')}(\kappa_0,\nu_0') \to \mathbb{C}^N$ of $(z,t) = (x+iy, \sigma+i\tau) \in \mathbb{C}^D \times \mathbb{C}$ such that $\mathbf{f}$ satisfies the following two conditions (cf. Definition 2.1):

(i) $\partial^{|\alpha|}\mathbf{f}/\partial z^\alpha \in \mathbf{L}^\infty(\Gamma_T^{(T_0')}(\kappa_0,\nu_0'))$ for $|\alpha| \le \ell$.

(ii) The restriction $t \in (0,T) \mapsto \mathbf{f}(\cdot,t) \in \mathbf{L}^\infty(\mathbb{R}^D)$ is a weakly* continuous function of t with a weak*-limit $\mathbf{f}(\cdot,0) \in \mathbf{L}^\infty(\mathbb{R}^D)$ as $t \to 0+$. The function $\mathbf{f}_0 = \mathbf{f}(\cdot,0)$ is called the *initial value* of $\mathbf{f}$.

Regarding these derivatives as distributions, we observe that $\mathcal{H}^{\ell,\infty}$ is isomorphic to a closed linear subspace of the L^∞-product space $(\mathcal{L}^\infty)^{1+D+\cdots+D^\ell}$, the isomorphism being defined by $\mathbf{f} \mapsto (\partial^{|\alpha|}\mathbf{f}/\partial z^\alpha)_{|\alpha|\le\ell}$. In particular, $\mathcal{H}^{\ell,\infty}$ is a complex Banach space endowed with the norm

$$\|\mathbf{f}\|_{\mathcal{H}^{\ell,\infty}} \overset{\text{def}}{=} \max_{|\alpha|\le\ell} \sup_{\Gamma_T^{(T_0')}(\kappa_0,\nu_0')} \left|\frac{\partial^{|\alpha|}\mathbf{f}}{\partial z^\alpha}\right|. \tag{3.1}$$

Notice that $\mathcal{H}^{0,\infty}$ is a proper, closed linear subspace of $\mathbf{H}^\infty(\Gamma_T^{(T_0')}(\kappa_0,\nu_0'))$.

Proof of Theorem 2.1. In the conclusion of Theorem 2.1 we claim that $T_0' \in (0,T_0]$ is a constant independent from $\mathbf{u}$, provided $\mathbf{u}$ satisfies the bounds (2.8) for $|\alpha| \le \ell$. In other words, only L^∞-bounds on the partial derivatives $\partial^{|\alpha|}\mathbf{u}/\partial x^\alpha$ in $\mathbb{R}^D \times (0,T)$, for $|\alpha| \le \ell$, influence the value of T_0', but not $\mathbf{u}$ itself. Consequently, if we can prove the local (in time) version of Theorem 2.1, claiming that $\mathbf{u}$ admits a holomorphic continuation with the desired properties into a region (cf. (2.5))

$$\begin{aligned}\Omega_{\mathbf{u}}' = \Gamma_{T_0'}^{(T_0')}(\kappa_0,\nu_0') = \Lambda^{(T_0')}(\kappa_0,\nu_0') = \{&(x+iy, \sigma+i\tau) \in \mathbb{C}^D \times \mathbb{C}\colon \\ &(x,\sigma) \in \mathbb{R}^D \times (0,T_0'),\ |y|^{2m} < \kappa_0\sigma \text{ and } \nu_0'|\tau| < \sigma\},\end{aligned} \tag{3.2}$$

for some $T_0' \in (0,T_0]$ sufficiently small, then by uniqueness (cf. Remark 2.4) we can extend $\mathbf{u}$ from $\Lambda^{(T_0')}(\kappa_0,\nu_0')$ to all its time r translations $\Lambda_r^{(T_0')}(\kappa_0,\nu_0')$, for $0 \le r \le T - T_0'$, in order to obtain the full conclusion of Theorem 2.1. Thus, it suffices to prove the following local result, where we may take $T = T_0 \in (0,\infty)$.

THEOREM 3.1. *Let* $\mathbf{P}$ *and* $\mathbf{F}$ *satisfy Hypotheses* (H1) *and* (H2), *where* $0 < T_0 = T < \infty$ *and* $\kappa_0, \nu_0, M_\alpha \in (0,\infty)$. *Then there exist three constants* $\nu_0' \in [\nu_0,\infty)$, $C_{\mathbf{P}} \equiv C_{\mathbf{P}}(\kappa_0,\nu_0') \in (1,\infty)$, *and* $T_0' \in (0,T_0]$ *with the following property:*

Assume that $\mathbf{u}_0 \in \mathbf{L}^\infty(\mathbb{R}^D)$ *satisfies* $\partial^{|\alpha|}\mathbf{u}_0/\partial x^\alpha \in \mathbf{L}^\infty(\mathbb{R}^D)$ *for* $|\alpha| \le \ell$, *together with*

$$\operatorname*{ess\,sup}_{\mathbb{R}^D}\left|\frac{\partial^{|\alpha|}\mathbf{u}_0}{\partial x^\alpha}\right| \le M_\alpha' \overset{\text{def}}{=} \frac{M_\alpha}{C_{\mathbf{P}}} \quad \textit{for } |\alpha| \le \ell. \tag{3.3}$$

Then there exists a unique weak L^∞*-solution* $\mathbf{u}$ *of our system* (1.1) *in* $\mathbb{R}^D \times (0,T_0')$ *such that the initial condition* $\mathbf{u}(\cdot,0) = \mathbf{u}_0 \in \mathbf{L}^\infty(\mathbb{R}^D)$ *and the bounds* $\|\partial^{|\alpha|}\mathbf{u}/\partial x^\alpha\|_\infty < M_\alpha$ *for* $|\alpha| \le \ell$ *are satisfied, and moreover, the function* $\mathbf{u}$ *has a holomorphic continuation* $\mathbf{u}(x+iy, \sigma+i\tau)$ *into the region* $\Omega_{\mathbf{u}}' = \Lambda^{(T_0')}(\kappa_0,\nu_0') \subset \mathbb{C}^D \times \mathbb{C}$ *satisfying the bounds* (2.9). *Furthermore, the continuation of* $\mathbf{u}$ *satisfies* (1.1) *in* $\Omega_{\mathbf{u}}'$.

The constant $\nu_0' \in [\nu_0, \infty)$ is completely determined by $\mathbf{P}$, whereas $C_{\mathbf{P}} \in (1, \infty)$ is completely determined by $\mathbf{P}$, κ_0, and ν_0'. In particular, both ν_0' and $C_{\mathbf{P}}$ are independent from the nonlinearity $\mathbf{F}$ and M_α.

Proof of Theorem 3.1. Let $T = T_0 \in (0, \infty)$ be fixed. We divide the proof into four steps.

Step 1. We start with the following three lemmas which are concerned with the Green function $\mathbf{G}(x, x'; t) \equiv \mathbf{G}(x - x'; t) \in \mathbb{C}^{N \times N}$, for $(x, x', t) \in \mathbb{R}^D \times \mathbb{R}^D \times (0, \infty)$, corresponding to the linear initial value problem

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} + \mathbf{P}\Big(\frac{1}{i}\frac{\partial}{\partial x}\Big)\mathbf{u} = \mathbf{0} \quad &\text{for } (x, t) \in \mathbb{R}^D \times (0, T), \\ \mathbf{u}(x, 0) = \mathbf{u}_0(x) \quad &\text{for } x \in \mathbb{R}^D. \end{aligned} \tag{3.4}$$

It is well known that $\mathbf{G}(x; t)$ is given by the inverse Fourier transform

$$\mathbf{G}(x; t) = (2\pi)^{-D} \int_{\mathbb{R}^D} e^{ix\cdot\xi} e^{-t\mathbf{P}(\xi)}\, d\xi \quad \text{for } (x, t) \in \mathbb{R}^D \times (0, \infty). \tag{3.5}$$

Recall that $\mathbf{P}(\xi) = \sum_{|\alpha| \le 2m} \xi^\alpha \mathbf{P}^{(\alpha)} \in \mathbb{C}^{N \times N}$ by (1.4), where $\xi^\alpha = \xi_1^{\alpha_1} \dots \xi_D^{\alpha_D}$. An extensive treatment of the Green function is offered in Hörmander [19, Chaps. 10 and 11].

LEMMA 3.2. *There exist constants $\nu_0' \in (0, \infty)$ and $c' \in \mathbb{R}$ with the following property: Given any $\kappa_0 \in (0, \infty)$, the function $\mathbf{G}\colon \mathbb{R}^D \times (0, \infty) \to \mathbb{C}^{N \times N}$ has a holomorphic continuation $\mathbf{G}(x + iy; \sigma + i\tau)$ into the region $\Lambda^{(\infty)}(\kappa_0, \nu_0')$ such that*

$$\Big|\frac{\partial^{|\alpha|}}{\partial z^\alpha}\mathbf{G}(z; t)\Big| \le C_n(\kappa_0, \nu_0') e^{c'\sigma} \sigma^{-(D+|\alpha|)/2m} \Big(1 + \frac{|x|}{\sigma^{1/2m}}\Big)^{-n} \tag{3.6}$$

$$\textit{for all } (z, t) = (x + iy, \sigma + i\tau) \in \Lambda^{(\infty)}(\kappa_0, \nu_0') \textit{ and } |\alpha| \le \ell,$$

where $C_n \equiv C_n(\kappa_0, \nu_0') \in (0, \infty)$ is a constant for every $n = 0, 1, 2, \dots$.

Proof. Making use of the Fourier representation (3.5), we obtain the conclusion of the present lemma from Proposition A.4 in the Appendix. For system (1.3) one obtains an easy proof of Lemma 3.2 directly from (2.15). □

We set

$$C_{\mathbf{G}}' \equiv C_{\mathbf{G}}'(\kappa_0, \nu_0') \overset{\text{def}}{=} C_{D+1}(\kappa_0, \nu_0') e^{c'T_0} \int_{\mathbb{R}^D} (1 + |x|)^{-D-1}\, dx \in (0, \infty). \tag{3.7}$$

Now let $T_0' \in (0, \infty)$ be a constant to be determined later. In the situation of Lemma 3.2 we define a mapping $\mathcal{G}\colon \mathbf{L}^\infty(\mathbb{R}^D) \to \mathbf{L}^\infty(\Lambda^{(T_0')}(\kappa_0, \nu_0'))$ by

$$(\mathcal{G}\mathbf{f})(z, t) \overset{\text{def}}{=} \int_{\mathbb{R}^D} \mathbf{G}(z - x'; t)\mathbf{f}(x')\, dx' \tag{3.8}$$

$$\text{for } (z, t) \in \Lambda^{(T_0')}(\kappa_0, \nu_0') \text{ and } \mathbf{f} \in \mathbf{L}^\infty(\mathbb{R}^D).$$

LEMMA 3.3. *The mapping $\mathcal{G}$ is an everywhere-defined, bounded linear operator from $\mathbf{L}^\infty(\mathbb{R}^D)$ to $\mathcal{H}^{0,\infty}(\Lambda^{(T_0')}(\kappa_0, \nu_0'))$ satisfying*

$$\Big|\frac{\partial^{|\alpha|}}{\partial z^\alpha}(\mathcal{G}\mathbf{f})(z, t)\Big| \le C_{\mathbf{G}}'(\kappa_0, \nu_0')\sigma^{-|\alpha|/2m} \operatorname*{ess\,sup}_{\mathbb{R}^D} |\mathbf{f}| \tag{3.9}$$

$$\textit{for } (z, t) = (z, \sigma + i\tau) \in \Lambda^{(T_0')}(\kappa_0, \nu_0'),\ |\alpha| \le \ell \textit{ and } \mathbf{f} \in \mathbf{L}^\infty(\mathbb{R}^D).$$

Proof. Complex analyticity of $\mathcal{G}\mathbf{f}$ as well as the bounds (3.9) follow from Lemma 3.2. It remains to show that the restriction of $\mathcal{G}\mathbf{f}$ to $\mathbb{R}^D \times (0, T_0')$ satisfies

$$\operatorname*{weak^*\text{-}lim}_{t\to 0+}[(\mathcal{G}\mathbf{f})(\cdot,t)] = \mathbf{f} \quad \text{in } \mathbf{L}^\infty(\mathbb{R}^D),$$

or equivalently,

$$\text{(3.10)} \quad \int_{\mathbb{R}^D} \mathbf{g}(x)\cdot(\mathcal{G}\mathbf{f})(x,t)\,dx \to \int_{\mathbb{R}^D} \mathbf{g}(x)\cdot\mathbf{f}(x)\,dx \quad \text{as } t\to 0+ \quad \text{for all } \mathbf{g}\in\mathbf{L}^1(\mathbb{R}^D).$$

Since we have already established the bounds (3.9), it suffices to prove (3.10) for all $\mathbf{g}$ from a (strongly) dense set $\mathbf{S}^1 \subset \mathbf{L}^1(\mathbb{R}^D)$ and for all $\mathbf{f}$ from a weak*-dense set $\mathbf{S}^\infty \subset \mathbf{L}^\infty(\mathbb{R}^D)$. Because the Fourier transformation $\mathcal{F}$ multiplied by $(2\pi)^{-D/2}$ is a unitary operator on the Hilbert space $\mathbf{L}^2(\mathbb{R}^D) = L^2(\mathbb{R}^D \to \mathbb{C}^N)$, we choose $\mathbf{S}^1 = \mathbf{L}^1(\mathbb{R}^D)\cap\mathbf{L}^2(\mathbb{R}^D)$ and $\mathbf{S}^\infty = \mathbf{L}^\infty(\mathbb{R}^D)\cap\mathbf{L}^2(\mathbb{R}^D)$. Then, for $(\mathbf{g},\mathbf{f})\in\mathbf{S}^1\times\mathbf{S}^\infty$, we compute as follows (cf. (3.5)):

$$\begin{aligned}\int_{\mathbb{R}^D} \mathbf{g}(x)\cdot(\mathcal{G}\mathbf{f})(x,t)\,dx &= \int_{\mathbb{R}^D}\int_{\mathbb{R}^D} \mathbf{g}(x)\cdot[\mathbf{G}(x-x';t)\mathbf{f}(x')]\,dx\,dx' \\ &= (2\pi)^{-D}\int_{\mathbb{R}^D}\int_{\mathbb{R}^D}\int_{\mathbb{R}^D} e^{i(x-x')\cdot\xi}\mathbf{g}(x)\cdot[e^{-t\mathbf{P}(\xi)}\mathbf{f}(x')]\,dx\,dx'\,d\xi \\ &= \int_{\mathbb{R}^D} (\mathcal{F}^{-1}\mathbf{g})(\xi)\cdot[e^{-t\mathbf{P}(\xi)}(\mathcal{F}\mathbf{f})(\xi)]\,d\xi \xrightarrow{t\to 0+} \int_{\mathbb{R}^D}(\mathcal{F}^{-1}\mathbf{g})(\xi)\cdot(\mathcal{F}\mathbf{f})(\xi)\,d\xi \\ &= \int_{\mathbb{R}^D} \mathbf{g}(x)\cdot\mathbf{f}(x)\,dx.\end{aligned}$$

All integrals in this computation are absolutely convergent. Consequently, we have also proved $\mathcal{G}f \in \mathcal{H}^{0,\infty}(\Lambda^{(T_0')}(\kappa_0,\nu_0'))$ for every $\mathbf{f}\in\mathbf{L}^\infty(\mathbb{R}^D)$. □

Imposing an analyticity hypothesis on $\mathbf{f}$ we can extend the estimate (3.9) from Lemma 3.3 as follows.

LEMMA 3.4. *For all $s\in(0,T_0')$ and $\mathbf{f}\in\mathbf{H}^\infty(\Pi^{(s)}(\kappa_0))$, we have*

$$\text{(3.11)} \quad \left|\frac{\partial^{|\alpha|}}{\partial z^\alpha}(\mathcal{G}\mathbf{f})\Big(z,\Big(1-\frac{s}{\sigma}\Big)t\Big)\right| \le C_{\mathbf{G}}'(\kappa_0,\nu_0')\sigma^{-|\alpha|/2m}\sup_{\Pi^{(s)}(\kappa_0)}|\mathbf{f}|$$

for $(z,t)\in\Lambda^{(T_0')}(\kappa_0,\nu_0')$ with $\sigma = \Re\mathfrak{e}\, t > s$, and for $|\alpha|\le\ell$.

Proof. Fix any $s\in(0,T_0')$ and $(z,t) = (x+iy,\sigma+i\tau)\in\Lambda^{(T_0')}(\kappa_0,\nu_0')$ with $s<\sigma$. Set $\theta = s/\sigma\in(0,1)$. By (3.8), the integrand

$$\text{(3.12)} \qquad x'\in\mathbb{R}^D \mapsto \frac{\partial^{|\alpha|}}{\partial z^\alpha}\mathbf{G}(z-x';(1-\theta)t)\mathbf{f}(x')\in\mathbb{C}^N$$

of the integral $\frac{\partial^{|\alpha|}}{\partial z^\alpha}(\mathcal{G}\mathbf{f})(z,(1-\theta)t)$ has a holomorphic continuation to the region $\Pi^{(s)}(\kappa_0)$, cf. (2.1). The continuation $\mathbf{G}(z-x'-iy';(1-\theta)t)$ of $\mathbf{G}(z-x';(1-\theta)t)$ to $\Pi^{(s)}(\kappa_0)$ satisfies the bounds (3.6) with κ_0 replaced by $(1-\theta)^{-1}\kappa_0$. (Notice that the constant $C_n((1-\theta)^{-1}\kappa_0,\nu_0')$, for $n=0,1,2,\ldots$, may be unbounded as $\theta\to 1-$, and so is $C_{\mathbf{G}}'((1-\theta)^{-1}\kappa_0,\nu_0')$.) Using the decay of the continuation of the integrand

(3.12) as $|x'| \to \infty$, uniformly in $|y'|^{2m} < (1-\theta)^{-1}\kappa_0(1-\theta)\sigma = \kappa_0\sigma$, where $z' = x' + iy' \in \mathbb{C}^D$, we are able to apply Cauchy's theorem (coordinatewise) to shift the domain of integration from $\mathbb{R}^D$ to $\mathbb{R}^D + iy'$, for any fixed $y' \in \mathbb{R}^D$ with $|y'|^{2m} < \kappa_0 s$. We arrive at

(3.13)
$$\begin{aligned}\frac{\partial^{|\alpha|}}{\partial z^\alpha}(\mathcal{G}\mathbf{f})(z,(1-\theta)t) &= \int_{\mathbb{R}^D}\left[\frac{\partial^{|\alpha|}}{\partial z^\alpha}\mathbf{G}(z-x';(1-\theta)t)\right]\mathbf{f}(x')\,dx' \\ &= \int_{\mathbb{R}^D}\left[\frac{\partial^{|\alpha|}}{\partial z^\alpha}\mathbf{G}(z-x'-iy';(1-\theta)t)\right]\mathbf{f}(x'+iy')\,dx'.\end{aligned}$$

Since $(z,t) = (x+iy,\sigma+i\tau) \in \Lambda^{(T_0')}(\kappa_0,\nu_0')$ implies $|y|^{2m} < \kappa_0\sigma$, we choose $y' = \theta^{1/2m}y = (s/\sigma)^{1/2m}y$. Hence $|y'|^{2m} < \kappa_0 s$, and consequently $z' = x' + iy' \in \Pi^{(s)}(\kappa_0)$ for all $x' \in \mathbb{R}^D$. On the other hand, we have

$$z - z' = x - x' + i(y-y') = x - x' + i(1-\theta^{1/2m})y \equiv x'' + iy'',$$

where

$$|y''|^{2m} = (1-\theta^{1/2m})^{2m}|y|^{2m} < (1-\theta^{1/2m})^{2m}\kappa_0\sigma \le (1-\theta)\kappa_0\sigma = \kappa_0(\sigma - s).$$

Thus $(z-z',(1-\theta)t) \in \Lambda^{(T_0')}(\kappa_0,\nu_0')$, so we may apply the bounds (3.6) to the integrand in the second integral in (3.13) above to obtain the desired estimate (3.11). □

Step 2. Next we treat two linear operators that appear in the variation-of-constants formula for a weak L^∞-solution of system (1.1). Given $0 < T_0' \le T_0 = T < \infty$, we recall and adjust our notation

$$\mathcal{L}^\infty \overset{\text{def}}{=} \mathbf{L}^\infty(\Gamma_{T_0'}^{(T_0')}(\kappa_0,\nu_0')) = \mathbf{L}^\infty(\Lambda^{(T_0')}(\kappa_0,\nu_0'))$$

and

$$\mathcal{H}^{\ell,\infty} \equiv \mathcal{H}^{\ell,\infty}(\Gamma_{T_0'}^{(T_0')}(\kappa_0,\nu_0')) = \mathcal{H}^{\ell,\infty}(\Lambda^{(T_0')}(\kappa_0,\nu_0')).$$

Recalling (2.2), we write

(3.14) $\Sigma_{\cup}^{(T_0')}(\nu_0') = \cup\{\Sigma^{(s)}(\nu_0') \colon s \in (0,T_0')\} = \{\sigma + i\tau \in \mathbb{C} \colon \nu_0'|\tau| < \sigma < T_0'\}$.

Making use of Lemmas 3.3 and 3.4, respectively, we define the mappings $\mathcal{K}_0, \mathcal{K} \colon \mathcal{H}^{0,\infty} \to \mathcal{L}^\infty$ by

(3.15)
$$(\mathcal{K}_0\mathbf{f})(z,t) \overset{\text{def}}{=} \int_{\mathbb{R}^D}\mathbf{G}(z-x';t)\mathbf{f}(x',0)\,dx' = (\mathcal{G}\mathbf{f}(\cdot,0))(z,t)$$

and

(3.16)
$$\begin{aligned}(\mathcal{K}\mathbf{f})(z,t) &\overset{\text{def}}{=} \int_0^t\int_{\mathbb{R}^D}\mathbf{G}(z-x';t-t')\mathbf{f}(x',t')\,dx'\,dt' \\ &= \int_0^t(\mathcal{G}\mathbf{f}(\cdot,t'))(z,t-t')\,dt',\end{aligned}$$

for $(z,t) \in \Lambda^{(T_0')}(\kappa_0,\nu_0')$ and $\mathbf{f} \in \mathcal{H}^{0,\infty}$. By Cauchy's theorem, the last integral $\int_0^t \cdots dt'$ is path independent provided $t' \in \Sigma_{\cup}^{(T_0')}(\nu_0')$, cf. (3.14). We choose the path

$\{t' = \theta t \in \mathbb{C}\colon \theta \in [0,1]\}$. The integrals in (3.15) and (3.16) above converge absolutely as Lebesgue integrals by Lemmas 3.3 and 3.4.

PROPOSITION 3.5. (a) *The mapping $\mathcal{K}_0$ is an everywhere defined, bounded linear operator from $\mathcal{H}^{\ell,\infty}$ into itself satisfying*

$$\text{(3.17)}\quad \left|\frac{\partial^{|\alpha|}}{\partial z^\alpha}(\mathcal{K}_0\mathbf{f})(z,t)\right| \le C'_{\mathbf{G}} \operatorname*{ess\,sup}_{x\in\mathbb{R}^D}\left|\frac{\partial^{|\alpha|}}{\partial x^\alpha}\mathbf{f}(x,0)\right| \le C'_{\mathbf{G}} \sup_{\Lambda^{(T'_0)}(\kappa_0,\nu'_0)}\left|\frac{\partial^{|\alpha|}\mathbf{f}}{\partial z^\alpha}\right|$$

$$\textit{for } (z,t) \in \Lambda^{(T'_0)}(\kappa_0,\nu'_0),\ |\alpha| \le \ell \textit{ and } \mathbf{f} \in \mathcal{H}^{\ell,\infty}.$$

(b) *The mapping $\mathcal{K}$ is an everywhere defined, bounded, linear operator from $\mathcal{H}^{0,\infty}$ into $\mathcal{H}^{\ell,\infty}$ satisfying*

$$\text{(3.18)}\quad \left|\frac{\partial^{|\alpha|}}{\partial z^\alpha}(\mathcal{K}\mathbf{f})(z,t)\right| \le C'_{\mathbf{G}}(1+(\nu'_0)^2)^{1/2}\left(1-\frac{|\alpha|}{2m}\right)^{-1}(T'_0)^{1-(|\alpha|/2m)} \sup_{\Lambda^{(T'_0)}(\kappa_0,\nu'_0)}|\mathbf{f}|$$

$$\textit{for } (z,t) \in \Lambda^{(T'_0)}(\kappa_0,\nu'_0),\ |\alpha| \le \ell \textit{ and } \mathbf{f} \in \mathcal{H}^{0,\infty}.$$

Proof. (a) Take any $\mathbf{f} \in \mathcal{H}^{\ell,\infty}$ and $|\alpha| \le \ell$. We have $\mathcal{K}_0\mathbf{f} \in \mathcal{H}^{0,\infty}$ by Lemma 3.3. Integration by parts in the convolution integral (3.15), combined with (3.9), yields

$$\text{(3.19)}\quad \left|\frac{\partial^{|\alpha|}}{\partial z^\alpha}(\mathcal{K}_0\mathbf{f})(z,t)\right| = \left|\left(\mathcal{K}_0\frac{\partial^{|\alpha|}\mathbf{f}}{\partial z^\alpha}\right)(z,t)\right| \le C'_{\mathbf{G}} \operatorname*{ess\,sup}_{x\in\mathbb{R}^D}\left|\frac{\partial^{|\alpha|}}{\partial x^\alpha}\mathbf{f}(x,0)\right|$$

$$\text{for } (z,t) \in \Lambda^{(T'_0)}(\kappa_0,\nu'_0).$$

By Remark 2.1, the function $t \in (0,T'_0) \mapsto \partial^{|\alpha|}\mathbf{f}(\cdot,t)/\partial x^\alpha \in \mathbf{L}^\infty(\mathbb{R}^D)$ is weakly* continuous with the weak*-limit $\partial^{|\alpha|}\mathbf{f}(\cdot,0)/\partial x^\alpha \in \mathbf{L}^\infty(\mathbb{R}^D)$ as $t \to 0+$. Hence, the duality between $\mathbf{L}^1(\mathbb{R}^D)$ and $\mathbf{L}^\infty(\mathbb{R}^D)$ yields

$$\text{(3.20)}\quad \operatorname*{ess\,sup}_{x\in\mathbb{R}^D}\left|\frac{\partial^{|\alpha|}}{\partial x^\alpha}\mathbf{f}(x,0)\right| \le \liminf_{t\to 0+}\ \operatorname*{ess\,sup}_{x\in\mathbb{R}^D}\left|\frac{\partial^{|\alpha|}}{\partial x^\alpha}\mathbf{f}(x,t)\right| \le \sup_{\Lambda^{(T'_0)}(\kappa_0,\nu'_0)}\left|\frac{\partial^{|\alpha|}\mathbf{f}}{\partial z^\alpha}\right|.$$

We combine (3.19) and (3.20) to obtain the estimate (3.17).

(b) Now take any $\mathbf{f} \in \mathcal{H}^{0,\infty}$ and $|\alpha| \le \ell$. By Lemma 3.4, we have $\mathcal{K}\mathbf{f} \in \mathcal{L}^\infty$ and

$$\text{(3.21)}\quad (\mathcal{K}\mathbf{f})(z,t) = t\int_0^1\int_{\mathbb{R}^D}\mathbf{G}(z-x';(1-\theta)t)\mathbf{f}(x',\theta t)\,dx'\,d\theta$$
$$= t\int_0^1(\mathcal{G}\mathbf{f}(\cdot,\theta t))(z,(1-\theta)t)\,d\theta \quad \text{for } (z,t) \in \Lambda^{(T'_0)}(\kappa_0,\nu'_0).$$

By Lemma 3.3, the function $(z,t) \in \Lambda^{(T'_0)}(\kappa_0,\nu'_0) \mapsto (\mathcal{G}\mathbf{f}(\cdot,\theta t))(z,(1-\theta)t) \in \mathbb{C}^N$ belongs to $\mathcal{H}^{0,\infty}$ for every fixed $\theta \in (0,1)$. Consequently, applying Lemma 3.4 to (3.21) we arrive at $\partial^{|\alpha|}(\mathcal{K}\mathbf{f})/\partial z^\alpha \in \mathcal{H}^{0,\infty}$, together with (3.18), as desired. Here we have used that $|\alpha| < 2m$ and $t = \sigma + i\tau \in \Sigma_{\mathrm{U}}^{(T'_0)}(\nu'_0)$ imply $|t|\sigma^{-|\alpha|/2m} \le (1+(\nu'_0)^2)^{1/2}(T'_0)^{1-(|\alpha|/2m)}$. □

Step 3. Making use of well-known arguments from the theory of distributions, we combine Definition 2.1 with Lemma 3.2 to obtain the following variation-of-constants formula: A function $\mathbf{u}\colon \mathbb{R}^D \times (0,T_0') \to \mathbb{C}^N$ is a weak L^∞-solution of system (1.1) if and only if $\mathbf{u}$ satisfies conditions (i), (ii), and (iii) from Definition 2.1, together with the following integral equation:

$$
\begin{aligned}
(3.22)\quad \mathbf{u}(x,t) &= \int_{\mathbb{R}^D} \mathbf{G}(x-x';t)\mathbf{u}(x',0)\,dx' \\
&+ \int_0^t \int_{\mathbb{R}^D} \mathbf{G}(x-x';t-t')\mathbf{F}\Big(x',t',\Big(\frac{\partial^{|\alpha|}}{\partial x^\alpha}\mathbf{u}(x',t')\Big)_{|\alpha|\le \ell}\Big)\,dx'\,dt' \\
&\qquad\qquad \text{for almost every } (x,t) \in \mathbb{R}^D \times (0,T_0').
\end{aligned}
$$

Step 4. Now we are ready to implement our contraction mapping argument. We choose $C_{\mathbf{P}} \equiv C_{\mathbf{P}}(\kappa_0,\nu_0') \stackrel{\text{def}}{=} 1 + C_{\mathbf{G}}'(\kappa_0,\nu_0') \in (1,\infty)$. Next we fix the initial value $\mathbf{u}_0 \in \mathbf{L}^\infty(\mathbb{R}^D)$ satisfying (3.3). For $(z,t) \in \Lambda^{(T_0')}(\kappa_0,\nu_0')$ and $\mathbf{u} \in \mathcal{W}^{\ell,\infty}$ we set

$$
\mathbf{u}_1(z,t) \stackrel{\text{def}}{=} \int_{\mathbb{R}^D} \mathbf{G}(z-x';t)\mathbf{u}_0(x')\,dx'
$$

and

$$
\begin{aligned}
(3.23)\quad (\mathcal{T}(\mathbf{u}))(z,t) &\stackrel{\text{def}}{=} \mathbf{u}_1(z,t) \\
&+ \int_0^t \int_{\mathbb{R}^D} \mathbf{G}(z-x';t-t')\mathbf{F}\Big(x',t',\Big(\frac{\partial^{|\alpha|}}{\partial x^\alpha}\mathbf{u}(x',t')\Big)_{|\alpha|\le \ell}\Big)\,dx'\,dt'.
\end{aligned}
$$

Here, $\mathcal{W}^{\ell,\infty} \equiv \mathcal{W}^{\ell,\infty}(\mathbb{R}^D \times (0,T_0'))$ is the Sobolev space defined in Remark 2.4. Thus, a function $\mathbf{u} \in \mathcal{W}^{\ell,\infty}$ satisfies (3.22) if and only if $\mathcal{T}(\mathbf{u}) = \mathbf{u}$. Combining (2.7) with (3.6) we deduce that $\mathcal{T}$ is a contraction from a closed ball $\overline{\mathcal{B}} \subset \mathcal{W}^{\ell,\infty}$ into itself, provided $T_0' \in (0,T_0]$ is sufficiently small. Our more sophisticated estimates (3.17) and (3.18) from Proposition 3.5 show that $\mathcal{T}$ is also a contraction from a closed ball $\overline{\mathcal{B}} \subset \mathcal{H}^{\ell,\infty}$ into itself, provided $T_0' \in (0,T_0]$ is sufficiently small. We choose $\overline{\mathcal{B}} = \{\mathbf{f} \in \mathcal{H}^{\ell,\infty} \colon \|\mathbf{f} - \mathbf{u}_1\|_{\mathcal{H}^{\ell,\infty}} \le \delta\}$, where $0 < \delta < \min\{M_\alpha' \colon |\alpha| \le \ell\}$. Notice that by (3.9),

$$
\begin{aligned}
(3.24)\qquad \sup_{\Lambda^{(T_0')}(\kappa_0,\nu_0')} \left|\frac{\partial^{|\alpha|}\mathbf{u}_1}{\partial z^\alpha}\right| + \delta &\le C_{\mathbf{G}}' \operatorname*{ess\,sup}_{\mathbb{R}^D}\left|\frac{\partial^{|\alpha|}\mathbf{u}_0}{\partial x^\alpha}\right| + \delta \le C_{\mathbf{G}}' M_\alpha' + \delta \\
&< C_{\mathbf{G}}' M_\alpha' + M_\alpha' = C_{\mathbf{P}} M_\alpha' = M_\alpha \qquad \text{for } |\alpha| \le \ell.
\end{aligned}
$$

In order that $\mathcal{T}$ be a contraction on $\overline{\mathcal{B}}$, the constant $T_0' \in (0,T_0]$ must be chosen so small that, by (2.7) and (3.18),

$$
(3.25)\qquad K C_{\mathbf{F}} \le \delta \quad\text{and}\quad K(1 + D + \cdots + D^\ell)C_{\mathbf{F}}' < 1,
$$

where

$$
K \equiv K(T_0') = C_{\mathbf{G}}'(1 + (\nu_0')^2)^{1/2} \max_{a\in[0,\ell]}\left\{\left(1 - \frac{a}{2m}\right)^{-1}(T_0')^{1-(a/2m)}\right\}.
$$

Observe that $0 < T_0' \le 1$ and $1 - (\ell/2m) > 0$ entail

$$
K(T_0') = C_{\mathbf{G}}'(1 + (\nu_0')^2)^{1/2}\left(1 - \frac{\ell}{2m}\right)^{-1}(T_0')^{1-(\ell/2m)}.
$$

Consequently, the integral equation (3.22) possesses a unique solution $\mathbf{u} \in \overline{\mathcal{B}}$ by Banach's fixed-point theorem. The restriction of $\mathbf{u}$ to $\mathbb{R}^D \times (0, T_0')$ is the unique weak L^∞-solution of system (1.1), the uniqueness being obtained by replacing $\mathcal{H}^{\ell,\infty}$ by $\mathcal{W}^{\ell,\infty}$ above. In the latter case we may replace $\nu_0' > 0$ by $\nu_0' = 0$ in (3.25).

Our proof of Theorem 3.1 is now complete. In Step 4 of this proof we have also justified the reduction of Theorem 2.1 to its local (in time) version, Theorem 3.1.

4. Some generalizations of the main result. We discuss and suggest possible generalizations of our results to a wider class of systems (1.1) involving a much more general pseudodifferential operator as their linear part.

Remark 4.1. It is clear from our auxiliary results (cf. the Appendix) about the Green function corresponding to the linear initial value problem (3.4) that $\mathbf{P}\left(\frac{1}{i}\frac{\partial}{\partial x}\right)$can be replaced by a much more general pseudodifferential operator $\mathbf{P}\left(x, t, \frac{1}{i}\frac{\partial}{\partial x}\right)$, with a matrix-valued symbol $\mathbf{P}(x,t,\xi) \in \mathbb{C}^{N\times N}$ of class C^∞ satisfying certain analyticity hypotheses in $(x,t) \in \mathbb{R}^D \times (0,T)$, growth conditions in $\xi \in \mathbb{R}^D$, and the uniform strong ellipticity condition (in place of inequality (1.5))

$$\Re\mathfrak{e}\left(\sum_{j=1}^{N}\sum_{k=1}^{N} P_{jk}(x,t,\xi)\eta_k\eta_j^*\right) \geq (c_1|\xi|^\mu - c_2)|\eta|^2 \tag{4.1}$$

$$\text{for all } \xi \in \mathbb{R}^D \text{ and } \eta \in \mathbb{C}^N,$$

where $c_1 > 0$, $c_2 \geq 0$ and $\mu > 1$ are some constants. We refer to Trèves [37, Chap. V, §5, pp. 288–299] for details. The nonlinearity $\mathbf{F}$ may depend upon the derivatives $\partial^{|\alpha|}\mathbf{u}/\partial x^\alpha$ of $\mathbf{u}$ up to order ℓ, where $0 \leq \ell < \mu$.

For instance, we may take $\mathbf{P}(x,t,\xi)$ to be an inhomogeneous polynomial in $\xi \in \mathbb{R}^D$ of the form

$$\mathbf{P}(x,t,\xi) = \sum_{n=1}^{D} \xi_n^{2m_n}\mathbf{P}^{(n)}(x,t), \quad \xi = (\xi_1, \dots, \xi_D) \in \mathbb{R}^D,$$

where $m_n \in \mathbb{N}$ and $\mathbf{P}^{(n)}(x,t) \in \mathbb{C}^{N\times N}$ are uniformly strictly positive-definite matrices, i.e., there is a constant $c > 0$ such that, for all $(x,t) \in \mathbb{R}^D \times (0,T)$,

$$\Re\mathfrak{e}\left(\sum_{j=1}^{N}\sum_{k=1}^{N} P_{jk}^{(n)}(x,t)\eta_k\eta_j^*\right) \geq c|\eta|^2 \quad \text{for all } \eta \in \mathbb{C}^N.$$

In this case, the nonlinearity $\mathbf{F}$ may depend upon the derivatives $\partial^{|\alpha|}\mathbf{u}/\partial x^\alpha = \frac{\partial^{|\alpha|}\mathbf{u}}{\partial x_1^{\alpha_1}\dots\partial x_D^{\alpha_D}}$ of $\mathbf{u}$ such that $\alpha_n < 2m_n$ for all $n = 1, \dots, D$.

5. Applications to the validity problem. Here we discuss the validity problem for the complex Ginzburg–Landau equation (1.2).

The main result of this paper (for $2m = 2$) justifies one of the basic assumptions made in van Harten [15] and Bollerman [3]. In these papers, as in e.g. Kirrmann, Schneider, and Mielke [22] and Schneider [31, 32], the validity of the Ginzburg–Landau equation as a universal modulation equation for nonlinear stability problems at near-critical conditions has been studied. In [15] the following general equation is considered:

$$\frac{\partial u}{\partial t} = -\mu\Big(\frac{1}{i}\frac{\partial}{\partial x}; R\Big)u + 2\pi\varrho\Big(\frac{1}{i}\frac{\partial}{\partial x}; R\Big)u^2 \quad \text{for } (x,t) \in \mathbb{R} \times [0,\infty), \tag{5.1}$$
$$u(x,0) = u_0(x) \quad \text{for } x \in \mathbb{R}.$$

Here $\mu\left(\frac{1}{i}\frac{\partial}{\partial x};R\right)$ is again a strongly elliptic linear operator in x of order $2m$, and $\varrho\left(\frac{1}{i}\frac{\partial}{\partial x};R\right)$ is a linear differential operator in x of order $\ell \leq 2m-1$. Both polynomials $\mu(\cdot;R)$ and $\varrho(\cdot;R)$ depend upon a real control parameter R possessing a critical value R_c above which the trivial solution $u \equiv 0$ becomes unstable. By assuming $R = R_c + \varepsilon^2$ with $\varepsilon > 0$ small enough and

$$u(x,t) = 2\varepsilon\,\mathfrak{Re}\big(A(\varepsilon x, \varepsilon^2 t)e^{i(k_c x+\omega t)}\big) + O(\varepsilon^2),$$

it is possible to derive rigorously that the amplitude $A(X,T)$ satisfies the complex Ginzburg–Landau equation

$$(5.2)\quad \frac{\partial A}{\partial T} = (\tau_2 - i\nu_2)\frac{\partial^2 A}{\partial X^2} + (\tau_0 + i\nu_0 - \beta|A|^2)A \quad \text{for } (X,T) \in \mathbb{R}\times(0,T_0).$$

This formal approach is based on an observation of certain physical effects leading to amplitude modulations on a long time scale $T = \varepsilon^2 t$ and a long space scale moving with group velocity $X = \varepsilon(x + \varepsilon\nu_1 t)$. The constants $k_c, \omega, \tau_0, \tau_2, \nu_0, \nu_2 \in \mathbb{R}$ are completely determined by μ, whereas $\beta \in \mathbb{C}$ is completely determined by μ, ρ, k_c, and ω. The constants τ_0 and τ_2 are obtained from the Taylor expansion at $k = k_c$, the critical wavelength, of the form

$$-\mathfrak{Re}\,\mu(k, R_c + \varepsilon^2) = \varepsilon^2\left(\tau_0 - \tau_2\Big(\frac{k-k_c}{\varepsilon}\Big)^2\right) + \cdots,$$

whereas ω, ν_0, and ν_2 are obtained from

$$-\mathfrak{Im}\,\mu(k, R_c + \varepsilon^2) = \omega + \varepsilon^2\left(\nu_0 + \nu_2\Big(\frac{k-k_c}{\varepsilon}\Big)^2\right) + \cdots.$$

In accordance with the physical situation we find out that $\tau_2 > 0$ (whence $-\mathfrak{Re}\,\mu(\cdot,R)$ attains a local maximum at $k = k_c$), which guarantees the strong parabolicity of (5.2). From the point of view of physics, we are studying a marginally unstable basic state; cf. Doelman [6]. In [15] a rigorous proof of the validity of this approximation has been given.

The main result in van Harten [15] can be stated as follows.

THEOREM 5.1. *Let $S_a = \{z = x + iy \in \mathbb{C}\colon |y| < a\}$ for some constant $a \in (0,\infty)$, and $R \equiv R(\varepsilon) = R_c + \varepsilon^2$ for $\varepsilon \in (0,\infty)$.*

(a) *There exists $T_0 > 0$ such that, for any initial value $A(\cdot,0) = A_0\colon \mathbb{R} \to \mathbb{C}$ satisfying*

(i$_0$) *A_0 is holomorphic in S_a and*

(ii$_0$) *$\sup_{Z\in S_a}|A_0(Z)| < \infty$,*

there exists a unique classical solution A of (5.2) *in $\mathbb{R}\times[0,T_0]$ such that*

(i) *$A(\cdot,T)$ is holomorphic in S_a for each $T \in [0,T_0]$, $A\colon S_a\times[0,T_0] \to \mathbb{C}$ is continuous, and*

(ii) *$\sup_{(Z,T)\in S_a\times[0,T_0]}|A(Z,T)| < \infty$.*

(b) *Let A be a solution of* (5.2) *as in* (a), *and for each $\varepsilon > 0$ set*

$$\psi_\varepsilon(z,t) = 2\varepsilon\,\mathfrak{Re}\left(A(\varepsilon z, \varepsilon^2 t)e^{i(k_c z+\omega t)}\right) \quad \textit{for } (z,t) \in S_{a/\varepsilon}\times[0, T_0/\varepsilon^2].$$

Then $\psi_\varepsilon(x+iy,t) \equiv \psi_\varepsilon(x,y,t)$ is a real-valued, real analytic function of $(x,y) \in S_{a/\varepsilon}$, for each $t \in [0,T_0/\varepsilon^2]$, and for each $\varepsilon > 0$ small enough, there exists a solution $u \equiv u_\varepsilon$ of (5.1) *in $\mathbb{R}\times[0,T_0/\varepsilon^2]$ with the following properties:*

(i$_\varepsilon$) *$u_\varepsilon(\cdot,t)$ ($t \in [0,T_0/\varepsilon^2]$) is real analytic in $S_{a'/\varepsilon}$ for $a' \in (0,a)$, $u_\varepsilon\colon S_{a'/\varepsilon} \times [0,T_0/\varepsilon^2] \to \mathbb{R}$ is continuous and bounded, and*

(ii$_\varepsilon$) *there is a constant $C > 0$ independent from ε such that*

$$\sup_{(z,t)\in S_{a'/\varepsilon}\times[0,T_0/\varepsilon^2]} |u_\varepsilon(z,t) - \psi_\varepsilon(z,t)| \le C\varepsilon^2.$$

The proof in [15] is based on an application of a contraction-mapping argument which proves the existence of the remainder term $\varepsilon^2 w_\varepsilon = u_\varepsilon - \psi_\varepsilon$ of order $O(\varepsilon^2)$ by subtracting the appropriate rescaled form of (1.2) from (5.1). The application of the contraction-mapping argument hinges upon making the correct estimates for the operators $\mathcal{K}$ and $\mathcal{K}_\varrho$, where

$$(\mathcal{K}f)(z,t) = \frac{1}{2\pi}\int_0^t \int_{\mathbb{R}} e^{iz\xi} e^{-(t-t')\mu(\xi)} \int_{\mathbb{R}} e^{-ix'\xi} f(x',t')\,dx'\,d\xi\,dt',$$

$$(\mathcal{K}_\varrho f)(z,t) = \frac{1}{2\pi}\int_0^t \int_{\mathbb{R}} e^{iz\xi} e^{-(t-t')\mu(\xi)} \varrho(\xi) \int_{\mathbb{R}} e^{-ix'\xi} f(x',t')\,dx'\,d\xi\,dt'.$$

There are two fundamental assumptions in this theorem.

Assumption 1. The first one is the assumption that A_0, the initial value of the solution A of (5.2), is holomorphic on a strip S_a. Due to the results of the present paper, this analyticity assumption can be weakened significantly, while the conclusion of the theorem can be improved in the following way.

THEOREM 5.2. (a) *There exist positive constants κ_0, ν_0', T_0', and T_0 with $0 < T_0' \le T_0$ and the following property: Given any initial value $A_0 \in L^\infty(\mathbb{R})$, there exists a unique weak L^∞-solution A of* (5.2) *in $\mathbb{R} \times (0,T_0)$ satisfying the initial condition $A(\cdot,0) = A_0$ and such that* $\operatorname{ess\,sup}_{(X,T)\in\mathbb{R}\times(0,T_0)} |A(X,T)| < \infty$. *Furthermore, A has a holomorphic continuation $A(X+iY, S+iT)$ into the region $\Omega_A' = \Gamma_{T_0}^{(T_0')}(\kappa_0,\nu_0';2) \subset \mathbb{C}\times\mathbb{C}$ for which $\sup_{\Omega_A'} |A| < \infty$ is satisfied.*

(b) *Let κ_0, ν_0', T_0', and T_0 be as in* (a), *let A be a solution of* (5.2), *and let*

$$\psi_\varepsilon(z,t) = 2\varepsilon\,\Re\mathfrak{e}\left(A(\varepsilon z,\varepsilon^2 t)e^{i(k_c z+\omega t)}\right)$$

be defined in $\Omega_{\psi_\varepsilon}' = \Gamma_{T_0/\varepsilon^2}^{(T_0'/\varepsilon^2)}(\kappa_0,\nu_0';2)$ for each $\varepsilon > 0$. Assume that

(i) *the initial distribution $A(\cdot,0) = A_0\colon \mathbb{R} \to \mathbb{C}$ has uniformly Lipschitz continuous derivatives up to order $\ell - 1$, and*

(ii) $\operatorname{ess\,sup}_{(X,T)\in\mathbb{R}\times(0,T_0)} |A(X,T)| < \infty$.

Then there exist positive constants $\tilde\kappa_0$, $\tilde\nu_0'$, and $\tilde T_0'$ with $0 < \tilde T_0' \le T_0$ and the following property: For each ε small enough, $0 < \varepsilon \le 1$, there exist a solution $u \equiv u_\varepsilon$ of (5.1) *holomorphic in $\Omega_{u_\varepsilon}' = \Gamma_{T_0/\varepsilon^2}^{(\tilde T_0')}(\tilde\kappa_0,\tilde\nu_0';2m)$ and a constant $C > 0$ independent from ε such that*

$$\sup_{(z,t)\in\Omega_{u_\varepsilon}'\cap\Omega_{\psi_\varepsilon}'} |u_\varepsilon(z,t) - \psi_\varepsilon(z,t)| \le C\varepsilon^2. \tag{5.3}$$

Since the (5.1) and (5.2) have possibly different orders $2m$ and 2, respectively, we have introduced the notation $\Gamma_T^{(s)}(\kappa_0,\nu_0;2m) \equiv \Gamma_T^{(s)}(\kappa_0,\nu_0)$ for $0 \le s \le T \le \infty$; cf. (2.5).

The following few remarks explain some statements made in Theorem 5.2.

Remark 5.1. Conditions (i) and (ii) from Theorem 5.2(b) imply that there exist constants C_n, $0 \leq n \leq \ell$, such that

$$\sup_{\Omega'_A} \left| \frac{\partial^n A}{\partial Z^n} \right| \leq C_n \quad \text{for } n = 0, 1, \ldots, \ell. \tag{5.4}$$

Indeed, assume that the initial distribution $A(\cdot, 0)$ has essentially bounded derivatives up to order ℓ, and A is essentially bounded in $\mathbb{R} \times (0, T_0)$. Define $A_n \stackrel{\text{def}}{=} \frac{\partial^n A}{\partial X^n}$, $n \leq \ell$. Differentiating (5.2) n times with respect to X, $1 \leq n \leq \ell$, we arrive at the following linear system of equations for $\mathfrak{Re}\, A_n$ and $\mathfrak{Im}\, A_n$:

$$\begin{aligned}
&\frac{\partial A_n}{\partial T} = (\tau_2 - i\nu_2) \frac{\partial^2 A_n}{\partial X^2} + (\tau_0 + i\nu_0) A_n \\
&- \beta \left[N_n \left((A_i)_{i \leq n-1}, (A_i^*)_{i \leq n-1} \right) A_n + N_n' \left((A_i)_{i \leq n-1}, (A_i^*)_{i \leq n-1} \right) A_n^* \right] && \text{in } \mathbb{R} \times (0, T_0), \\
&A_n(\cdot, 0) = \frac{\partial^n A}{\partial X^n}(\cdot, 0) && \text{in } \mathbb{R}.
\end{aligned}$$

Here, both $N_n \left((A_i)_{i \leq n-1}, (A_i^*)_{i \leq n-1} \right)$ and $N_n' \left((A_i)_{i \leq n-1}, (A_i^*)_{i \leq n-1} \right)$ are polynomials in all their variables. Hence, we use induction on $n = 1, \ldots, \ell$ to obtain

$$\operatorname*{ess\,sup}_{(X,T) \in \mathbb{R} \times (0, T_0)} \left| \frac{\partial^n A}{\partial X^n}(X, T) \right| \leq C_n' = \text{const} < \infty \quad \text{for } n = 0, 1, \ldots, \ell. \tag{5.5}$$

Finally, we can apply our Theorem 2.1 to conclude (5.4), as desired.

Remark 5.2. In Theorem 5.2(a) above we are allowed to take $T_0 = \infty$ for the spatial dimension $D \leq 2$, provided $\mathfrak{Re}\, \beta > 0$ holds, by the global existence results in Bartuccelli et al. [2]. However, we can prove the estimate (5.3) for $0 < T_0 < \infty$ only.

Remark 5.3. Observe that the intersection $\Omega'_{u_\varepsilon} \cap \Omega'_{\psi_\varepsilon}$ contains the region

$$\Omega' = \Gamma_{T_0/\varepsilon^2}^{(\bar{T}_0')}(\bar{\kappa}_0, \bar{\nu}_0'; 2), \quad 0 < \varepsilon \leq 1,$$

where $\bar{\kappa}_0 = \min\{\kappa_0, \tilde{\kappa}_0\}$, $\bar{\nu}_0' = \max\{\nu_0', \tilde{\nu}_0'\}$ and $\bar{T}_0' = \min\{T_0', \tilde{T}_0', 1/\tilde{\kappa}_0\}$.

In [22] and [31, 32], it is shown that it is possible to prove the validity of (5.2) for initial values A_0 that are not necessarily holomorphic, but are sufficiently smooth. However, in these papers it was necessary to impose a certain rate of decay as $|x| \to \infty$ on the initial values $A_0(x)$. As a consequence it turned out to be impossible to prove validity for such interesting basic cases as periodic and front solutions of (5.2) studied, e.g. in Doelman [6], Doelman and Titi [7], and Takáč [35].

Assumption 2. The second fundamental assumption in van Harten's theorem [15] (Theorem 5.1 above) is on the structure of the initial values u_0 for (5.1),

$$u_0(x) = 2\varepsilon\, \mathfrak{Re}\big(A_0(\varepsilon x) e^{i k_c x}\big) + O(\varepsilon^2). \tag{5.6}$$

In Eckhaus [10] it has been shown that, on a time scale faster than the $1/\varepsilon^2$ Ginzburg–Landau time scale, any solution of (5.1) with L^∞ initial values small enough (like $O(\varepsilon)$) first collapses toward a solution having the special structure of (5.6) (clustered mode distribution) before this solution starts to evolve on the Ginzburg–Landau time scale. In [10] this phenomenon is called "the attractivity of the Ginzburg–Landau manifold."

It should be possible to obtain the same results by applying the main result of this paper. From the estimates (3.6) it is clear that in the linearized case the only values of $\xi \in \mathbb{R}$ for which the Fourier transform $\hat{u}(\xi, t) \equiv (\mathcal{F}u)(\xi, t)$ has any chance of retaining its order of magnitude on such time scales,

$$0 < t < \tilde{T}/\varepsilon^{\nu} \quad \text{for some } \nu \in (0, 2) \text{ and } \tilde{T} = O(1),$$

are those in the $O(\varepsilon)$-neighborhood of $\pm k_c$ in which $-\,\mathfrak{Re}\,\mu(\xi)$ assumes positive values, while for the other values, $\hat{u}(\xi, t)$ seems to become exponentially small. Adding the nonlinearity to the equation will only result in preventing $\hat{u}(\xi, t)$ from descending to that order of magnitude in $O(\varepsilon)$-neighborhoods of multiples of k_c. For the remaining values of ξ we find that $\hat{u}(\xi, t)$ becomes exponentially small, which is a slightly stronger result than the one obtained in [10].

6. Discussion. We conclude our article with a brief discussion about further applications of our results to some current issues in mathematical physics and dynamical systems.

Our main result simplifies the proof of the validity of the time-dependent amplitude approximation for the Swift–Hohenberg equation given in Collet and J.-P. Eckmann [4]. Analyticity of spatially periodic solutions to the Kuramoto–Sivashinsky equation was studied in Collet et al. [5, Thm. 1.2] and Liu [25]. Our results cover this case also.

An important model for convecting fluids between poorly conducting slabs was derived by Proctor [30, Eq. (4.15)]. Our analyticity result applies to the simplified version of the Proctor equation studied in Hoyle [20].

Our results apply also to the generalized Ginzburg–Landau equation studied in Duan and Holmes [8].

A stronger version of analyticity in the space variable is obtained by considering the Gevrey classes of functions (cf. Lions and Magenes [23, 24]). For the Ginzburg–Landau equation, a priori Gevrey-class regularity of solutions on bounded domains was shown in Doelman and Titi [7] and in Duan, Holmes, and Titi [9] for a generalized Ginzburg–Landau equation. For the Navier-Stokes equations, it was proved by Foias and Temam [12] (see also Henshaw, Kreiss, and Reyna [16]). For the Euler equations, i.e., the Navier–Stokes equations with vanishing kinematic viscosity (no diffusion), the domain of holomorphy (in the space variable) of solutions on bounded domains was investigated in Bardos and Benachour [1].

Besides the analyticity results of Friedman [13] for solutions of nonlinear parabolic systems mentioned in §1 above, a number of analyticity results for solutions of linear parabolic systems can be found in Friedman [14, Chap. 3, §3, pp. 212–216], Hörmander [18, Chap. 9, §4, pp. 346–353], Lions and Magenes [23, 24], and Tanabe [36].

The analyticity of solutions to analytic, strongly elliptic systems of both linear and quasilinear partial differential equations in a bounded open domain, with analytic boundary and assigned analytic Dirichlet boundary data, is treated in the monograph by Morrey [28, Chap. 6, §§6 and 7, pp. 258–277]. Shorter alternate proofs of some of these results are given in Masuda [27].

Our last remark concerns possible extensions of our main result, Theorem 2.1, to

(a) an arbitrary bounded open domain $\Omega \subset \mathbb{R}^D$ with analytic boundary $\partial\Omega$ and assigned analytic Dirichlet boundary data, and to

(b) an arbitrary (unbounded) cylindrical domain $\Omega = \mathbb{R}^{D_1} \times \Omega_2$, where $\Omega_2 \subset \mathbb{R}^{D_2}$ is a bounded open domain with analytic boundary $\partial\Omega_2$, $D = D_1 + D_2$, and the assigned Dirichlet boundary data on $\mathbb{R}^{D_1} \times \partial\Omega_2$ are analytic.

The latter case is of interest in the validity problem for the Ginzburg–Landau equation (5.2); cf. Bollerman [3].

It is not difficult to see from our proof of Theorem 2.1 (and its local-in-time version, Theorem 3.1) that all we need in this proof are the analyticity and decay-at-infinity properties of the Green's function $\mathbf{G}(z, z'; t) \in \mathbb{C}^{N\times N}$ corresponding to the linear initial value problem (3.4). However, we can apply the analyticity results of Friedman [14, Chap. 3, §3, pp. 212–216] to obtain all desired properties of the Green's function $\mathbf{G}(z, z'; t)$ in any domain $\Omega \subset \mathbb{R}^D$ of type (a) or (b) specified above, in place of $\mathbb{R}^D$.

Moreover, knowing the analyticity results for solutions to analytic, strongly elliptic systems in a bounded open domain, cf. Morrey [28, Chap. 6, §§6 and 7, pp. 258–277], we may restrict ourselves to Dirichlet (zero) boundary data. Consequently, our main trick in the proof of Lemma 3.4, shifting the domain of integration from $\mathbb{R}^D$ to $\mathbb{R}^D + iy'$, for any fixed $y' \in \mathbb{R}^D$ with $|y'|^{2m} < \kappa_0 s$, can be used in any domain $\Omega \subset \mathbb{R}^D$ of type (a) or (b). Therefore *we conjecture* that our main result, Theorem 2.1, is valid also in any such domain Ω. We leave the details to the reader.

Appendix. The Green function. We establish some standard results about the inverse Fourier transform

$$\text{(A.1)} \qquad \mathbf{G}(x;t) \overset{\text{def}}{=} (2\pi)^{-D} \int_{\mathbb{R}^D} e^{ix\cdot\xi} e^{-t\mathbf{P}(\xi)}\, d\xi \quad \text{for } (x,t) \in \mathbb{R}^D \times (0,\infty).$$

The reader is referred to Stein and Weiss [34] regarding the Fourier transformation. Recall that $\mathbf{P}(\xi) = \sum_{|\alpha|\le 2m} \xi^\alpha \mathbf{P}^{(\alpha)} \in \mathbb{C}^{N\times N}$ is a polynomial in $\xi = (\xi_1, \dots, \xi_D) \in \mathbb{R}^D$ by (1.4), where $\xi^\alpha = \xi_1^{\alpha_1} \dots \xi_D^{\alpha_D}$ for $\alpha = (\alpha_1, \dots, \alpha_D) \in (\mathbb{Z}_+)^D$. Each $\mathbf{P}^{(\alpha)} = (P_{jk}^{(\alpha)})_{j,k=1}^N$ is an $N \times N$ real (or complex) matrix. We assume that $\mathbf{P}$ is strongly elliptic, i.e., inequality (1.5) is valid. Consequently, we can find a constant $\nu_0' \in (0,\infty)$ so large that there exist two additional constants $c_1 > 0$ and $c_2 \ge 0$ such that the following inequality holds for all $1 + i\tau \in \Sigma^{(1)}(\nu_0')$, $\xi = (\xi_1, \dots, \xi_D) \in \mathbb{R}^D$ and $\boldsymbol{\eta} = (\eta_1, \dots, \eta_N) \in \mathbb{C}^N$:

$$\text{(A.2)} \qquad \mathfrak{Re}\left((1+i\tau)\sum_{j=1}^N \sum_{k=1}^N \sum_{|\alpha|\le 2m} P_{jk}^{(\alpha)} \xi^\alpha \eta_k \eta_j^*\right) \ge (c_1|\xi|^{2m} - c_2)|\boldsymbol{\eta}|^2.$$

Here $\Sigma^{(s)}(\nu_0') = \{t = \sigma + i\tau \in \mathbb{C} : \nu_0'|\tau| < \sigma = s\}$ for $0 < s < \infty$, by (2.2), and $|\boldsymbol{\eta}| = \left(\sum_{j=1}^N |\eta_j|^2\right)^{1/2}$ is the Euclidean norm in $\mathbb{C}^N$. We set

$$\Sigma_\cup^{(\infty)}(\nu_0') = \cup\{\Sigma^{(s)}(\nu_0') : s \in (0,\infty)\} = \{t = \sigma + i\tau \in \mathbb{C} : \nu_0'|\tau| < \sigma\}.$$

We will investigate the holomorphic continuation of $\mathbf{G}(x;t)$ into the region

$$\Lambda^{(\infty)}(\kappa_0, \nu_0') = \{(x+iy, \sigma+i\tau) \in \mathbb{C}^D \times \mathbb{C} : \\ (x,\sigma) \in \mathbb{R}^D \times (0,\infty),\ |y|^{2m} < \kappa_0\sigma \text{ and } \nu_0'|\tau| < \sigma\},$$

where $\kappa_0 \in (0,\infty)$ is an arbitrary constant, cf. (2.5).

As usual, we denote by $\mathbf{I} = (\delta_{jk})_{j,k=1}^N$ the $N \times N$ identity matrix. We define the ℓ^2-operator norm of a matrix $\mathbf{M} \in \mathbb{C}^{N\times N}$ by

$$\|\mathbf{M}\| \overset{\text{def}}{=} \sup_{|\mathbf{f}|\le 1} |\mathbf{M}\mathbf{f}| \quad \text{for } \mathbf{f} \in \mathbb{C}^N.$$

We need the following lemma in order to be able to estimate the norm of $e^{-t\mathbf{P}(\xi)}$ in (A.1).

LEMMA A.1. *For all $\xi \in \mathbb{R}^D$, and for all $t, \lambda \in \mathbb{C}$ satisfying $\nu_0' |\Im\mathfrak{m}\, t| < \Re\mathfrak{e}\, t$ and $\Re\mathfrak{e}\,\lambda > c_2 \Re\mathfrak{e}\, t$, the matrix $\lambda\mathbf{I} + t\mathbf{P}(\xi) \in \mathbb{C}^{N\times N}$ is invertible with the norm*

$$\|(\lambda\mathbf{I} + t\mathbf{P}(\xi))^{-1}\| \le (\Re\mathfrak{e}\,\lambda + (c_1|\xi|^{2m} - c_2)\Re\mathfrak{e}\, t)^{-1}. \tag{A.3}$$

Proof. Let $1+i\tau \in \Sigma^{(1)}(\nu_0')$, $\lambda' \in \mathbb{C}$ and $\xi = (\xi_1, \dots, \xi_D) \in \mathbb{R}^D$ be fixed. Applying the Cauchy–Schwartz inequality to (A.2) we arrive at

$$\begin{aligned}(c_1|\xi|^{2m} + \Re\mathfrak{e}\,\lambda' - c_2)|\boldsymbol{\eta}|^2 &\le \Re\mathfrak{e}\left(\sum_{j=1}^N\sum_{k=1}^N\left(\lambda'\delta_{jk} + (1+i\tau)\sum_{|\alpha|\le 2m} P_{jk}^{(\alpha)}\xi^\alpha\right)\eta_k\eta_j^*\right)\\ &\le \left(\sum_{j=1}^N\left|\sum_{k=1}^N\left(\lambda'\delta_{jk} + (1+i\tau)\sum_{|\alpha|\le 2m} P_{jk}^{(\alpha)}\xi^\alpha\right)\eta_k\right|^2\right)^{1/2}\left(\sum_{j=1}^N|\eta_j|^2\right)^{1/2}\\ &= |(\lambda'\mathbf{I} + (1+i\tau)\mathbf{P}(\xi))\boldsymbol{\eta}|\,|\boldsymbol{\eta}| \quad \text{for all } \boldsymbol{\eta} = (\eta_1, \dots, \eta_N) \in \mathbb{C}^N.\end{aligned}$$

In particular, taking $\tau = (\Im\mathfrak{m}\, t)/(\Re\mathfrak{e}\, t)$ and $\lambda' = \lambda/(\Re\mathfrak{e}\, t)$ we obtain $\Re\mathfrak{e}\,\lambda' > c_2$ together with (A.3). □

Now we can estimate the norm of $e^{-t\mathbf{P}(\xi)}$ in $(A.1)$.

LEMMA A.2. *For all $\xi \in \mathbb{R}^D$ and $t \in \mathbb{C}$ satisfying $\nu_0'|\Im\mathfrak{m}\, t| < \Re\mathfrak{e}\, t$, we have*

$$\|e^{-t\mathbf{P}(\xi)}\| \le e^{(c_2 - c_1|\xi|^{2m})\Re\mathfrak{e}\, t}. \tag{A.4}$$

Proof. Let us fix any $\xi \in \mathbb{R}^D$ and $t \in \mathbb{C}$ such that $\nu_0'|\Im\mathfrak{m}\, t| < \Re\mathfrak{e}\, t$. By $(A.3)$ in Lemma A.1, for every $\lambda \in \mathbb{C}$ satisfying $\Re\mathfrak{e}\,\lambda > c_2 \Re\mathfrak{e}\, t$, we have

$$\|(\lambda\mathbf{I} + t\mathbf{P}(\xi))^{-n}\| \le (\Re\mathfrak{e}\,\lambda + (c_1|\xi|^{2m} - c_2)\Re\mathfrak{e}\, t)^{-n}, \quad n = 1, 2, 3, \dots.$$

Taking $\lambda = n$ for $n > c_2 \Re\mathfrak{e}\, t$ we arrive at

$$\|(\mathbf{I} + (t/n)\mathbf{P}(\xi))^{-n}\| \le (1 + (c_1|\xi|^{2m} - c_2)(\Re\mathfrak{e}\, t/n))^{-n}.$$

Finally, we let $n \to \infty$ in order to obtain (A.4) for the limits

$$(\mathbf{I} + (t/n)\mathbf{P}(\xi))^{-n} \to e^{-t\mathbf{P}(\xi)}$$

and

$$(1 + (c_1|\xi|^{2m} - c_2)\Re\mathfrak{e}(t/n))^{-n} \to e^{(c_2 - c_1|\xi|^{2m})\Re\mathfrak{e}\, t}. \qquad \square$$

Remark A.1. Making use of Lemma A.2, for any fixed $t \in \mathbb{C}$ with $\nu_0'|\Im\mathfrak{m}\, t| < \Re\mathfrak{e}\, t$, we easily deduce that the function $\xi \in \mathbb{R}^D \mapsto e^{-t\mathbf{P}(\xi)} \in \mathbb{C}^{N\times N}$ is in the Schwartz space $\mathcal{S}(\mathbb{R}^D \to \mathbb{C}^{N\times N})$ and its inverse Fourier transform $\mathbf{G}(x,t)$ has a holomorphic continuation to the entire complex space $\mathbb{C}^D$. Recall that $\mathcal{S}(\mathbb{R}^D \to \mathbb{C}^{N\times N})$ denotes the Schwartz space of all functions $M\colon \mathbb{R}^D \to \mathbb{C}^{N\times N}$ that are infinitely many times continuously differentiable with all partial derivatives having faster than polynomial decay at intinity. Also $\mathcal{F}^{\pm 1}(\mathcal{S}(\mathbb{R}^D \to \mathbb{C}^{N\times N})) = \mathcal{S}(\mathbb{R}^D \to \mathbb{C}^{N\times N})$, cf. Edwards [11, §5.15.1, p. 375].

In order to establish certain upper bounds on the norm of the holomorphic continuation of the Green function $\mathbf{G}$ and all its partial derivatives $\partial^{|\alpha|}\mathbf{G}/\partial x^\alpha$ to the region $\Lambda^{(\infty)}(\kappa_0, \nu_0')$, we need the following estimate.

LEMMA A.3. *Given any multiindex $\beta \in (\mathbb{Z}_+)^D$, there exists a constant $C' \equiv C'(\beta) \in (0,\infty)$ such that*

$$\text{(A.5)} \quad \left\| \frac{\partial^{|\beta|}}{\partial(\xi')^\beta} e^{-s(1+i\tau)\mathbf{P}(s^{-1/2m}\xi')} \right\| \le C'(\beta)(1+s^{1/2m}+|\xi'|)^{(2m-1)|\beta|} e^{c_2 s - c_1|\xi'|^{2m}}$$

for all $\xi' \in \mathbb{R}^D$, $s \in (0,\infty)$ and $1+i\tau \in \mathbb{C}$ satisfying $\nu_0'|\tau| < 1$.

Proof. By induction on $|\beta|$ we derive the formula

$$\text{(A.6)} \quad \frac{\partial^{|\beta|}}{\partial(\xi')^\beta} e^{-s(1+i\tau)\mathbf{P}(s^{-1/2m}\xi')} = e^{-s(1+i\tau)\mathbf{P}(s^{-1/2m}\xi')} \sum_{n=0}^{(2m-1)|\beta|} s^{n/2m}\mathbf{Q}^{(n)}(\xi'),$$

where each

$$\mathbf{Q}^{(n)}(\xi') = \sum_{|\alpha| \le (2m-1)|\beta|-n} (\xi')^\alpha \mathbf{Q}^{(n,\alpha)}, \quad n = 0,1,\dots,(2m-1)|\beta|$$

is a polynomial of order $\le (2m-1)|\beta| - n$ in $\xi' \in \mathbb{R}^D$ with the coefficients $\mathbf{Q}^{(n,\alpha)} \in \mathbb{C}^{N\times N}$. Consequently, in order to obtain (A.5), we estimate the sum in (A.6) as follows, for all $\xi' \in \mathbb{R}^D$ and $s \in (0,\infty)$:

$$\left\| \sum_{n=0}^{(2m-1)|\beta|} s^{n/2m}\mathbf{Q}^{(n)}(\xi') \right\| \le C'(\beta)(1+s^{1/2m}+|\xi'|)^{(2m-1)|\beta|}.$$

The exponential in (A.6) can be estimated by (A.4) with $\xi = s^{-1/2m}\xi'$ and $t = s(1+i\tau)$. □

From (A.1) combined with Lemma A.2, we deduce that the formula

$$\text{(A.7)} \quad \mathbf{G}(z;t) = (2\pi)^{-D} \int_{\mathbb{R}^D} e^{iz\cdot\xi} e^{-t\mathbf{P}(\xi)}\, d\xi,$$

$$\text{for } (z,t) = (x+iy, \sigma+i\tau) \in \Lambda^{(\infty)}(\kappa_0, \nu_0')$$

defines the holomorphic continuation $\mathbf{G}(x+iy,\sigma+i\tau)$ of the Green function $\mathbf{G}$ into the region $\Lambda^{(\infty)}(\kappa_0,\nu_0')$. Notice that in (A.7) we have

$$\mathfrak{Re}(iz\cdot\xi) = -y\cdot\xi \le \kappa_0\sigma^{1/2m}|\xi|.$$

Next we set

$$\Lambda'(\kappa_0,\nu_0') = \Pi^{(1)}(\kappa_0) \times \Sigma_U^{(\infty)}(\nu_0') = \{(x'+iy', \sigma+i\tau) \in \mathbb{C}^D \times \mathbb{C}:$$
$$(x',\sigma) \in \mathbb{R}^D \times (0,\infty),\ |y'|^{2m} < \kappa_0 \text{ and } \nu_0'|\tau| < \sigma\}.$$

In order to estimate the norms of the functions $z^\beta(\partial^{|\alpha|}\mathbf{G}/\partial z^\alpha)$, for $\alpha,\beta \in (\mathbb{Z}_+)^D$, we make the substitution

$$\text{(A.8)} \quad \mathbf{G}(z;t) = \sigma^{-D/2m}\mathbf{G}'(\sigma^{-1/2m}z;t)$$

for $(z,t) = (x+iy,\sigma+i\tau) \in \Lambda^{(\infty)}(\kappa_0,\nu_0')$, where

$$\text{(A.9)} \quad \mathbf{G}'(z';t) \overset{\text{def}}{=} (2\pi)^{-D} \int_{\mathbb{R}^D} e^{iz'\cdot\xi'} e^{-t\mathbf{P}(\sigma^{-1/2m}\xi')}\, d\xi'$$

$$\text{for } (z',t) = (x'+iy', \sigma+i\tau) \in \Lambda'(\kappa_0,\nu_0').$$

Combining integration by parts with Lemma A.3, we arrive at

$$(z')^{\beta}\frac{\partial^{|\alpha|}}{\partial(z')^{\alpha}}\mathbf{G}'(z';t)=(2\pi)^{-D}i^{|\alpha|+|\beta|}\int_{\mathbb{R}^D}e^{iz'\cdot\xi'}(\xi')^{\alpha}\frac{\partial^{|\beta|}}{\partial(\xi')^{\beta}}e^{-t\mathbf{P}(\sigma^{-1/2m}\xi')}\,d\xi'.$$

Taking $(z',t)=(x'+iy',\sigma+i\tau)\in\Lambda'(\kappa_0,\nu_0')$ we obtain, by Lemma A.3,

$$\begin{aligned}
&|(z')^{\beta}|\left\|\frac{\partial^{|\alpha|}}{\partial(z')^{\alpha}}\mathbf{G}'(z';t)\right\|\\
&\le(2\pi)^{-D}\int_{\mathbb{R}^D}e^{-y'\cdot\xi'}|(\xi')^{\alpha}|\left\|\frac{\partial^{|\beta|}}{\partial(\xi')^{\beta}}e^{-t\mathbf{P}(\sigma^{-1/2m}\xi')}\right\|d\xi'\\
&\le(2\pi)^{-D}C'(\beta)\int_{\mathbb{R}^D}e^{\kappa_0|\xi'|}|\xi'|^{|\alpha|}(1+\sigma^{1/2m}+|\xi'|)^{(2m-1)|\beta|}e^{c_2\sigma-c_1|\xi'|^{2m}}\,d\xi'\\
&\le(2\pi)^{-D}C'(\beta)\int_{\mathbb{R}^D}(1+\sigma^{1/2m}+|\xi'|)^{|\alpha|+(2m-1)|\beta|}e^{c_2\sigma-c_1|\xi'|^{2m}+\kappa_0|\xi'|}\,d\xi'.
\end{aligned}$$

Consequently, given any $\kappa_0\in(0,\infty)$, $c'\in(c_2,\infty)$, and $\alpha,\beta\in(\mathbb{Z}_+)^D$, there exists a constant $C'(\kappa_0,c',|\alpha|,|\beta|)\in(0,\infty)$ such that

$$\text{(A.10)}\quad |(z')^{\beta}|\left\|\frac{\partial^{|\alpha|}}{\partial(z')^{\alpha}}\mathbf{G}'(z';t)\right\|\le C'(\kappa_0,c',|\alpha|,|\beta|)e^{c'\sigma}$$
$$\text{for } (z',t)=(x'+iy',\sigma+i\tau)\in\Lambda'(\kappa_0,\nu_0').$$

Furthermore, by (A.8) we have

$$\text{(A.11)}\quad (\sigma^{-1/2m}z)^{\beta}\frac{\partial^{|\alpha|}}{\partial z^{\alpha}}\mathbf{G}(z;t)=\sigma^{-(D+|\alpha|)/2m}(z')^{\beta}\frac{\partial^{|\alpha|}}{\partial(z')^{\alpha}}\mathbf{G}'(z';t),$$
$$\text{where } z'=\sigma^{-1/2m}z \text{ for } (z,t)=(x+iy,\sigma+i\tau)\in\Lambda^{(\infty)}(\kappa_0,\nu_0').$$

Finally, we combine (A.10) with (A.11) to derive the main result of the Appendix.

PROPOSITION A.4. *Formula* (A.7) *defines a holomorphic function* $\mathbf{G}\colon\Lambda^{(\infty)}(\kappa_0,\nu_0')\to\mathbb{C}^{N\times N}$. *Recall that* $\nu_0'\in(0,\infty)$ *is fixed, whereas* $\kappa_0\in(0,\infty)$ *is an arbitrary constant. Choose any* $c'\in(c_2,\infty)$, *where* $c_2\ge0$ *appears in* (A.2).

Then, given any $\kappa_0\in(0,\infty)$, $\alpha\in(\mathbb{Z}_+)^D$, *and* $n\in\mathbb{Z}_+$, *there exists a constant* $C'_{\alpha,n}\equiv C'_{\alpha,n}(\kappa_0,\nu_0')\in(0,\infty)$ *such that*

$$\text{(A.12)}\quad \left(1+\frac{|x|}{\sigma^{1/2m}}\right)^{n}\left|\frac{\partial^{|\alpha|}}{\partial z^{\alpha}}\mathbf{G}(z;t)\right|\le C'_{\alpha,n}(\kappa_0,\nu_0')e^{c'\sigma}\sigma^{-(D+|\alpha|)/2m}$$
$$\textit{for all } (z,t)=(x+iy,\sigma+i\tau)\in\Lambda^{(\infty)}(\kappa_0,\nu_0').$$

Remark A.2. We recall that $|\mathbf{M}|=\left(\sum_{j,k=1}^{N}|M_{jk}|^2\right)^{1/2}$ is the Euclidean norm, whereas $\|\mathbf{M}\|$ denotes the ℓ^2-operator norm of a matrix $\mathbf{M}=(M_{jk})_{j,k=1}^{N}\in\mathbb{C}^{N\times N}$. Of course, both these norms are equivalent in $\mathbb{C}^{N\times N}$.

REFERENCES

[1] C. BARDOS AND S. BENACHOUR, *Domaine d'analyticité des solutions de l'équation d'Euler dans un ouvert de R^n*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 4 (1977), pp. 647–687.

[2] M. Bartuccelli, P. Constantin, C. R. Doering, J. D. Gibbon, and M. Gisselfält, *On the possibility of soft and hard turbulence in the complex Ginzburg–Landau equation*, Phys. D, 44 (1990), pp. 421–444.

[3] P. Bollerman, *Validity Results for Ginzburg–Landau's Equation*, Tech. Report 800, Mathematisch Instituut, Rijksuniversiteit Utrecht, Utrecht, The Netherlands, 1993.

[4] P. Collet and J.-P. Eckmann, *The time dependent amplitude equation for the Swift–Hohenberg problem*, Comm. Math. Phys., 132 (1990), pp. 139–153.

[5] P. Collet, J.-P. Eckmann, H. Epstein, and J. Stubbe, *Analyticity for the Kuramoto–Sivashinsky equation*, Phys. D, 67 (1993), pp. 321–326.

[6] A. Doelman, *On the Nonlinear Evolution of Patterns; Modulation Equations and Their Solutions*, Ph.D. thesis, Mathematisch Instituut, Rijksuniversiteit Utrecht, Utrecht, The Netherlands 1990.

[7] A. Doelman and E. S. Titi, *Regularity of solutions and the convergence of the Galërkin method in the Ginzburg–Landau equation*, Numer. Funct. Anal. Optim., 14 (1993), pp. 299–321.

[8] J. Duan and P. Holmes, *On the Cauchy problem for a generalized Ginzburg–Landau equation*, Nonlinear Anal., 22 (1994), pp. 1033–1040.

[9] J. Duan, P. Holmes, and E. S. Titi, *Regularity, approximation and asymptotic dynamics for a generalized Ginzburg–Landau equation*, Nonlinearity, 6 (1993), pp. 915–933.

[10] V. Eckhaus, *The Ginzburg–Landau manifold is an attractor*, J. Nonlinear Sci., 3 (1993), pp. 329–348.

[11] R. E. Edwards, *Functional Analysis: Theory and Applications*, Holt, Rinehart, and Winston, New York, London, 1965.

[12] C. Foias and R. Temam, *Gevrey class regularity for the solutions of the Navier–Stokes equations*, J. Funct. Anal., 84 (1989), pp. 359–369.

[13] A. Friedman, *On the regularity of the solutions of nonlinear elliptic and parabolic systems of partial differential equations*, J. Math. Mech., 7 (1958), pp. 43–59.

[14] ———, *Partial Differential Equations*, Holt, Rinehart, and Winston, New York, London, 1969.

[15] A. van Harten, *On the validity of the Ginzburg–Landau equation*, J. Nonlinear Sci., 1 (1991), pp. 397–422.

[16] W. D. Henshaw, H.-O. Kreiss, and L. G. Reyna, *Smallest scale estimates for the Navier–Stokes equations for incompressible fluids*, Arch. Rational Mech. Anal., 112 (1990), pp. 21–44.

[17] L. Hörmander, *An Introduction to Complex Analysis in Several Variables*, 3rd ed., North–Holland, Amsterdam, New York, 1990.

[18] ———, *The Analysis of Linear Partial Differential Operators* I, 2nd ed., Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1990.

[19] ———, *The Analysis of Linear Partial Differential Operators* II, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1983.

[20] R. B. Hoyle, *Long wavelength instabilities of square patterns*, Phys. D, 67 (1993), pp. 198–223.

[21] F. John, *Partial Differential Equations*, 4th ed., Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1982.

[22] P. Kirrmann, G. Schneider, and A. Mielke, *The validity of modulation equations for extended systems with cubic nonlinearities*, Proc. Roy. Soc. Edinburgh Sect. A, 122 (1992), pp. 85–91.

[23] J. L. Lions and E. Magenes, *Espaces de fonctions et distributions du type de Gevrey et problemes aux limites paraboliques*, Ann. Mat. Pura Appl., 68 (1965), pp. 341–418.

[24] ———, *Espaces du type de Gevrey et problemes aux limites pour diverses classes d'equations d'evolution*, Ann. Mat. Pura Appl., 72 (1966), pp. 343–394.

[25] X. Liu, *Gevrey class regularity and approximate inertial manifolds for the Kuramoto–Sivashinsky equation*, Phys. D, 50 (1991), pp. 135–151.

[26] K. Masuda, *On the regularity of solutions of the nonstationary Navier–Stokes equations*, Lecture Notes in Math., 771, Springer-Verlag, New York, Heidelberg, Berlin, 1980, pp. 360–370.

[27] ———, *Regularity of solutions of partial differential equations*, I, Comment. Math. Univ. St. Paul, 42 (1993), pp. 93–100.

[28] C. B. Morrey, Jr., *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, New York, Heidelberg, Berlin, 1966.

[29] A. C. Newell and J. A. Whitehead, *Finite bandwidth, finite amplitude convection*, J. Fluid Mech., 38 (1969), pp. 279–303.

[30] M. R. E. Proctor, *Planform selection by finite-amplitude thermal convection between poorly conducting slabs*, J. Fluid Mech., 113 (1981), pp. 469–485.

[31] G. Schneider, *Die Gültigkeit der Ginzburg–Landau-approximation*, Ph.D. Thesis, Institut für Mathematik, Universität Stuttgart, Stuttgart, Germany, 1992 (in German).

[32] ———, *Error estimates for the Ginzburg–Landau approximation*, Z. Angew. Math. Phys., 45 (1994), pp. 1–24.

[33] E. M. STEIN, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

[34] E. M. STEIN AND G. WEISS, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.

[35] P. TAKÁČ, *Invariant 2-tori in the time-dependent Ginzburg–Landau equation*, Nonlinearity, 5 (1992), pp. 289–321.

[36] H. TANABE, *On differentiability and analyticity of solutions of weighted elliptic boundary value problems*, Osaka J. Math., 2 (1965), pp. 163–190.

[37] F. TRÈVES, *Introduction to Pseudodifferential and Fourier Integral Operators*, Vol. 1: *Pseudodifferential Operators*; Vol. 2: *Fourier Integral Operators*, Plenum Press, New York, London, 1980.

SIAM J. MATH. ANAL.
Vol. 27, No. 2, pp. 449–475, March 1996

LARGE-TIME BEHAVIOR IN INCOMPRESSIBLE NAVIER–STOKES EQUATIONS*

ANA CARPIO†

Abstract. We give a development up to the second order for strong solutions u of incompressible Navier–Stokes equations in $\mathbb{R}^n$, $n \geq 2$. By combining estimates obtained from the integral equation with a scaling technique, we prove that, for initial data satisfying some integrability conditions (and small enough, if $n \geq 3$), u behaves like the solution of the heat equation taking the same initial data as u plus a corrector term that we compute explicitely.

Key words. incompressible Navier–Stokes equations, strong solutions, large-time behavior, asymptotic development, heat equation

AMS subject classifications. 76D05, 35B40, 35K05

Introduction and main results. This paper is devoted to the study of the large-time behavior of the solutions of the incompressible Navier–Stokes equations in the whole space $\mathbb{R}^n$, $n \geq 2$,

$$\text{(NS)}\qquad \begin{cases} u_t - \Delta u + u^i \partial_i u + \nabla p = 0 & \text{in } \mathbb{R}^+ \times \mathbb{R}^n, \\ u(x) \to 0, & |x| \to \infty, \\ \operatorname{div} u = 0 & \text{in } \mathbb{R}^+ \times \mathbb{R}^n, \\ u(x,0) = u_0, \ \operatorname{div} u_0 = 0 & \text{in } \mathbb{R}^n, \end{cases}$$

where $u = (u^1, \ldots, u^n)$ stands for the velocity of the fluid and p for its pressure.

Let us first recall some facts on solutions of (NS). It is known that for any initial data $u_0 \in (L^2(\mathbb{R}^n))^n$ with div $u_0 = 0$, global weak solutions of (NS) exist. This was first proved by Leray ([18], [19]) for $n \leq 3$ and then by Hopf [13] for all n by means of a Galerkin method. By a weak solution of (NS), we mean a function u such that

$$u \in C_{\text{weak}}([0,\infty); (L^2(\mathbb{R}^n))^n) \quad \text{with} \quad \operatorname{div} u = 0,$$

$$\partial_i u \in L^2(0,\infty; (L^2(\mathbb{R}^n))^n), \quad i = 1, \ldots, n,$$

$$\langle u(0), \varphi(0) \rangle = -\int_0^\infty \langle u, \varphi_t \rangle dt + \int_0^\infty \langle \nabla u, \nabla \varphi \rangle dt + \int_0^\infty \langle u \cdot \nabla u, \varphi \rangle dt$$

for every $\varphi \in (C_c^\infty([0,\infty) \times \mathbb{R}^n))^n$ with div $\varphi = 0$, where $\langle\ \rangle$ denotes the scalar product in $(L^2(\mathbb{R}^n))^n$. From now on, we shall drop the superscript n and denote by X both the spaces X and X^n.

Weak solutions are known to be unique and smooth (hence, strong) when $n = 2$. For higher dimensions, uniqueness and smoothness remain open problems.

Besides the Leray–Hopf construction, there are several methods for proving the existence of weak solutions in $\mathbb{R}^n$ (see [3], [17], [26]). They all construct strong solutions u_k of some approximating problems which converge weakly in $L^2_{\text{loc}}(0,\infty; H^1(\mathbb{R}^n))$ and

*Received by the editors October 8, 1993; accepted for publication (in revised form) July 26, 1994. This research was partially supported by DGICYT (Spain) project PB90-0245.

†Departamento de Matemática Aplicada, Universidad Complutense, 28040 Madrid, Spain (carpio @sunma4.mat.ucm.es).

strongly in $L^2_{\text{loc}}(\mathbb{R}^+ \times \mathbb{R}^n)$ to some weak solution u of (NS). Those u_k fulfill the energy inequality, that is,

$$\|u_k(t)\|^2_{L^2(\mathbb{R}^n)} + 2\int_s^t \|\nabla u_k(\sigma)\|^2_{L^2(\mathbb{R}^n)} d\sigma \leq \|u_k(s)\|^2_{L^2(\mathbb{R}^n)}$$

for $0 \leq s \leq t$. In case $n \leq 4$, the energy inequality also holds for the limit u for every $t > s$ and a.e. $s > 0$ and $s = 0$ (see [14], [26]). Following Leray's terminology, those weak solutions verifying the energy inequality are called turbulent.

For initial data $u_0 \in H^{\frac{1}{2}}(\mathbb{R}^n)$ (see [16]), $u_0 \in L^p(\mathbb{R}^n)$, $p > n$ (see [10], [11], [2], [8]), or $u_0 \in L^n(\mathbb{R}^n)$ (see [14]), strong local solutions are known to exist. They turn out to be global if the norm of the initial data is small. We call them strong since they belong to classes where regularity and uniqueness hold. Besides, they satisfy the equation in the classical sense and both the energy inequality and the associated integral equation are satisfied.

There are some uniqueness criteria allowing to relate strong and weak solutions, provided that they both verify the energy inequality. For instance, if a weak solution u is known to fulfill the energy inequality and we have a strong solution $w \in C([0,T];(L^n(\mathbb{R}^n))^n)$ (see [25]) or $w \in L^r(0,T;L^q(\mathbb{R}^n))$ (see [23]) for some adequate q and r, then w agrees with the weak solution u on $[0,T]$.

We must distinguish the cases $n > 2$ and $n = 2$. For $n = 2$, we shall study the asymptotic behavior of weak solutions of (NS) without smallness assumptions on the data. When $n > 2$, we are concerned with the study of the asymptotic behavior of global strong solutions of (NS) like those constructed in [14] with data in $L^n(\mathbb{R}^n)$ of small norm. If u_0 also belongs to some $L^p(\mathbb{R}^n)$ with $1 < p \leq n$, Kato obtains decay rates for the $L^q(\mathbb{R}^n)$ norms, $q \geq p$, similar to those which hold for the heat equation (see also [2]), provided that $\frac{n}{2}(\frac{1}{p} - \frac{1}{q}) < 1$, which always holds when $n = 2$. We shall remove this restriction and extend the decay estimates to reach the case $p = 1$.

Let us consider first the case $n = 2$. In this case, weak solutions of (NS) turn out to also be strong. For data in $L^p \cap L^2(\mathbb{R}^n)$, $1 \leq p \leq 2$, an argument involving the use of Fourier transforms allows to prove (see [26], [17]) that the weak solutions of (NS) behave in L^2 like the solutions of the heat equation with the same initial data when $t \to \infty$. We extend this result to L^q with $q \neq 2$.

THEOREM 0.1. *Let u be a weak solution of the two-dimensional* (NS) *with initial data $u_0 \in L^p \cap L^2(\mathbb{R}^2)$, $1 \leq p \leq 2$, such that* div $u_0 = 0$. *Then for any $q \geq p$,*

(i) *if $1 < p < 2$,*

$$\|G(t) * u_0 - u(t)\|_q \leq Ct^{-\frac{2}{p}+\frac{1}{q}+\frac{1}{2}}, \quad t > 0,$$

(ii) *if $p = 1$,*

$$\|G(t) * u_0 - u(t)\|_q \leq Ct^{\frac{-3}{2}+\frac{1}{q}} \log\, t, \quad t > 0,$$

(iii) *if $p = 2$,*

$$t^{\frac{1}{2}-\frac{1}{q}}\|u(t)\|_q \to 0 \quad as\ t \to \infty,$$

where we denote the heat kernel by $G(t)$.

In this theorem and what follows, C denotes a positive constant independent of time.

In cases (i) and (ii) $\|G(t)*u_0\|_q$ decays at a slower rate than the powers appearing in the right-hand side when $t \to \infty$. Therefore, we may say that $G(t) * u_0$ is the first term in the asymptotic development of u when $t \to \infty$.

Both (i) and (ii) follow easily from the integral equation satisfied by u thanks to the decay estimates on the L^q norms of u obtained by Kato for data of small L^2 norm. Since the L^2 norm of u is known to tend to 0 as $t \to \infty$, we can spare this smallness hypothesis.

Kato's estimates together with the covergence to zero of the L^2 norm yield (iii), which also holds for $\|G(t) * u_0\|_q$. In fact, at least when $q = 2$, this decay estimate turns out to be optimal for both the heat and Navier–Stokes equations in the sense that no uniform decay rate can be found for the L^2 norm of solutions with initial data $u_0 \in L^2(\mathbb{R}^2)$ (see [21]). However, we ignore whether it is possible to find functions $g(t)$ and $\delta(t)$ with $\delta(t) \to \infty$ as $t \to \infty$ such that

$$\delta(t)\, t^{\frac{1}{2}-\frac{1}{q}} \|u(t) - g(t)\|_q \to 0 \quad \text{as } t \to \infty$$

under the only assumption $u_0 \in L^2(\mathbb{R}^2)$ (see remarks in §1.2). The first term in this case (other than 0) is unknown.

In some cases we can make the above result more precise.

THEOREM 0.2. *Let u be a solution of the two-dimensional* (NS) *with initial data $u_0 \in L^p \cap L^2(\mathbb{R}^2)$, $\frac{6}{5} < p < 2$, and v a solution of*

$$(\mathcal{L}_2) \qquad \begin{cases} v_t - \Delta v = -h^i\partial_i h - \partial_j \nabla E_2 * h^i \partial_i h^j & \text{in } \mathbb{R}^+ \times \mathbb{R}^2, \\ \operatorname{div} v = 0 & \text{in } \mathbb{R}^+ \times \mathbb{R}^2, \\ v(x,0) = u_0, \ \operatorname{div} u_0 = 0 & \text{in } \mathbb{R}^2, \end{cases}$$

where $h(t) = G(t) * u_0$ is the solution of the heat equation with data u_0 and E_2 stands for the fundamental solution of $-\Delta$ in $\mathbb{R}^2$. Then for any $q \geq p$, we have

$$\|(u - v)(t)\|_q \leq Ct^{-\frac{3}{p}+\frac{1}{q}+1}.$$

We see that, when $\frac{6}{5} < p < 2$, the function v, that can be written as

$$v(t) = G(t) * u_0 - \int_0^t G(t-s) * (h^i\partial_i h + \partial_j \nabla E_2 * h^i \partial_i h^j)(s)ds,$$

that is, $G(t) * u_0$ plus a corrector term $I(t)$ approaches u better that $G(t) * u_0$. The restriction on p is needed to make some integrals finite when estimating the difference by using the integral equations.

For initial data $u_0 \in L^p \cap L^2(\mathbb{R}^2)$ with $1 < p \leq \frac{6}{5}$, Theorem 0.2 implies

$$\|(u - v)(t)\|_q \leq Ct^{-\frac{3}{r}+\frac{1}{q}+1}$$

when $q \geq r$, for any $\frac{6}{5} < r < 2$. This decay is faster that the decay $t^{-\frac{2}{p}+\frac{1}{q}+\frac{1}{2}}$ observed for $\|G(t) * u_0 - u(t)\|_q$. Therefore, the second term in the development, in norm L^q, $q \geq \frac{6}{5}$, is again $I(t)$.

When $p = 1$, Theorem 0.1 yields the decay rate $Ct^{\frac{-3}{2}+\frac{1}{q}} \log\, t$ for $\|G(t) * u_0 - u(t)\|_q$, $q \geq 1$. By Theorem 0.2, we get the slower decay rate $Ct^{-\frac{3}{r}+\frac{1}{q}+1}$ with $r > \frac{6}{5}$ for $\|v(t) - u(t)\|_q$ $q \geq r$. However, in this case $p = 1$ and, provided some integrability hypotheses on the data are added, we can use a scaling technique to find the second term in the development of u. We have the following theorem.

THEOREM 0.3. *Set $q \geq 1$. Let u be a solution of the two-dimensional* (NS) *with initial data $u_0 \in (L^2 \cap L^1)(\mathbb{R}^2, 1+|x|)$ such that* $\operatorname{div} u_0 = 0$ *and $u_0 \in L^{2r}(\mathbb{R}^2)$ for some*

r satisfying $q \geq r > \frac{2q}{q+2}$. We set $M = \int_{\mathbb{R}^2} u_0(x)dx = (M^1, M^2)$ and $E_2 = \frac{1}{2\pi}\log|x|$. Then,

$$\frac{t^{\frac{3}{2}-\frac{1}{q}}}{\log t}\|u(t) - MG(t) + R(t)\|_q \to 0 \quad as\ t \to \infty,$$

where

$$R(t) = \log t\ \left(\frac{M^iM}{2}\partial_i G(t) + \frac{M^iM^j}{2}\partial_i G(t) * \nabla\partial_j E_2\right).$$

This result remains true when replacing u with the solution v of $(\mathcal{L}_2)$ so that $R(t)$ is also the second term in the development of v. The term $MG(t)$ comes from $G(t) * u_0$, while $R(t)$ is the contribution due to the integral term $I(t)$. Note that, formally, we have set $h(t) = MG(t)$ in the expresion of $I(t)$. It is clear then that, for u_0 as in Theorem 0.3,

$$\frac{t^{\frac{3}{2}-\frac{1}{q}}}{\log t}\|u(t) - v(t)\|_q \to 0 \quad \text{as } t \to \infty$$

so that, again, v approaches u better than $G(t) * u_0$.

When the mass of the initial data is zero, Theorem 0.3 reduces to

$$\frac{t^{\frac{3}{2}-\frac{1}{q}}}{\log t}\|u(t)\|_q \to 0 \quad \text{as } t \to \infty$$

It was proved in [22] that, if $u_0 \in L^1(\mathbb{R}^2, 1 + |x|^2) \cap L^2(\mathbb{R}^2, |x|^{\frac{1}{2}}) \cap H(\mathbb{R}^2)$, where $H(\mathbb{R}^2)$ denotes the closure in L^2 of $C_0^\infty(\mathbb{R}^2) \cap \{u \text{ s.t. div } u = 0\}$ and the mass of u_0 is 0, that is, $\mathcal{F}u_0 = 0$, then $\|u(t)\|_2 \leq C(1+t)^{-1}$. Solutions may decay faster, even exponentially, depending on the order of the zero of $\mathcal{F}u$ at time 0.

We shall also extend the above theorems to higher dimensions.

THEOREM 0.4. *For $n \geq 3$, let u be a global strong solution of* (NS) *with data $u_0 \in L^p \cap L^n(\mathbb{R}^n)$, $1 \leq p < n$, of L^n norm small enough and such that* div $u_0 = 0$. *Then for $q \geq p$, we have*

$$\|G(t) * u_0 - u(t)\|_q \leq Ct^{(-\frac{2}{p}+\frac{1}{q})\frac{n}{2}+\frac{1}{2}}.$$

For weak solutions satisfying the energy inequality (or that can be approached by solutions of approximating problems verifying it), the above decay estimate on $\|G(t) * u_0 - u(t)\|_q$ was known to hold for $q = 2$ and $1 \leq p \leq 2$ (see [17], [26]). For divergence free data u_0 belonging to $L^2 \cap L^n(\mathbb{R}^n)$ with small $L^n(\mathbb{R}^n)$ norm, there exists a unique strong global solution and at least one global weak solution. They both agree when the weak solution satisfies the energy inequality. These kinds of weak solutions are known to exist when $n \leq 4$ (see, for instance, [26]). Therefore, Theorem 0.4 extends the results known for weak solutions.

Denoting the n-dimensional analogues of problem $(\mathcal{L}_2)$ by $(\mathcal{L}_n)$, we get that in some cases the solution v of $(\mathcal{L}_n)$ approaches u better than $G(t) * u_0$, furnishing the second term in the development of u.

THEOREM 0.5. *For $n \geq 3$, let u be a global strong solution of* (NS) *with data $u_0 \in L^p \cap L^n(\mathbb{R}^n)$, $\frac{3n}{n+3} < p < n$, whose L^n norm is small enough and such that* div $u_0 = 0$. *Then for any $q \geq p$, we have*

$$\|(u - v)(t)\|_q \leq Ct^{(-\frac{3}{p}+\frac{1}{q})\frac{n}{2}+1},$$

where v is the solution of $(\mathcal{L}_n)$ with data u_0.

Taking $u_0 \in L^p \cap L^n(\mathbb{R}^n)$, $\frac{2n}{n+2} < p \leq \frac{3n}{n+3}$, of small L^n norm such that div $u_0 = 0$, we conclude that

$$\|(u-v)(t)\|_q \leq Ct^{(-\frac{3}{r}+\frac{1}{q})\frac{n}{2}+1}$$

for $q \geq r$ and $\frac{3n}{n+3} < r < n$. This decay rate is faster than that observed for $\|G(t) * u_0 - u(t)\|_q$. However, we ignore what happens when $p \leq q \leq \frac{3n}{n+3}$ or $1 < p \leq \frac{2n}{n+2}$.

Theorems 0.4 and 0.5 are obtained estimating the norms by using the integral equations and the known decay estimates. We can handle the case $p = 1$ by using a scaling technique, provided that some integrability hypotheses are added.

THEOREM 0.6. *For $n \geq 3$, let u be a strong solution of* (NS) *with data $u_0 \in L^1(\mathbb{R}^n, 1+|x|) \cap L^n(\mathbb{R}^n)$, $1 \leq p \leq n$, of L^n norm small enough and such that* div $u_0 = 0$ *and $q \geq 1$. If $u_0 \in L^{2r}(\mathbb{R}^n)$ for some $q \geq r > \frac{nq}{q+n}$, then*

$$t^{\frac{1}{2}+\frac{n}{2}(1-\frac{1}{q})}\|u(t) - MG(t) + m^i\partial_i G(t) + R(t)\|_q \to 0$$

as $t \to \infty$, where

$$M = \int_{\mathbb{R}^n} u_0(x)dx; \quad m_i = \int_{\mathbb{R}^n} x^i u_0(x)dx, \ i = 1, \ldots, n,$$

$$R(t) = \left(\int_0^\infty \int_{\mathbb{R}^n} u^i u(\sigma, y)dyd\sigma\right) \partial_i G(t) + \left(\int_0^\infty \int_{\mathbb{R}^n} u^i u^j(\sigma, y)dyd\sigma\right) \partial_i G(t) * \nabla\partial_j E_n,$$

and E_n stands for the fundamental solution of $-\Delta$ in $\mathbb{R}^n$.

The same result holds if we replace u by v so that, in particular, we see that v approaches u better than $G(t) * u_0$. Both Theorem 0.3 and Theorem 0.6 extend to (NS) the results proved in [27] for the following scalar convection-diffusion equations:

$$u_t - \Delta u + a|u| \cdot \nabla u = 0 \quad \text{in } \mathbb{R}^+ \times \mathbb{R}^n, \quad a \in \mathbb{R}^n,$$

where the same difference between the case $n = 2$ and $n \geq 3$ appears. We remark that the solutions we are dealing with satisfy the decay estimate $\|u(t)\|_2 \leq C(1+t)^{\frac{-n}{4}}$. When $n \geq 3$, this decay ensures the existence of $\int_0^\infty \int_{\mathbb{R}^n} u^i u^j(\sigma, y)dyd\sigma$ for $i, j = 1, \ldots, n$. For $n = 2$, we obtain an upper bound for $\int_0^t \int_{\mathbb{R}^n} u^i u^j(\sigma, y)dyd\sigma$ which grows as log t grows. The results in [27] were proved changing to self-similar variables and then making eigenvalue expansions in some weighted Sobolev spaces. Our technique can be adapted to yield another proof of these results.

Theorem 0.6 is related to a result obtained by Schonbek in [22]. When $n = 3$, she proved that $k_1(1+t)^{\frac{-5}{4}} \leq \|u(t)\|_2 \leq k_2(1+t)^{\frac{-5}{4}}$ for a special class of data. Theorem 0.6 implies that $u(t,x) = MG(t,x) + m^i\partial_i G(t,x) + R(t,x) + r(t,x)$ where $\|r(t,x)\|_2 = o(t^{\frac{-5}{4}})$ as $t \to \infty$ and $\|MG(t,x) + m^i\partial_i G(t,x) + R(t,x)\|_2 = Ct^{\frac{-5}{4}}$.

The paper is organized as follows. In §1, we briefly recall some known results which will be of use to us in what follows. The next two sections are devoted to the proof of Theorems 0.1 and 0.2, respectively. In §4, we study some related linear problems and prove Theorem 0.3. The last section deals with the asymptotic behavior in higher dimensions.

1. Known results.

1.1. Strong solutions. The following results are taken from [14]. They are established by using an iterative scheme that goes back to Leray. The idea is to convert (NS) to an integral equation

$$u(t) = G(t) * u_0 + \int_0^t \partial_i G(t-s) * P(u_i u(s))ds = Su(t),$$

where $G(t)$ denotes the heat kernel and P the projection on the space of divergence free vectors, which is a bounded operator from L^p to L^p for $1 < p < \infty$. Taking $u_0 \in L^n(\mathbb{R}^n)$, the sequence $u^{m+1} = Su^m$, $m \geq 1$, with $u^1 = G(t) * u_0$ converges strongly to a solution u of the integral equation on some time interval $[0,T]$. If $\|u_0\|_n$ is small enough, we can take $T = \infty$. The construction also yields the estimates below.

THEOREM 1.1 (see [14]). *Let* $u_0 \in L^n(\mathbb{R}^n)$ *be such that* div $u_0 = 0$ *and* $\|u_0\|_n$ *is small enough. Then there exists a unique solution* u *of* (NS) *such that*

(i)

$$t^{\frac{n}{2}(\frac{1}{n}-\frac{1}{q})}u \in BC([0,\infty);L^q(\mathbb{R}^n)) \quad \forall n \leq q \leq \infty,$$

(ii)

$$t^{\frac{n}{2}(\frac{1}{n}-\frac{1}{q})+\frac{1}{2}}\nabla u \in BC([0,\infty);L^q(\mathbb{R}^n)) \quad \forall n \leq q < \infty.$$

Moreover,

(iii)

$$u \in L^r((0,\infty);L^q(\mathbb{R}^n)), \quad \frac{1}{r} = \frac{n}{2}\left(\frac{1}{n}-\frac{1}{q}\right), \quad n < q < \frac{n^2}{n-2},$$

(iv)

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \|u(s)\|_n ds = 0.$$

Remark 1.1. When $n = 2$, $\|u(t)\|_2$ is monotonically nonincreasing (the energy inequality holds) and (iv) implies that $\|u(t)\|_2 \to 0$ as $t \to \infty$ if $\|u_0\|_2$ is small.

THEOREM 1.2 (see [14]). *Let* $u_0 \in L^n(\mathbb{R}^n) \cap L^p(\mathbb{R}^n)$ *such that* div $u_0 = 0$ *and* $\|u_0\|_n$ *is small enough, with* $1 < p < n$. *Then*

(i)

$$t^{\frac{n}{2}(\frac{1}{p}-\frac{1}{q})}u \in BC([1,\infty);L^q(\mathbb{R}^n)) \quad \forall p \leq q \leq \infty,$$

(ii)

$$t^{\frac{n}{2}(\frac{1}{p}-\frac{1}{q})+\frac{1}{2}}\nabla u \in BC([1,\infty);L^q(\mathbb{R}^n)) \quad \forall p \leq q < \infty$$

if $\frac{n}{2}(\frac{1}{p}-\frac{1}{q}) < 1$ *in* (i) *or* $\frac{n}{2}(\frac{1}{p}-\frac{1}{q}) + \frac{1}{2} < 1$ *in* (ii); *otherwise, we replace them by any positive number less than* 1.

(iii)

$$\|u(t) - G(t) * u_0\|_p \leq C_\delta t^{\frac{-\delta}{2}}, \quad 0 < \delta < \text{Min}\left(1, n\left(1-\frac{1}{p}\right), \frac{n}{p}-1\right).$$

Remarks 1.2.

- In view of Theorem 1.1(i) and the fact that $u \in BC([0,\infty);L^p(\mathbb{R}^n))$ (see [14]) when $u_0 \in L^p, 1 < p < n$, we can replace $BC([1,\infty);L^q(\mathbb{R}^n))$ in Theorem 1.2(i) by $BC([0,\infty);L^q(\mathbb{R}^n))$ for any $q \geq p$.
- If $\|u_0\|_n$ is small then $\|u(t)\|_n \to 0$ as $t \to \infty$ (see Masuda's remark in [14]).
- Strong solutions have been also obtained for data $u_0 \in L^r(\mathbb{R}^n)$, $r > n$ (see [10], [8])
- In [2], strong solutions with data $u_0 \in L^2 \cap L^n(\mathbb{R}^n)$ are constructed in a different way. A decay estimate $\|u(t)\|_n \leq Ct^{\frac{n}{2}(\frac{1}{2}-\frac{1}{n})}$ without restrictions on the size of $\frac{n}{2}(\frac{1}{2}-\frac{1}{n})$ is also obtained.
- When $n = 2$, Theorem 1.2(i) holds for any $q \geq p$. If $n \geq 3$ it holds for any $q \geq p$ when $p \geq \frac{n}{2}$ and for $p \leq q \leq \frac{pn}{n-2p}$ when $p < \frac{n}{2}$.

1.2. Weak solutions. The use of Fourier transform allows us to obtain decay estimates for weak solutions with initial data $u_0 \in L^2(\mathbb{R}^n)$ without smallness hypotheses, provided that they satisfy the energy inequality or can be approached by solutions of related problems verifying it. The first results in that direction are due to Schonbek [20]. The idea consists in taking the Fourier transform of the equation and splitting the frequency domain in order to get differential inequalities yielding some decay. We do this first for some approached solutions and in the limit we get a decay estimate for u. This result has been successively improved and extended in [21], [22], [17], and [26]. The following theorem is taken from [26].

THEOREM 1.3 (see [26]). *Let u be a weak solution of the incompressible* (NS) *equations which satisfies the energy inequality (or can be approached by solutions of approximated problems satisfying it) for any $n \geq 2$. Then for every $u_0 \in L^2(\mathbb{R}^n)$ with* $\operatorname{div} u_0 = 0$,

(i) $\|u(t)\|_2 \to 0$ *as* $t \to \infty$.

If, further, $\|G(t) * u_0\|_2^2 \leq C(1+t)^{-\alpha_0}$ *for all* $t \geq 0$, *then*

(ii) $\|u(t)\|_2^2 \leq C(1+t)^{-\alpha}$ *with* $\alpha = \operatorname{Min}(\frac{n}{2}+1, \alpha_0)$, $t \geq 0$, *and*

(iii) $\|u(t) - G(t) * u_0\|_2^2 \leq h(t)(1+t)^{-d}$ *for all* $t \geq 0$, *where*

$$d = \frac{n}{2} + 1 - 2\operatorname{Max}(1-\alpha_0, 0) \quad \left(d > \alpha = \alpha_0 \; \textit{if} \; \frac{n}{2} + 1 > \alpha_0\right),$$

$$h(t) \;=\; \begin{cases} \varepsilon(t) \to 0, \quad t \to \infty & \textit{if}\; \alpha = 0, \\ C \ln^2(t+c) & \textit{if}\; \alpha = 1, \\ C & \textit{if}\; \alpha \neq 0, 1. \end{cases}$$

Remark 1.3. Therefore, if $u_0 \in L^p \cap L^2$, $1 \leq p \leq 2$, we have $\alpha_0 = \frac{n}{2}(\frac{1}{p} - \frac{1}{2})$ and

$$\|u(t) - G(t) * u_0\|_{L^2(\mathbb{R}^2)} \;\leq\; g(t)(1+t)^{-\frac{n}{2}(\frac{1}{p}-\frac{1}{2})} \quad \forall t \geq 0$$

with

$$g(t) \;=\; \begin{cases} \varepsilon(t) \to 0, \quad t \to \infty & \text{if } p = 2, \\ C \ln(t+c)(1+t)^{\frac{-1}{2}} & \text{if } p = 1, \\ (1+t)^{(\frac{1}{2} - \frac{n}{2p})} & \text{if } 1 < p < 2. \end{cases}$$

Remark 1.4. Some results on the behavior of weak solutions in exterior domains are also known. See, for instance, [1] and the references therein.

Remark 1.5. As we said in the introduction, lower bounds for the decay of the L^2 norm and faster decay rates when $n = 2$ and the mass of the data is 0 have been obtained in [22].

Remark 1.6. Concerning the case $p = 2$ in Remark 1.3, it is known that both $\|G(t) * u_0\|_2$ and $\|u(t)\|_2$ tend to zero as t goes to infinity. Moreover, in both cases no uniform decay rate can be found. This is well known for the heat equation, where $\|G(t) * u_0\|_2$ can decay at an arbitrarily slow algebraic rate or even exponentially by choosing an adequate $u_0 \in L^2(\mathbb{R}^n)$ (see [21], for instance). For Navier–Stokes equations in dimensions two and three, the proof of the lack of uniformity is due to Schonbek [21]. When $n \geq 3$ a function $\delta(t) \to \infty$, $t \to \infty$ can be found in such a way that

$$\delta(t)\|G(t) * u_0 - u(t)\|_2 \to 0 \quad \text{as } t \to \infty.$$

By Theorem 2 (iii) of [17], we may choose $\delta(t) = t^{\frac{n}{4}-\frac{1}{2}}$. However, we ignore whether such a $\delta(t)$ exists when $n = 2$.

1.3. Solutions with singular data. When $n = 2$, solutions of this kind have been constructed in [11] and its asymptotic behavior is studied in [12] and [6], where the following result is proved.

THEOREM 1.4. *Let u be a solution of* (NS) *with initial data $u_0 \in L^{2,\infty}(\mathbb{R}^2)$ such that* div $u_0 = 0$ *and $v_0 =$* curl $u_0 \in M(\mathbb{R}^2)$. *If the total variation of v_0 is small (see* [12]) *or, more generally, the mass of v_0, $|\int_{\mathbb{R}^2} v_0|$ is small (see* [6]), *then*

$$t^{\frac{1}{2}-\frac{1}{q}} \,\|u(t) - G(t) * u_0\|_q \to 0, \quad t \to \infty \quad \forall q > 2,$$

$$t^{1-\frac{1}{q}} \,\|\nabla u(t) - \nabla G(t) * u_0\|_q \to 0, \quad t \to \infty, \quad \forall 1 < q < \infty.$$

We denote the space of finite measures by $M(\mathbb{R}^2)$ and the usual Lorenz space by $L^{2,\infty}(\mathbb{R}^2)$.

The energy of these solutions is not necessarily finite. If we take $u_0 \in L^2(\mathbb{R}^2) \subset L^{2,\infty}(\mathbb{R}^2)$ with div $u_0 = 0$ and without hypothesis on curl u_0, then Theorem 1.3 asserts that

$$\|u(t) - G(t) * u_0\|_{L^2(\mathbb{R}^2)} \to 0 \quad \text{as } t \to \infty$$

but it gives no information on L^q norms with $q \neq 2$ or on the behavior of $\nabla u(t)$.

In higher dimensions, solutions with data in Morrey spaces have been constructed in [24] and [15]. However, little is known about the asymptotic behavior.

2. Dimension two: First term.

Let us take $u_0 \in L^2(\mathbb{R}^2)$ such that div $u_0 = 0$ and let u be the corresponding weak solution of (NS), which is known to be unique and smooth. Since $\|u(t)\|_2 \to 0$ as $t \to \infty$ (Theorem 1.3) we can choose t_0 such that $\|u(t_0)\|_2$ is small enough to apply the estimates of Theorem 1.1.

If we also assume $u_0 \in L^p(\mathbb{R}^2)$ for some $1 \leq p < 2$, then we know that $u(t) \in L^p$ for $t \geq 0$ (see [17]). Therefore, the decay estimates furnished by Theorem 1.2 also apply.

First, we are going to extend Theorem 1.2 to the case $p = 1$. In order to do that, we need the following result.

LEMMA 2.1. *Let $G(t)$ be the n-dimensional heat kernel. Then, for every $i = 1, \ldots, n$ and every $t > 0$, $\partial_i G(t)$ belongs to the Hardy space $\mathcal{H}^1(\mathbb{R}^n)$ and*

$$\|\partial_i G(t)\|_{\mathcal{H}^1(\mathbb{R}^n)} \leq C t^{\frac{-1}{2}}.$$

Proof. There are several equivalent definitions of $\mathcal{H}^1$ (see [9]).

$$\begin{aligned}\mathcal{H}^1(\mathbb{R}^n) &= \{u \in L^1(\mathbb{R}^n) \text{ s.t. } R_i * u \in L^1(\mathbb{R}^n),\ i = 1, \ldots, n\} \\ &= \{u \in L^1(\mathbb{R}^n) \text{ s.t. } \operatorname*{Sup}_{s>0} |h_s * u| \in L^1(\mathbb{R}^n)\},\end{aligned}$$

where $R_i(x) = \frac{x_i}{|x|^n}$ and $h_s(x) = s^{-n} h(\frac{x}{s})$ with $h \in \mathcal{S}(\mathbb{R}^n)$ such that $0 \leq h \leq 1$ and $\int h = 1$. We may endow $\mathcal{H}^1(\mathbb{R}^n)$ with either of the equivalent norms

$$\|u\|_{L^1(\mathbb{R}^n)} + \sum_{i=1}^{n} \|R_i * u\|_{L^1(\mathbb{R}^n)}$$

or

$$\|u\|_{L^1(\mathbb{R}^n)} + \left\| \operatorname*{Sup}_{s>0} |h_s * u| \right\|_{L^1(\mathbb{R}^n)}.$$

Let us fix any $i \in 1, \dots, n$. In order to prove that $\partial_i G(t) \in \mathcal{H}^1$, it suffices to verify that $\mathrm{Sup}_{s>0}|h_s * \partial_i G(t)| \in L^1(\mathbb{R}^n)$, where $h_s(x) = s^{-n}h(\frac{x}{s})$ and $h \in \mathcal{S}(\mathbb{R}^n)$ is such that $0 \le h \le 1$ and $\int_{\mathbb{R}^n} h(x)dx = 1$. We take $h = G(1)$ and then

$$|h_s * \partial_i G(t)(x)| = |\partial_i(G(t+s,x))| = \frac{|x_i|}{(4\pi(t+s))^{\frac{n}{2}}} e^{\frac{-|x|^2}{4(t+s)}}$$

so that

$$\operatorname*{Sup}_{s>0} |h_t * \partial_i G(t)| = |\partial_i G(t)| \in L^1(\mathbb{R}^n)$$

and

$$\|\partial_i G(t)\|_{\mathcal{H}^1} \le C t^{\frac{-1}{2}} \qquad \square$$

PROPOSITION 2.2. *Let u be a solution of* (NS) *in dimension two with initial data $u_0 \in L^1 \cap L^2(\mathbb{R}^2)$ such that* $\operatorname{div} u_0 = 0$. *Then*

(i) $u(t) \in L^q, \quad t > 0, \quad 1 \le q \le 2,$

(ii) $\|u(t)\|_q \le C t^{-1+\frac{1}{q}}, \quad t > 0, \quad q > 1,$

(iii) $\|u(t)\|_1 \le C, \quad t \ge 0.$

Remark 2.1. These estimates extend the estimates known for $q = 2$ (see [26], [17]).

Proof. (i) It is well known that $u \in L^\infty([0,\infty); L^2(\mathbb{R}^2))$. By taking the divergence of the equation we get the following equation for the pressure:

$$-\Delta\, p = \partial_j(u^i \partial_i u^j),$$

so that, up to a function of time, the pressure is given by $p = E_2 * \partial_j(u^i\partial_i u^j)$, where E_2 denotes the fundamental solution of $-\Delta$ in $\mathbb{R}^2$. Let us write the associated integral equation

$$\begin{aligned} u(t) = {} & G(t) * u_0 + \int_0^t \partial_i G(t-s) * u^i u(s) ds \\ & + \int_0^t \partial_i G(t-s) * \partial_j \nabla E_2 * u^i u^j(s) ds. \end{aligned}$$

Since $u_0 \in L^1$, $G(t) * u_0 \in L^q$ for all $q \ge 1$ and $t > 0$. On the other hand, $u(s) \in L^2$ implies that $u^i u(s) \in L^1$ and

$$\left\| \int_0^t \partial_i G(t-s) * u^i u(s) ds \right\|_q \le C \int_0^t (t-s)^{-1+\frac{1}{q}-\frac{1}{2}} \|u(s)\|_2^2 ds \le C t^{\frac{1}{q}-\frac{1}{2}},$$

provided that $1 \le q < 2$. Therefore, the first integral belongs to L^q for all $1 \le q < 2$.

As far as the second integral is concerned, since $\partial_i G(t)$ (Lemma 2.1) belongs to the Hardy space $\mathcal{H}^1(\mathbb{R}^2)$ and $\partial_j \nabla E_2$ is a Calderon–Zygmund kernel, we conclude that $\partial_i G(t-s) * \partial_j \nabla E_2 \in L^1$ and

$$\|\partial_i G(t-s) * \partial_j \nabla E_2\|_1 \le C \|\partial_i G(t-s)\|_{\mathcal{H}^1} \le C(t-s)^{\frac{-1}{2}}$$

(see [5]). Then

$$\left\| \int_0^t \partial_i G(t-s) * \partial_j \nabla E_2 * u^i u^j(s) ds \right\|_1 \le \int_0^t (t-s)^{\frac{-1}{2}} \|u(s)\|_2^2 ds \le C t^{\frac{1}{2}}.$$

In the same way, since $\partial_j \nabla E_2$ is a Calderon–Zygmund kernel, we have

$$\|\partial_i G(t-s) * \partial_j \nabla E_2\|_q \leq C_q \|\partial_i G(t-s)\|_q, \quad 1 < q < \infty$$

(see [5]) so that we get

$$\left\| \int_0^t \partial_j \nabla E_2 * \partial_i G(t-s) * u^i u^j(s) ds \right\|_q \leq \int_0^t (t-s)^{-1+\frac{1}{q}-\frac{1}{2}} \|u(s)\|_2^2 ds \leq C t^{\frac{1}{q}-\frac{1}{2}}$$

for $1 < q \leq 2$. Thus, the second integral belongs also to L^q for all $1 \leq q < 2$.

Remark 2.2. It is known that if $A \in (L^p)^n$, $B \in (L^{p'})^n$ are such that $\operatorname{div} A = 0$, $\operatorname{curl} B = 0$ (see [4]). Then, $A \cdot B$ belongs to the Hardy space $\mathcal{H}^1$ and

$$\|A \cdot B\|_{\mathcal{H}^1} \leq C \|A\|_p \|B\|_{p'}.$$

In our case, for a.e. $s \geq 0$ and every $j = 1, \dots, n$, we have $u(s), \nabla u^j(s) \in L^2$ with $\operatorname{div} u = 0$ and $\operatorname{curl}(\nabla\, u^j) = 0$. Therefore, $u^i \partial u^j(s) \in \mathcal{H}^1$. Taking into account that $\partial_j \nabla E_2$ is a Calderon–Zygmund kernel, we conclude that $\partial_j \nabla E_2 * u^i \partial_i u^j(s) \in L^1$ for a.e. $s > 0$.

(ii) Taking norms in the integral equation,

$$\begin{aligned} u(t) = {} & G(t) * u_0 + \int_0^t \partial_i G(t-s) * u^i u(s) ds \\ & + \int_0^t \partial_i G(t-s) * \partial_j \nabla E_2 * u^i u^j(s) ds, \end{aligned}$$

we get

$$\|u(t)\|_q \leq \|G(t) * u_0\|_q + \int_0^t \|\partial_i G(t-s) * u^i u(s)\|_q ds.$$

Taking into account some classical estimates on the heat kernel when $n = 2$,

$$\|\partial^\alpha G(t) * a\|_q \leq C t^{-\frac{|\alpha|}{2}-\frac{1}{r}+\frac{1}{q}} \|a\|_r, \quad q \geq r$$

together with the fact that

$$\|\partial_j \nabla E_2 * u^i u^j(s)\|_r \leq C \|u^i u^j(s)\|_r, \quad 1 < r < \infty,$$

it follows that

$$\|u(t)\|_q \leq C t^{-1+\frac{1}{q}} + C \int_0^t (t-s)^{-\frac{1}{2}-\frac{1}{r}+\frac{1}{q}} s^{\frac{-2}{p}+\frac{1}{r}} ds$$

thanks to the estimates

$$\|u(t)\|_{2r}^2 \leq C t^{2(\frac{-1}{p}+\frac{1}{2r})}$$

valid for $2 > p > 1$ if $2r \geq p$. To prove these estimates, it suffices to observe that (Theorem 1.3) the L^2 norm of u tends to 0 as t goes to infinity and then apply Theorem 1.1 and Remark 1.2.

We split the integral appearing in the inequality as follows:

(a)

$$\int_{\frac{t}{2}}^t (t-s)^{-\frac{1}{2}-\frac{1}{r}+\frac{1}{q}} s^{\frac{-2}{p}+\frac{1}{r}} \leq C\, t^{\frac{1}{2}+\frac{1}{q}-\frac{2}{p}},$$

choosing r such that $\frac{1}{2} - \frac{1}{r} + \frac{1}{q} > 0$, that is, $q \geq r > \frac{2q}{q+2}$.

(b)

$$\int_0^{\frac{t}{2}} (t-s)^{-\frac{1}{2}-\frac{1}{r}+\frac{1}{q}} s^{\frac{-2}{p}+\frac{1}{r}} \; \leq \; C\, t^{\frac{1}{2}+\frac{1}{q}-\frac{2}{p}},$$

choosing r such that $\frac{-2}{p} + \frac{1}{r} + 1 > 0$, that is, $1 < r < \frac{p}{2-p}$, which is possible if $p > 1$.

Therefore,

$$\|u(t)\|_q \; \leq \; Ct^{-1+\frac{1}{q}} + Ct^{\frac{1}{2}+\frac{1}{q}-\frac{2}{p}} \; \leq \; Ct^{-1+\frac{1}{q}}$$

if $1 < p \leq \frac{4}{3}$, where C is a constant depending on q and on the data.

(iii) Taking norms in

$$u(t) = G(t) * u_0 \; + \; \int_0^t \partial_i G(t-s) * u^i u(s) ds$$
$$+ \int_0^t \partial_i G(t-s) * \partial_j \nabla E_2 * u^i u^j(s) ds,$$

we get

$$\|G(t) * u_0\|_1 \leq C,$$

$$\left\| \int_0^t \partial_i G(t-s) * u^i u(s) ds \right\|_1 \leq C \int_0^t (t-s)^{\frac{-1}{2}} s^{\frac{-2}{p}+1} \leq Ct^{\frac{3}{2}-\frac{2}{p}} \leq C$$

if $p = \frac{4}{3}$ and also

$$\left\| \int_0^t \partial_i G(t-s) * \partial_j \nabla E_2 * u^i u^j(s) ds \right\|_1 \leq \int_0^t \|\partial_i G(t-s) * \partial_j \nabla E_2\|_1 \|u^i u^j(s) ds\|_1 \leq C$$

when $p = \frac{4}{3}$ since $\partial_i G$ belongs to the Hardy space $\mathcal{H}^1$. □

We prove now that, in a first approximation and for some classes of initial data, the solutions of the incompressible Navier–Stokes equations behave like the solutions of the heat equation with the same initial data.

THEOREM 2.3. *Let u be a weak solution of* (NS) *in dimension two with initial data $u_0 \in L^p \cap L^2(\mathbb{R}^2)$, $1 \leq p \leq 2$, such that* div $u_0 = 0$. *Then for any $q \geq p$,*

(i) *if $1 < p < 2$,*

$$\|G(t) * u_0 - u(t)\|_q \leq Ct^{-\frac{1}{p}+\frac{1}{q}} t^{-\frac{1}{p}+\frac{1}{2}}, \quad t > 0,$$

(ii) *if $p = 1$,*

$$\|G(t) * u_0 - u(t)\|_q \leq Ct^{-1+\frac{1}{q}} t^{-\frac{1}{2}} \log\, t, \quad t > 0,$$

(iii) *if $p = 2$,*

$$t^{\frac{1}{2}-\frac{1}{q}} \|u(t)\|_q \to 0 \quad as\, t \to \infty.$$

Remark 2.3. All the estimates were known when $q = 2$ for any $1 \leq p \leq 2$ ([26], [17]). We can replace the powers of t by powers of $t+1$ (and also log t by log $(t+1)$) when $p \leq q \leq 2$. In case $p = 2$, (iii) is known to hold for the solutions of the heat equation; hence, it holds for the difference $G(t) * u_0 - u(t)$, but this gives no extra information.

Proof. (i) From the integral equation we get

$$\|G(t) * u_0 - u(t)\|_q \leq C \int_0^t (t-s)^{-\frac{1}{2}-\frac{1}{r}+\frac{1}{q}} s^{\frac{-2}{p}+\frac{1}{r}}$$

for $q \geq r > 1$ and $2r \geq p$. As in the proof of (ii) below, we split the integral in two intervals $[0, \frac{t}{2}]$ and $[\frac{t}{2}, t]$. By choosing an adequate r we conclude that

$$\|G(t) * u_0 - u(t)\|_q \leq C t^{-\frac{1}{p}+\frac{1}{q}} t^{-\frac{1}{p}+\frac{1}{2}}.$$

(ii) Taking norms in the integral equation, we get

$$\begin{aligned} \|G(t) * u_0 - u(t)\|_q \leq & \int_0^t \|\partial_i G(t-s) * u^i u(s)\|_q ds \\ & + \int_0^t \|\partial_i G(t-s) * \partial_j \nabla E_2 * u^i u^j(s)\|_q ds. \end{aligned}$$

Since

$$\|\partial_i G(t-s) * \partial_j \nabla E_2\|_{r'} \leq \begin{cases} C\|\partial_i G(t-s)\|_{r'}, & 1 < r' < \infty, \\ C\|\partial_i G(t-s)\|_{\mathcal{H}^1}, & r' = 1, \end{cases}$$

both integrals are bounded by

$$C \int_0^t (t-s)^{-\frac{1}{2}-\frac{1}{r}+\frac{1}{q}} s^{-2+\frac{1}{r}}$$

for $q \geq r,\ 2r \geq 1,\ r \geq 1$. We split this integral as follows:

(a)

$$\int_{\frac{t}{2}}^t (t-s)^{-\frac{1}{2}-\frac{1}{r}+\frac{1}{q}} s^{-2+\frac{1}{r}} \ \leq\ C\, t^{\frac{1}{2}+\frac{1}{q}-2},$$

choosing r such that $\frac{1}{2} - \frac{1}{r} + \frac{1}{q} > 0$, that is, $q \geq r > \frac{2q}{q+2}$.

(b)

$$\int_0^{\frac{t}{2}} (t-s)^{-\frac{1}{2}-\frac{1}{r}+\frac{1}{q}} s^{-2+\frac{1}{r}} \ \leq\ C\, t^{\frac{1}{2}+\frac{1}{q}-2} \log\, t,$$

choosing $r = 1$. Therefore, (ii) holds.

(iii) It is known that $\|u(t)\|_2 \to 0$ as $t \to \infty$. Interpolating and taking into account that $t^{\frac{1}{2}-\frac{1}{r}}\|u(t)\|_r \leq C$ (see Theorem 1.1), we get

$$t^{\frac{1}{2}-\frac{1}{q}}\|u(t)\|_q \leq C(\|u(t)\|_2)^{1-\alpha}(t^{\frac{1}{2}-\frac{1}{r}}\|u(t)\|_r)^\alpha \leq C(\|u(t)\|_2)^{1-\alpha}$$

for any $2 < q < r$, which yields the result. $\square$

3. Dimension two: Second term. Let u be again a weak solution of (NS) and v a solution of

$$(\mathcal{L}_2) \quad \begin{cases} v_t - \Delta v = -h^i \partial_i h - \partial_j \nabla E_2 * h^i \partial_i h^j & \text{in } \mathbb{R}^+ \times \mathbb{R}^2, \\ \operatorname{div} v = 0 & \text{in } \mathbb{R}^+ \times \mathbb{R}^2, \\ v(x,0) = u_0, \ \operatorname{div} u_0 = 0 & \text{in } \mathbb{R}^2, \end{cases}$$

where $h(t) = G(t) * u_0$ so that v can be written as h plus a corrector term. Let us see whether the difference $u(t) - v(t)$ tends to 0 faster than $u(t) - h(t)$. We assume again that $u_0 \in L^2(\mathbb{R}^2) \cap L^p(\mathbb{R}^2)$, $1 \leq p < 2$.

Taking norms in the integral equation satisfied by the difference $u(t) - v(t)$ we get

$$\begin{aligned}
\|u(t) - v(t)\|_q \leq & \, \| \int_0^t \partial_i G(t-s) * (u^i(u-h) + h(u^i - h^i))(s) ds \|_q \\
& + \| \int_0^t \partial_i G(t-s) * \partial_j \nabla E_2 * (u^i(u-h) + h(u^i - h^i))(s) ds \|_q \\
\leq & \int_0^t \|\partial_i G(t-s)\|_{r'} (\|u(s)\|_{2r} + \|h(s)\|_{2r}) \|(u-h)(s)\|_{2r} ds \\
& + \int_0^t \|\partial_i G(t-s) * \partial_j \nabla E_2\|_{r'} (\|u(s)\|_{2r} + \|h(s)\|_{2r}) \|(u-h)(s)\|_{2r} ds
\end{aligned}$$

for $q \geq r$, $r' \geq 1$ such that $\frac{1}{r} + \frac{1}{r'} = \frac{1}{q}$. In view of the estimates

$$\begin{aligned}
\|u(t)\|_{2r} &\leq Ct^{-\frac{1}{p}+\frac{1}{2r}}, \\
\|h(t)\|_{2r} &\leq Ct^{-\frac{1}{p}+\frac{1}{2r}}, \\
\|(u-h)(t)\|_{2r} &\leq Ct^{-\frac{1}{p}+\frac{1}{2r}} t^{-\frac{1}{p}+\frac{1}{2}},
\end{aligned}$$

valid when $q \geq p$, $q \geq r \geq 1$, and $2r \geq p$, and the fact that

$$\|\partial_i G(t-s) * \partial_j \nabla E_2\|_{r'} \leq \begin{cases} C\|\partial_i G(t-s)\|_{r'}, & 1 < r' < \infty, \\ C\|\partial_i G(t-s)\|_{\mathcal{H}^1}, & r' = 1, \end{cases}$$

we get

$$\|u(t) - v(t)\|_q \leq C \int_0^t (t-s)^{-\frac{1}{2}-\frac{1}{r}+\frac{1}{q}} s^{-\frac{2}{p}+\frac{1}{r}} s^{-\frac{1}{p}+\frac{1}{2}} ds.$$

We split the integral as follows:

(a)

$$\int_{\frac{t}{2}}^t (t-s)^{-\frac{1}{2}-\frac{1}{r}+\frac{1}{q}} s^{-\frac{3}{p}+\frac{1}{r}+\frac{1}{2}} \leq Ct^{-\frac{3}{p}+\frac{1}{q}+1}$$

if $r > \frac{2q}{q+2}$;

(b)

$$\int_0^{\frac{t}{2}} (t-s)^{-\frac{1}{2}-\frac{1}{r}+\frac{1}{q}} s^{-\frac{3}{p}+\frac{1}{r}+\frac{1}{2}} \leq Ct^{-\frac{3}{p}+\frac{1}{q}+\frac{1}{2}}$$

if $-\frac{3}{p} + \frac{1}{r} + \frac{3}{2} > 0$, that is, $r \leq q$ and $1 \leq r < \frac{2p}{3(2-p)}$, provided that $p > \frac{6}{5}$.

Therefore,

$$\|(u-v)(t)\|_q \leq Ct^{-\frac{1}{p}+\frac{1}{q}} t^{-\frac{1}{p}+\frac{1}{2}} t^{-\frac{1}{p}+\frac{1}{2}}$$

if $2 > p > \frac{6}{5}$. This decay is faster than that corresponding to $u(t) - h(t)$.

Thus, we have proved the following theorem.

THEOREM 3.1. *Let u be a solution of the two-dimensional* (NS) *with initial data* $u_0 \in L^p \cap L^2(\mathbb{R}^2)$ $\frac{6}{5} < p < 2$ *and v a solution of* $(\mathcal{L}_2)$. *Then for any $q \geq p$, we have*

$$\|(u-v)(t)\|_q \leq Ct^{-\frac{3}{p}+\frac{1}{q}+1}.$$

4. Dimension two: Explicit second term. In this section, we shall obtain an explicit approximation up to the second term of both u and v. Let us consider the following three problems:

$$(\mathcal{P}_1) \qquad \begin{cases} w_t - \Delta w = 0 & \text{in } \mathbb{R}^+ \times \mathbb{R}^2, \\ \operatorname{div} w = 0 & \text{in } \mathbb{R}^+ \times \mathbb{R}^2, \\ w(x,0) = u_0, \ \operatorname{div} u_0 = 0 & \text{in } \mathbb{R}^2, \end{cases}$$

$$(\mathcal{P}_2) \qquad \begin{cases} w_t - \Delta w = -f^i \partial_i f & \text{in } \mathbb{R}^+ \times \mathbb{R}^2, \\ \operatorname{div} w = 0 & \text{in } \mathbb{R}^+ \times \mathbb{R}^2, \\ w(x,0) = 0 & \text{in } \mathbb{R}^2, \end{cases}$$

$$(\mathcal{P}_3) \qquad \begin{cases} w_t - \Delta w = -\partial_j \nabla E_2 * f^i \partial_i f^j & \text{in } \mathbb{R}^+ \times \mathbb{R}^2, \\ \operatorname{div} w = 0 & \text{in } \mathbb{R}^+ \times \mathbb{R}^2, \\ w(x,0) = 0 & \text{in } \mathbb{R}^2, \end{cases}$$

where $w = (w^1, w^2)$ and $f = (f^1, f^2)$. We assume that $u_0 \in L^1(\mathbb{R}^2)$, $\partial_j \nabla E_2 * f^i \partial_i f^j$, $f^i \partial_i f \in L^1(0,\infty; L^1(\mathbb{R}^2))$, and $\operatorname{div} f = 0$. Let us denote by w_i the solution of each $(\mathcal{P}_i)$. Then, $w = w_1 + w_2 + w_3$ is a solution of

$$(\mathcal{P}) \qquad \begin{cases} w_t - \Delta w = -f^i \partial_i f - \partial_j \nabla E_2 * f^i \partial_i f^j & \text{in } \mathbb{R}^+ \times \mathbb{R}^2, \\ \operatorname{div} w = 0 & \text{in } \mathbb{R}^+ \times \mathbb{R}^2, \\ w(x,0) = u_0, \ \operatorname{div} u_0 = 0 & \text{in } \mathbb{R}^2, \end{cases}$$

and satisfies the integral equation

$$w(t) = G(t) * u_0 - \int_0^t G(t-s) * f^i \partial_i f(s) ds - \int_0^t G(t-s) * \nabla \partial_j E_2 * f^i \partial_i f^j(s) ds.$$

Therefore, we may think of the solutions u of (NS) or v of $(\mathcal{L}_2)$ as being the sum of the following three terms:

- a term w_1 which solves $(\mathcal{P}_1)$,
- a term w_2 which solves $(\mathcal{P}_2)$ with $f = u$ (resp., $f = h$),
- a term w_3 which solves $(\mathcal{P}_3)$ with $f = u$ (resp., $f = h$).

We will study the asymptotic behavior of the solutions of these problems in order to get information on u and v with initial data $u_0 \in L^1 \cap L^2(\mathbb{R}^2)$.

4.1. Problem $(\mathcal{P}_1)$. The solution of this heat equation is $w_1 = G(t) * u_0$, whose asymptotic behavior is well known. In case $u_0 \in L^1(\mathbb{R}^2)$,

$$t^{1-\frac{1}{q}} \|G(t) * u_0 - MG(t)\|_q \to 0 \quad \text{as } t \to \infty$$

for any $1 \leq q \leq \infty$, where $M = \int_{\mathbb{R}^2} u_0$. Thus, the first term in the development of w_1 when $t \to \infty$ is $MG(t)$. If $u_0 \in L^1(1+|x|;\mathbb{R}^2)$, we know further that

$$t^{1-\frac{1}{q}+\frac{1}{2}}\|G(t) * u_0 - MG(t)\|_q \leq C\|u_0\|_{L^1(|x|,\,\mathbb{R}^2)}.$$

In fact, setting $M^i = \int_{\mathbb{R}^2} x^i u_0$, we have

$$t^{\frac{3}{2}-\frac{1}{q}}\|G(t) * u_0 - MG(t) + M^i\partial_i G(t)\|_q \to 0 \quad \text{as } t \to \infty.$$

We can obtain more terms when more moments of the initial data are finite (see [7]).

In case $u_0 \in L^p(\mathbb{R}^2)$, $1 < p < \infty$, we have

$$t^{\frac{1}{p}-\frac{1}{q}}\|G(t) * u_0\|_q \to 0 \quad \text{as } t \to \infty$$

for any $p \leq q \leq \infty$, so that the first term in the development is 0.

This convergence is clear when $u_0 \in \mathcal{D}(\mathbb{R}^n)$, since $u_0 \in L^1(\mathbb{R}^n)$ and then $\|G(t) * u_0\|_p \leq C(1+t)^{-1+\frac{1}{p}}$. Thanks to the fact that $\|G(t) * u_0\|_p$ decreases as time grows, we can extend the result to any $u_0 \in L^p(\mathbb{R}^n)$ by density. Given $u_0 \in L^p(\mathbb{R}^n)$, we take a sequence $u_{0,k} \subset \mathcal{D}(\mathbb{R}^n)$ such that $u_{0,k} \to u_0$ in $L^p(\mathbb{R}^n)$. Let u_k be the solution of the heat equation with data $u_{0,k}$. Then

$$\|u(t)\|_p \leq \|u_k(t)\|_p + \|u(t) - u_k(t)\|_p \leq \|u_k(t)\|_p + \|u_0 - u_{0,k}\|_p.$$

Given $\varepsilon > 0$, we can choose k large enough to have

$$\|u_0 - u_{0,k}\|_p \leq \frac{\varepsilon}{2}$$

and, fixing that k, we get

$$\|u_k(t)\|_p \leq \frac{\varepsilon}{2}$$

for $t \geq t_\varepsilon$. Therefore, $\|u(t)\|_p \to 0$ as $t \to \infty$. Once this is proved, it is clear that

$$t^{\frac{1}{p}-\frac{1}{q}}\|G(t) * u_0\|_q \to 0 \quad \text{as } t \to \infty$$

for any $p \leq q \leq \infty$.

4.2. Problem ($\mathcal{P}_2$). In order to describe the asymptotic behavior of w_2, we are going to use a scaling technique. In the following, we shall drop the subscript 2 and write only w. Since we want to take $f = u$, u being a weak solution of (NS) with data $u_0 \in L^1 \cap L^2(\mathbb{R}^2)$, we shall assume that

$$f(t) \in BC(0,\infty;L^2(\mathbb{R}^2)), \quad \|f(t)\|_2 \leq C(1+t)^{\frac{-1}{2}}, \quad t \geq 0,$$

so that we can rewrite the integral expression for w,

$$w(t) = -\int_0^t \partial_i G(t-s) * f^i f(s) ds.$$

By rescaling, we see that the functions $w_\lambda(t,x) = \lambda^2 w(\lambda^2 t, \lambda x)$ satisfy

$$w_\lambda(t) \;=\; -\lambda^{-1}\int_0^t \partial_i G(t-s) * f^i_\lambda f_\lambda(s) ds.$$

We remark that

$$\|w_\lambda(t)\|_q = \lambda^{2(1-\frac{1}{q})}\|w(\lambda^2 t)\|_q.$$

Thus, if we want to make precise the asymptotic behavior of w when $t \to \infty$, it suffices to find a function g such that

$$\|w_\lambda(t) - g_\lambda(t)\|_q \delta(\lambda) \to 0 \quad \text{as } \lambda \to \infty$$

for some $\delta(\lambda)$ tending to ∞ as $\lambda \to \infty$, where $g_\lambda(t,x) = \lambda^2 g(\lambda^2 t, \lambda x)$. That implies

$$\|w(t) - g(t)\|_q t^{1-\frac{1}{q}} \delta(t^{\frac{1}{2}}) \to 0 \quad \text{as } t \to \infty.$$

It is easy to prove that

$$\|w_\lambda\|_q \;=\; \lambda^{-1}\| \int_0^t \partial_i G(t-s) * f_\lambda^i f_\lambda(s) ds\|_q$$

is bounded by $C \ \log \lambda^2 \ \lambda^{-1}$. Therefore, the same kind of bound should hold for g_λ, but the difference $\|g_\lambda(t) - w_\lambda(t)\|_q$ should go to 0 faster. Under certain conditions, it is possible to take $g(t) = (\frac{M^i M}{2}) \log \ t \ \partial_i G(t)$ with $M = (M^1, M^2)$ and $\delta(t) = t^{\frac{1}{2}} / \log \ t$. More precisely, we prove the following theorem.

THEOREM 4.1. *Let w be a solution of $(\mathcal{P}_2)$ and $q \geq 1$. Let us assume that*

(i)

$$(1+t)^{\frac{1}{2}+\varepsilon}\|f(t) - G * u_0(t)\|_2 \leq C, \quad t \geq 0$$

for some $\varepsilon > 0$, and

(ii)

$$\|f(t)\|_{2r} \;\leq\; C(1+t)^{-1+\frac{1}{2r}}, \qquad t \geq 0,$$

where $q \geq r > \frac{2q}{q+2}$.

Then

$$\frac{t^{\frac{3}{2}-\frac{1}{q}}}{\log(t)} \left\| w(t) - \log \ t \ \partial_i G(t) \left(\frac{M^i M}{2}\right) \right\|_q \to 0 \quad \textit{as } t \to \infty,$$

where $M = (M^1, M^2) = \int_{\mathbb{R}^2} u_0(x) dx$, provided that $|x|^{\frac{\alpha}{2}} u_0 \in L^1$ and $|x|^{\frac{\alpha}{2}} u_0 \in L^2$ for some $\alpha \geq 2$.

Remark 4.1. It follows from (i) that (ii) holds, replacing $2r$ by 2, and that

$$t^{\frac{1}{2}}\|f(t) - MG(t)\|_2 \to 0 \quad \text{as } t \to \infty.$$

Proof. Setting

$$g_\lambda(1) = \lambda^{-1} \ \log \ \lambda^2 \ \partial_i G(1) \left(\frac{M^i M}{2}\right) = \left(\log(\cdot) \ \partial_i G(\cdot) \left(\frac{M^i M}{2}\right)\right)_\lambda (1),$$

we see that

$$\frac{t^{1+\frac{1}{2}-\frac{1}{q}}}{\log(t)} \left\| \int_0^t \partial_i G(t-s) * f^i f(s) ds - \log(t) \ \partial_i G(t) \left(\frac{M^i M}{2}\right) \right\|_q \to 0 \quad \text{as } t \to \infty$$

is equivalent to

$$\left\| \lambda^{-1} \int_0^1 \partial_i G(1-s) * f_\lambda^i f_\lambda(s) ds - g_\lambda(1) \right\|_q \;=\; o(1) \log(\lambda^2) \lambda^{-1} \quad \text{as } \lambda \to \infty.$$

Making the change of variables $z(s,y) = e^s f(e^s - 1, e^{\frac{s}{2}} y)$, it follows from (i) that

$$\lim_{\sigma\to\infty}\left\{\frac{1}{\sigma}\int_0^\sigma\int_{\mathbb{R}^2} z^i z\right\} = M^iM\int G(1)^2 = \frac{M^iM}{2}.$$

Thus, we must prove

$$\left\|\int_0^1 \partial_i G(1-s) * \frac{f^i_\lambda f_\lambda(s)}{\log\,\lambda^2}ds - \partial_i G(1)\lim_{\sigma\to\infty}\left\{\frac{1}{\sigma}\int_0^\sigma\int_{\mathbb{R}^2} z^i z\right\}\right\|_q \to 0 \quad \text{as } \lambda\to\infty$$

We remark first that

$$\begin{aligned}
&\int_0^1\left(\partial_i G(1-s) * \frac{f^i_\lambda f_\lambda(s)}{\log\,\lambda^2}\right)(x)ds\\
&\qquad = \int_0^1\int_{\mathbb{R}^2}\partial_i G(1-s, x-y)\frac{f^i f(\lambda^2 s, \lambda x)}{\log\,\lambda^2}\lambda^4 dsdy\\
&\qquad = \int_0^{\lambda^2}\int_{\mathbb{R}^2}\partial_i G\left(1-\frac{s}{\lambda^2}, x-\frac{y}{\lambda}\right)\frac{f^i f(s,y)}{\log\,\lambda^2}dsdy\\
&\qquad = \int_0^{\log(\lambda^2+1)}\int_{\mathbb{R}^2}\partial_i G\left(1-\frac{e^s-1}{\lambda^2}, x-\frac{ye^{\frac{s}{2}}}{\lambda}\right)\frac{z^i z(s,y)}{\log\,\lambda^2}dsdy.
\end{aligned}$$

Therefore, it is enough to prove that

$$\int_0^{\lambda^2}\int_{\mathbb{R}^2}\partial_i G\left(1-\frac{s}{\lambda^2}, x-\frac{y}{\lambda}\right)\frac{f^i f(s,y)}{\log\,\lambda^2}dsdy \;-\; \partial_i G(1,x)\int_0^{\lambda^2}\int_{\mathbb{R}^2}\frac{f^i f(s,y)}{\log\,\lambda^2}dsdy$$

and

$$\partial_i G(1,x)\int_0^{\lambda^2}\int_{\mathbb{R}^2}\frac{f^i f(s,y)}{\log\,\lambda^2}dsdy \;-\; \partial_i G(1,x)\lim_{\sigma\to 0}\left\{\frac{1}{\sigma}\int_0^\sigma\int_{\mathbb{R}^2} z^i z(s,y)\right\}$$

tend to 0 in L^q_x when $\lambda\to\infty$. Since

$$\int_0^{\lambda^2}\int_{\mathbb{R}^2}\frac{f^i f(s,y)}{\log\,\lambda^2}dsdy = \frac{1}{\log\,\lambda^2}\int_0^{\log(1+\lambda^2)}\int_{\mathbb{R}^2} z^i z(s,y)dsdy,$$

the last convergence is clear. We must prove only the first one. In order to do that, we split the integral as follows:

$$\begin{aligned}
&\int_0^{\lambda^2}\int_{\mathbb{R}^2}\left(\partial_i G\left(1-\frac{s}{\lambda^2}, x-\frac{y}{\lambda}\right) - \partial_i G(1,x)\right)\frac{f^i f(s,y)}{\log\,\lambda^2}dsdy\\
&\qquad = \int_0^{\lambda^2\delta}\int_{|y|\le\lambda\delta}\left(\partial_i G\left(1-\frac{s}{\lambda^2}, x-\frac{y}{\lambda}\right) - \partial_i G(1,x)\right)\frac{f^i f(s,y)}{\log\,\lambda^2}dsdy\\
&+ \int_{\lambda^2\delta}^{\lambda^2}\int_{\mathbb{R}^2}\partial_i G\left(1-\frac{s}{\lambda^2}, x-\frac{y}{\lambda}\right)\frac{f^i f(s,y)}{\log\,\lambda^2}dsdy + \int_{\lambda^2\delta}^{\lambda^2}\int_{\mathbb{R}^2}\partial_i G(1,x)\frac{f^i f(s,y)}{\log\,\lambda^2}dsdy\\
&+ \int_0^{\lambda^2\delta}\int_{|y|>\lambda\delta}\partial_i G\left(1-\frac{s}{\lambda^2}, x-\frac{y}{\lambda}\right)\frac{f^i f(s,y)}{\log\,\lambda^2}dsdy\\
&+ \int_0^{\lambda^2\delta}\int_{|y|>\lambda\delta}\partial_i G(1,x)\frac{f^i f(s,y)}{\log\,\lambda^2}dsdy\\
&\qquad = I^1_{\delta,\lambda} + I^2_{\delta,\lambda} + J^2_{\delta,\lambda} + I^3_{\delta,\lambda} + J^3_{\delta,\lambda}.
\end{aligned}$$

Estimate $I^1_{\delta,\lambda}$.

$$\|I^1_{\delta,\lambda}\|_{L^q_x} \le C \int_0^{\lambda^2\delta} \int_{|y|<\lambda\delta} \left\| \partial_i G\left(1-\frac{s}{\lambda^2}, x-\frac{y}{\lambda}\right) - \partial_i G(1,x)\right\|_{L^q_x} \frac{|f^i||f|(s,y)}{\log \lambda^2} dsdy.$$

Thanks to the continuity of translations in L^q_x and the continuity with respect to t, given $\varepsilon > 0$ we can choose $\delta > 0$ such that

$$\operatorname*{Sup}_{s\le\lambda^2\delta,\ |y|\le\lambda\delta} \left\| \partial_i G\left(1-\frac{s}{\lambda^2}, x-\frac{y}{\lambda}\right) - \partial_i G(1,x)\right\|_{L^q_x} \le \varepsilon.$$

Therefore,

$$\|I^1_{\delta,\lambda}\|_{L^q_x} \le \varepsilon \int_0^{\lambda^2\delta}\int_{|y|\le\lambda\delta} \frac{|f|^2(s,y)}{\log\lambda^2} dsdy \le \varepsilon \int_0^{\log(1+\lambda^2\delta)} \int_{\mathbb{R}^2} \frac{|z|^2(s,y)}{\log\lambda^2} dsdy \le \varepsilon C$$

uniformly with respect to λ, since

$$\|z(s)\|_2 = e^{\frac{s}{2}}\ \|f(e^s-1)\|_2 \le C.$$

Estimate $J^2_{\delta,\lambda}$.

$$\begin{aligned}\|J^2_{\delta,\lambda}\|_{L^q_x} &\le C\|\nabla G(1)\|_{L^q_x} \int_{\lambda^2\delta}^{\lambda^2}\int_{\mathbb{R}^2} \frac{|f|^2(s,y)}{\log\lambda^2} dsdy \\ &\le \|\nabla G(1)\|_{L^q_x} \int_{\log(1+\lambda^2\delta)}^{\log(1+\lambda^2)} \int_{\mathbb{R}^2} \frac{|z|^2(s,y)}{\log\lambda^2} dsdy \le C\frac{\log(\frac{1+\lambda^2}{1+\lambda^2\delta})}{\log\lambda^2}.\end{aligned}$$

Therefore, it tends to 0 when $\lambda \to \infty$ and δ is fixed.

Estimate $I^2_{\delta,\lambda}$.

$$\|I^2_{\delta,\lambda}\|_{L^q_x} \le \frac{1}{\log(\lambda^2)} \int_{\log(1+\lambda^2\delta)}^{\log(1+\lambda^2)} \left\| \int_{\mathbb{R}^2} \partial_i G\left(1-\frac{e^s-1}{\lambda^2}, x-\frac{e^{\frac{s}{2}}y}{\lambda}\right) z^i z(y,s) dy\right\|_{L^q_x} ds.$$

We first observe that

$$\begin{aligned}&\int_{\mathbb{R}^2} \partial_i G\left(1-\frac{e^s-1}{\lambda^2}, x-\frac{e^{\frac{s}{2}}y}{\lambda}\right) z^i z(y,s)dy \\ &= \left(\partial_i G\left(1-\frac{e^s-1}{\lambda^2}, y\right) *_y (\lambda e^{\frac{-s}{2}})^2\ z^i z(s,\lambda e^{\frac{-s}{2}}y)\right)(x).\end{aligned}$$

Thus,

$$\begin{aligned}&\left\| \int_{\mathbb{R}^2} \partial_i G\left(1-\frac{e^s-1}{\lambda^2}, x-\frac{e^{\frac{s}{2}}y}{\lambda}\right) z^i z(y,s)dy\right\|_{L^q_x} \\ &\le C\left(1-\frac{e^s-1}{\lambda^2}\right)^{-(\frac{1}{r}-\frac{1}{q})-\frac{1}{2}} \|(\lambda e^{\frac{-s}{2}})^2 z^i z(s,\lambda e^{\frac{-s}{2}}y)\|_{L^r_x}\end{aligned}$$

for $r \le q$. Since

$$\|z(s)\|_{2r} = e^{\frac{s}{2}(2-\frac{1}{r})}\ \|f(e^s-1)\|_{2r}\ \le C,$$

it follows that

$$\|(\lambda e^{\frac{-s}{2}})^2 z^i z(\lambda e^{\frac{-s}{2}} y, s)\|_{L^r_x} \;\leq\; \|z(s)\|^2_{L^{2r}_x}(\lambda e^{\frac{-s}{2}})^{2(1-\frac{1}{r})} \;\leq\; C(\lambda e^{\frac{-s}{2}})^{2(1-\frac{1}{r})}.$$

Therefore,

$$\begin{aligned}\|I^2_{\delta,\lambda}\|_{L^q_x} &\leq \frac{C\lambda^{2(1-\frac{1}{r})}}{\log(\lambda^2)} \int_{\log(1+\lambda^2\delta)}^{\log(1+\lambda^2)} \left(1-\frac{e^s-1}{\lambda^2}\right)^{-(\frac{1}{r}-\frac{1}{q})-\frac{1}{2}} (e^{\frac{-s}{2}})^{2(1-\frac{1}{r})} ds \\ &\leq \frac{C\lambda^{2(1-\frac{1}{r})}(\lambda^2\delta+1)^{(-1+\frac{1}{r})}}{\log(\lambda^2)} \int_{\log(1+\lambda^2\delta)}^{\log(1+\lambda^2)} \left(1-\frac{e^s-1}{\lambda^2}\right)^{-(\frac{1}{r}-\frac{1}{q})-\frac{1}{2}} ds \\ &\leq \frac{C}{\log\ \lambda^2}\left(\frac{\lambda^2}{1+\delta\lambda^2}\right)^{2-\frac{1}{r}} \int_\delta^1 (1-t)^{-(\frac{1}{r}-\frac{1}{q})-\frac{1}{2}} dt \;\leq\; \frac{C}{\log\ \lambda^2},\end{aligned}$$

where C is a constant depending on δ but not on λ, taking $r > \frac{2q}{q+2}$ in order to have $-(\frac{1}{r}-\frac{1}{q})+\frac{1}{2} > 0$.

We conclude that

$$\|I^2_{\delta,\lambda}\|_{L^q_x} \;\leq\; \frac{C}{\log \lambda^2} \longrightarrow 0 \quad \text{as } \lambda \longrightarrow \infty$$

for a fixed δ.

Estimate $J^3_{\delta,\lambda}$.

$$\|J^3_{\delta,\lambda}\|_{L^q_x} \;\leq\; \|\nabla G(1)\|_{L^q_x} \int_0^{\lambda^2\delta} \int_{|y|\geq\delta\lambda} \frac{|f|^2(s,y)}{\log\ \lambda^2}.$$

We must prove that

$$\int_0^{\lambda^2\delta} \int_{|y|\geq\delta\lambda} \frac{|f|^2(s,y)}{\log\ \lambda^2} \longrightarrow 0 \quad \text{as } \lambda \longrightarrow \infty$$

for δ fixed. For any $\alpha \geq 2$ we have

$$\||y|^{\frac{\alpha}{2}} G * u_0(t,y)\|_2^2 \;\leq\; C(t^{\frac{\alpha}{2}-1} + (t+1)^{-1}) \;\leq\; C(t^{\frac{\alpha}{2}-1})$$

provided that u_0, $|x|^{\frac{\alpha}{2}} u_0 \in L^1$ and $|x|^{\frac{\alpha}{2}} u_0 \in L^2$. It follows that if $f = G * u_0$,

$$\int_0^{\lambda^2\delta} \int_{|y|\geq\delta\lambda} \frac{|y|^\alpha |f|^2(y,s)}{\lambda^\alpha\ \log\ \lambda^2} \;\leq\; C\frac{(\lambda^2\delta)^{\frac{\alpha}{2}}}{\lambda^\alpha\ \log\ \lambda^2} \;\leq\; \frac{C}{\log\ \lambda^2}.$$

Since $\|f(t) - G * u_0(t)\|_2 \leq C(1+t)^{\frac{-1}{2}-\varepsilon}$,

$$\int_0^{\lambda^2\delta} \int_{|y|\geq\delta\lambda} \frac{|f(s,y) - G * u_0(s,y)|^2}{\log\ \lambda^2} \;\leq\; \frac{C}{\log\ \lambda^2}.$$

Therefore, for a fixed δ,

$$\|J^3_{\delta,\lambda}\|_{L^q_x} \longrightarrow 0 \quad \text{as } \lambda \longrightarrow \infty.$$

Estimate $I^3_{\lambda,\delta}$.

$$\|I^3_{\lambda,\delta}\|_{L^q_x} \le \int_0^{\lambda^2\delta} \left\| \int_{|y|\ge\delta\lambda} \partial_i G\left(1-\frac{s}{\lambda^2}, x-\frac{y}{\lambda}\right) \frac{f^i f(s,y)}{\log\ \lambda^2} \right\|_{L^q_x} .$$

As in Estimate $I^2_{\lambda,\delta}$, but taking $r = 1$, we get

$$\begin{aligned}\|I^3_{\lambda,\delta}\|_{L^q_x} &\le C \int_0^{\lambda^2\delta} \left(1-\frac{s}{\lambda^2}\right)^{-1+\frac{1}{q}-\frac{1}{2}} \left\|\frac{f^i f(s,y)}{\log\ \lambda^2}\right\|_{L^1(|y|\ge\lambda\delta)} \\ &\le C \int_0^{\lambda^2\delta} \int_{|y|\ge\delta\lambda} \frac{|f|^2(s,y)}{\log\ \lambda^2} dyds\end{aligned}$$

and we finish in the same way as in estimate $J^3_{\lambda,\delta}$. □

4.3. Problem ($\mathcal{P}_3$). In the following, we shall drop the subscript 3 and write only w. As before, we shall assume that

$$f(t) \in BC(0,\infty; L^2(\mathbb{R}^2)), \quad \|f(t)\|_2 \le C(1+t)^{\frac{-1}{2}}, \quad t \ge 0,$$

so that we can rewrite the integral expression for w,

$$w(t) \ = \ -\int_0^t \partial_i G(t-s) * \nabla\partial_j E_2 * f^i f^j(s) ds.$$

By rescaling, we see that the functions $w_\lambda(t,x) = \lambda^2 w(\lambda^2 t, \lambda x)$ satisfy

$$w_\lambda(t) \ = \ -\lambda^{-1} \int_0^t \partial_i G(t-s) * \nabla\partial_j E_2 * f^i_\lambda f^j_\lambda(s) ds.$$

Since $\|\partial_i G(t-s) * \nabla\partial_j E_2\|_1 \le C\|\partial_i G(t-s)\|_{\mathcal{H}^1}$, it is easy to prove that

$$\| \int_0^t \partial_i G(t-s) * \nabla\partial_j E_2 * f^i_\lambda f^j_\lambda(s) ds\|_q$$

is bounded by $C\ \log\lambda^2\ \lambda^{-1}$.

It suffices to rewrite step by step the proof of Theorem 4.1, replacing the L^q norms of $\partial_i G(t-s)$ by the L^q norms of $\partial_i G(t-s) * \nabla\partial_j E_2$ to get the following theorem.

THEOREM 4.2. *Let w be a solution of ($\mathcal{P}_3$) and $q \ge 1$. Let us assume that*

(i)

$$(1+t)^{\frac{1}{2}+\varepsilon}\|f(t) - G * u_0(t)\|_2 \le C, \quad t \ge 0$$

for some $\varepsilon > 0$, and

(ii)

$$\|f(t)\|_{2r} \ \le \ C(1+t)^{-1+\frac{1}{2r}}, \quad t \ge 0,$$

where $q \ge r > \frac{2q}{q+2}$. Then

$$\frac{t^{\frac{3}{2}-\frac{1}{q}}}{\log\ t} \left\| w(t) - \left(\frac{M^i M^j}{2}\right) \log\ t\ \partial_i G(t) * \nabla\partial_j E_2 \right\|_q \to 0 \quad as\, t \to \infty,$$

provided that $u_0 \in L^1 \cap L^2(\mathbb{R}^2, |x|)$.

Remark 4.2. If we take u to be a solution of (NS) and v a solution of $(\mathcal{L}_2)$ with initial data $u_0 \in (L^1 \cap L^2)(\mathbb{R}^2, 1+|x|) \cap L^{2r}(\mathbb{R}^2)$, then we can apply the theorem with $f = u$ or $f = G(t) * u_0$ and $M = \int_{\mathbb{R}^2} u_0(x)dx$ to obtain the behavior of the third term, which is the same for both of them.

4.4. Conclusion. Putting together the previous results we obtain the following theorem.

THEOREM 4.3. *Let u be a solution of the two-dimensional* (NS) *with initial data $u_0 \in (L^1 \cap L^2)(\mathbb{R}^2, 1+|x|)$ such that* div $u_0 = 0$ *and set $M = \int_{\mathbb{R}^2} u_0(x)dx$. Then for a given $q \geq 1$,*

$$\frac{t^{\frac{3}{2}-\frac{1}{q}}}{\log(t)} \left\| u(t) - MG(t) + \log(t) \, \left(\partial_i G(t) \left(\frac{M^i M}{2} \right) + \partial_i G(t) * \nabla \partial_j E_2 \left(\frac{M^i M^j}{2} \right) \right) \right\|_q$$

tends to 0 *as t goes to infinity provided that $u_0 \in L^{2r}(\mathbb{R}^2)$ for some $q \geq r > \frac{2q}{q+2}$.*

Remark 4.3. When $u_0 \in (L^1 \cap L^2)(\mathbb{R}^2, 1+|x|)$ we can take $r = 1$ and then the result holds for $1 \leq q < 2$.

5. Higher dimensions: $n > 2$. In what follows, we shall be concerned with solutions u of (NS) taking data $u_0 \in L^p \cap L^n$ such that div $u_0 = 0$, $\|u_0\|_n$ is small and $1 \leq p \leq n$. For that kind of data, unique global strong solutions are known to exist.

5.1. Decay estimates. We improve Theorem 1.2(i) here by proving the following result.

THEOREM 5.1. *Let u be a strong solution of* (NS) *with data $u_0 \in L^p \cap L^n(\mathbb{R}^n)$, $1 \leq p \leq n$, of small L^n norm such that* div $u_0 = 0$. *Then*

$$\|u(t)\|_q \leq Ct^{(-\frac{1}{p}+\frac{1}{q})\frac{n}{2}}$$

if $q \geq p$.

Proof. Taking norms in the integral equation associated to (NS) we get

$$\|u(t)\|_q \;\leq\; Ct^{(-\frac{1}{p}+\frac{1}{q})\frac{n}{2}} \;+\; C\int_0^t (t-s)^{-\frac{1}{2}+(-\frac{1}{r}+\frac{1}{q})\frac{n}{2}} s^{(\frac{-2}{k}+\frac{1}{r})\frac{n}{2}}$$

for $q \geq p$, $q \geq r$, $2r \geq k$, $r \geq 1$. Here we have used the fact that

$$\|\partial_i G * \partial_j \nabla E_2\|_r \;\leq\; \begin{cases} C\|\partial_i G\|_r, & 1 < r < \infty, \\ C\|\partial_i G\|_{\mathcal{H}^1}, & r = 1 \end{cases}$$

and the estimates (see Theorem 1.2 and Remarks 1.2)

$$\|u(t)\|_{2r}^2 \;\leq\; Ct^{2(\frac{-1}{k}+\frac{1}{2r})\frac{n}{2}},$$

known to be valid for $1 \leq k \leq n$, $2r \geq k \geq p$ and, if $k < \frac{n}{2}$, $2r \leq \frac{kn}{n-2k}$. We have also used some classical estimates on the heat kernel,

$$\|\partial^\alpha G(t) * a\|_q \;\leq\; Ct^{-\frac{|\alpha|}{2}+(-\frac{1}{r}+\frac{1}{q})\frac{n}{2}} \|a\|_r$$

for $q \geq r$. We remark that $2r \geq k$ yields a restriction on r if $k > 2$.

We split the integral as follows:

(a)

$$\int_{\frac{t}{2}}^{t} (t-s)^{-\frac{1}{2}+(-\frac{1}{r}+\frac{1}{q})\frac{n}{2}} s^{(\frac{-2}{k}+\frac{1}{r})\frac{n}{2}} \leq C\, t^{\frac{1}{2}+(\frac{1}{q}-\frac{2}{k})\frac{n}{2}},$$

choosing r such that $\frac{1}{2}+(-\frac{1}{r}+\frac{1}{q})\frac{n}{2} > 0$, that is, $q \geq r > \frac{nq}{q+n}$.

(b)

$$\int_{0}^{\frac{t}{2}} (t-s)^{-\frac{1}{2}+(-\frac{1}{r}+\frac{1}{q})\frac{n}{2}} s^{(\frac{-2}{k}+\frac{1}{r})\frac{n}{2}} \leq C\, t^{\frac{1}{2}+(\frac{1}{q}-\frac{2}{k})\frac{n}{2}},$$

choosing r such that $(\frac{-2}{k}+\frac{1}{r})\frac{n}{2}+1 > 0$, that is, $1 \leq r < \frac{nk}{2(n-k)}$.

Those conditions imply $k > \frac{2n}{n+2}$ and $q \geq \frac{k}{2}$.

Supposing the above restrictions to be verified, we should get

$$\|u(t)\|_q \leq Ct^{(-\frac{1}{p}+\frac{1}{q})\frac{n}{2}} + Ct^{\frac{1}{2}+(\frac{1}{q}-\frac{2}{k})\frac{n}{2}} \leq Ct^{(-\frac{1}{p}+\frac{1}{q})\frac{n}{2}}$$

if $k = \frac{2pn}{n+p}$, where C is a constant depending on q and on the data. This is valid for any $q \geq \frac{k}{2} = \frac{pn}{n+p}$. Since $\frac{np}{n+p} \leq p$, it is valid for any $q \geq q$. On the other hand $k = \frac{2pn}{n+p} > \frac{2n}{n+2}$. It remains to check that, when $k < \frac{n}{2}$, the conditions $q \geq r > \frac{nq}{q+n}$ and $1 \leq r < \frac{nk}{2(n-k)}$ are compatible with the restriction $2r \leq \frac{kn}{n-2k}$. The following possibilities arise:

- $p \geq \frac{n}{3}$. Since $k \geq p$, the restriction is unnecessary so that we can find an adequate r in both cases.
- $p < \frac{n}{3}$. We have $\frac{nk}{2(n-k)} < \frac{kn}{2(n-2k)}$ so that we can find an adequate r for case (b). It remains that $\frac{nq}{q+n} \leq \frac{kn}{2(n-2k)}$, that is, $q(2n-5k) \leq kn$. When $p \geq \frac{n}{4}$, we get $2n-5k \leq 0$ and the inequality holds for any $q \geq p$. When $p < \frac{n}{4}$, we need to take $q \leq \frac{pn}{n-4p}$ in order to find some r for (a).

Therefore, the decay estimate

$$\|u(t)\|_q \leq Ct^{(-\frac{1}{p}+\frac{1}{q})\frac{n}{2}} \tag{$*$}$$

now holds for $q \geq p$ if $p \geq \frac{n}{4}$ and for $\frac{pn}{n-4p} \geq q \geq p$ when $1 \leq p < \frac{n}{4}$. If $n \geq 4$, we have concluded since $p < \frac{n}{4}$ is excluded.

If we iterate this process using these new decay rates when estimating the integrals appearing in the integral equation, we obtain that $(*)$ holds for $q \geq p$ if $p \geq \frac{n}{8}$ and for $\frac{pn}{n-8p} \geq q \geq p$ when $1 \leq p < \frac{n}{8}$. The last possibility is again excluded when $n \geq 8$.

In general, assuming that $(*)$ holds for $q \geq p$ if $p \geq \frac{n}{z}$ and for $\frac{pn}{n-zp} \geq q \geq p$ when $1 \leq p < \frac{n}{z}$, we get from the integral equation that $(*)$ also holds for $q \geq p$ if $p \geq \frac{n}{2z}$ and for $\frac{pn}{n-2zp} \geq q \geq p$ when $1 \leq p < \frac{n}{2z}$, so that when $n \leq 2z$, we are done. Thus, we can get the right decay estimate for any $q \geq p$ by repeating this procedure. The number of iterations we need depends on the dimension. □

5.2. First term. The integral equation yields

$$\|G(t) * u_0 - u(t)\|_q \leq C \int_0^t (t-s)^{-\frac{1}{2}+(-\frac{1}{r}+\frac{1}{q})\frac{n}{2}} s^{(\frac{-2}{p}+\frac{1}{r})\frac{n}{2}}$$

for $q \geq p$, $q \geq r \geq 1$ and $2r \geq p$, that is, $q \geq p$. As we did in the above proof, we split the integral in two intervals $[0, \frac{t}{2}]$ and $[\frac{t}{2}, t]$. By choosing an adequate r, we conclude the following.

THEOREM 5.2. *Let u be a strong solution of* (NS) *with data* $u_0 \in L^p \cap L^n(\mathbb{R}^n)$, $\frac{nq}{n+q} < p \leq n$, *of small* L^n *norm such that* $\operatorname{div} u_0 = 0$. *Then*

$$\|G(t) * u_0 - u(t)\|_q \leq Ct^{(-\frac{1}{p}+\frac{1}{q})\frac{n}{2}} t^{-\frac{n}{2p}+\frac{1}{2}}$$

if $q \geq p$.

5.3. Second term. Let us define u, h, and v as in §3. From the integral equation, we get

$$\begin{aligned}\|u(t) - v(t)\|_q &\leq \| \int_0^t \partial_i G(t-s) * (u^i(u-h) + h(u^i - h^i))(s) ds\|_q \\ &\quad + \| \int_0^t \partial_i G(t-s) * \partial_j \nabla E_2 * (u^i(u-h) + h(u^i - h^i))(s) ds\|_q \\ &\leq C \int_0^t (t-s)^{-\frac{1}{2}+(-\frac{1}{r}+\frac{1}{q})\frac{n}{2}} s^{(-\frac{1}{p}+\frac{1}{2r})\frac{n}{2}} s^{(-\frac{1}{p}+\frac{1}{2r})\frac{n}{2}} s^{-\frac{n}{2p}+\frac{1}{2}}\end{aligned}$$

for $q \geq r \geq 1$, $2r \geq p$, where we have used the estimates

$$\|u(t)\|_{2r} \leq Ct^{(-\frac{1}{p}+\frac{1}{2r})\frac{n}{2}},$$
$$\|(u-h)(t)\|_{2r} \leq Ct^{(-\frac{1}{p}+\frac{1}{2r})\frac{n}{2}} t^{-\frac{n}{2p}+\frac{1}{2}}$$

when $\frac{2n}{n+2} < p \leq n$. We split the integral as follows:

(a)

$$\int_{\frac{t}{2}}^t (t-s)^{-\frac{1}{2}+(-\frac{1}{r}+\frac{1}{q})\frac{n}{2}} s^{(-\frac{3}{p}+\frac{1}{r})\frac{n}{2}+\frac{1}{2}} \ \leq \ Ct^{(-\frac{3}{p}+\frac{1}{q})\frac{n}{2}+1}$$

if $r > \frac{nq}{q+n}$.

(b)

$$\int_0^{\frac{t}{2}} (t-s)^{-\frac{1}{2}+(-\frac{1}{r}+\frac{1}{q})\frac{n}{2}} s^{(-\frac{3}{p}+\frac{1}{r})\frac{n}{2}+\frac{1}{2}} \ \leq \ Ct^{(-\frac{3}{p}+\frac{1}{q})\frac{n}{2}+1}$$

if $(-\frac{3}{p} + \frac{1}{q})\frac{n}{2} + \frac{3}{2} > 0$, that is, $r \leq q$ and $1 \leq r < \frac{np}{3(n-p)}$, provided $p > \frac{3n}{n+3}$.

Therefore, we have the following theorem.

THEOREM 5.3. *Let u be a strong solution of* (NS) *with data* $u_0 \in L^p \cap L^n(\mathbb{R}^n)$, $\frac{3n}{n+3} < p < n$ *of small* L^n *norm such that* $\operatorname{div} u_0 = 0$. *Then*

$$\|(u-v)(t)\|_q \leq Ct^{(-\frac{1}{p}+\frac{1}{q})\frac{n}{2}} t^{-\frac{n}{2p}+\frac{1}{2}} t^{-\frac{n}{2p}+\frac{1}{2}}$$

if $q \geq p$, *where v is the solution of* $(\mathcal{L}_n)$ *with data* u_0.

This decay is faster than that corresponding to $u(t) - h(t)$.

5.4. Explicit second term. Let u be a strong solution of (NS) with data $u_0 \in L^1 \cap L^n(\mathbb{R}^n)$ of small L^n norm such that $\operatorname{div} u_0 = 0$. Making the change of variables

$$u_\lambda(x,t) = \lambda^n(\lambda^2 t, \lambda x,), \quad \lambda > 0,$$

we obtain the integral equation

$$\begin{aligned}u_\lambda(t) &= G(t) * u_{\lambda,0} - \lambda^{1-n} \int_0^t \partial_i G(t-s) * u_\lambda^i \partial_i u_\lambda(s) ds \\ &\quad - \lambda^{1-n} \int_0^t G(t-s) * \nabla \partial_j E_n * u_\lambda^i \partial u_\lambda^j(s) ds \\ &= w_{1,\lambda} + w_{2,\lambda} + w_{3,\lambda}\end{aligned}$$

for u_λ, where E_n stands for the fundamental solution of $-\Delta$ in $\mathbb{R}^n$. We have denoted by $w_{i,\lambda}$ the rescaled solutions of the n-dimensional analogues of problems $(\mathcal{P}_i)$ with $f = u$.

Problem $(\mathcal{P}_1)$. The first term w_1 is the solution of the heat equation with data u_0. If, for instance, $u_0 \in L^1(1+|x|;\mathbb{R}^n)$, we know that

$$t^{\frac{1}{2}+\frac{n}{2}(1-\frac{1}{q})}\|G(t) * u_0 - MG(t) + m^i\partial_i G(t)\|_q \to 0 \quad \text{as } t \to \infty,$$

where $M = \int_{\mathbb{R}^n} u_0$ and $m^i = \int_{\mathbb{R}^n} x^i u_0$ for $i = 1, \ldots, n$ (see [7]).

Problem $(\mathcal{P}_2)$. It is easy to prove that

$$\|w_{2,\lambda}\|_q = \lambda^{-1}\left\|\int_0^t \partial_i G(t-s) * \frac{f_\lambda^i f_\lambda(s)}{\lambda^{n-2}} ds\right\|_q$$

is bounded by $C\lambda^{-1}$. Keeping the notation of §4.2, we shall see that it is possible to take $g(t) = (\int_0^\infty \int_{\mathbb{R}^n} u^i u(y,\sigma) d\sigma dy)\ \partial_i G(t)$ and $\delta(t) = t^{\frac{1}{2}}$. More precisely, we have the following proposition.

PROPOSITION 5.4. *Let w_2 be the solution of $(\mathcal{P}_2)$ and $q \geq 1$. Then*

$$t^{\frac{1}{2}+\frac{n}{2}(1-\frac{1}{q})}\left\|w(t) + \left(\int_0^\infty \int_{\mathbb{R}^n} u^i u(\sigma,y) dy d\sigma\right) \partial_i G(t)\right\|_q \to 0 \quad \textit{as } t \to \infty,$$

provided that $u_0 \in L^{2r}(\mathbb{R}^n)$ for some $q \geq r > \frac{nq}{q+n}$.

Proof. We must prove that

$$\left\|\int_0^1 \partial_i G(1-s) * \frac{u_\lambda^i u_\lambda}{\lambda^{n-2}} - \left(\int_0^\infty \int_{\mathbb{R}^n} u^i u(\sigma,y) d\sigma dt\right) \partial_i G(1)\right\|_q \to 0 \quad \text{as } \lambda \to \infty.$$

Since

$$\int_0^1 \partial_i G(1-s) * \frac{u_\lambda^i u_\lambda}{\lambda^{n-2}} = \int_0^{\lambda^2} \int_{\mathbb{R}^n} \partial_i G\left(1 - \frac{s}{\lambda^2}, x - \frac{y}{\lambda}\right) u^i u,$$

this is equivalent to proving

$$\left\|\int_0^{\lambda^2} \int_{\mathbb{R}^n} \partial_i G\left(1 - \frac{s}{\lambda^2}, x - \frac{y}{\lambda}\right) u^i u - \left(\int_0^\infty \int_{\mathbb{R}^n} u^i u(\sigma,y) d\sigma dt\right) \partial_i G(1)\right\|_q \to 0$$
$$\text{as } \lambda \to \infty.$$

On the other hand,

$$\left(\int_0^\infty \int_{\mathbb{R}^n} u^i u(\sigma,y) d\sigma dt - \int_0^{\lambda^2} \int_{\mathbb{R}^n} u^i u(\sigma,y) d\sigma dt\right) \partial_i G(1,x) \to 0$$

in L_x^q as $\lambda \to \infty$, so that it suffices to prove that

$$\left\|\int_0^{\lambda^2} \int_{\mathbb{R}^n} \left(\partial_i G\left(1 - \frac{s}{\lambda^2}, x - \frac{y}{\lambda}\right) - \partial_i G(1,x)\right) u^i u\right\|_q \to 0 \quad \text{as } \lambda \to \infty.$$

We split the integral as in the proof of Theorem 4.1. We have

$$\|I_{\delta,\lambda}^1\|_{L_x^q} \leq C\varepsilon \int_0^{\lambda^2\delta} \int_{|y|\leq\lambda\delta} |u|^2(s,y) ds dy \ \leq\ \varepsilon C$$

if δ is small enough since

$$\int_0^\infty \int_{\mathbb{R}^n} |u|^2(s,y)dsdy \;\le\; C\int_0^\infty (1+s)^{\frac{-n}{2}}ds \;\le\; C$$

for $n > 2$ (see Theorem 5.1 and Remark 1.2). Next,

$$\|J^2_{\delta,\lambda}\|_{L^q_x} \le C\|\nabla G(t)\|_{L^q_x} \int_{\lambda^2\delta}^{\lambda^2} \int_{\mathbb{R}^n} |u|^2(s,y)dsdy \to 0$$

as $\lambda \to \infty$ for a fixed δ and

$$\|I^2_{\delta,\lambda}\|_{L^q_x} \le C\int_{\lambda^2\delta}^{\lambda^2} \left\| \int_{\mathbb{R}^n} \partial_i G\left(1-\frac{s}{\lambda^2}, x-\frac{y}{\lambda}\right) u^i u(s,y)\; dy \right\|_{L^q_x} ds$$

$$\le C\lambda^{n(1-\frac{1}{r})} \int_{\lambda^2\delta}^{\lambda^2} \left(1-\frac{s}{\lambda^2}\right)^{-\frac{n}{2}(\frac{1}{r}-\frac{1}{q})-\frac{1}{2}} s^{-\frac{n}{2}(2-\frac{1}{r})} \;\le\; C\lambda^{2-n}$$

if $q \ge r > \frac{nq}{q+n}$ and $u_0 \in L^{2r}(\mathbb{R}^n)$ (see Theorem 5.1 and Remark 1.2). Since $n > 2$, it tends to 0 as $\lambda \to \infty$ for a fixed δ.

Concerning $\|J^3_{\delta,\lambda}\|_{L^q_x}$, we have

$$\|J^3_{\delta,\lambda}\|_{L^q_x} \;\le\; C\|\nabla G(t)\|_{L^q_x} \int_0^{\lambda^2\delta} \int_{|y|\ge\delta\lambda} |u|^2(s,y)dsdy.$$

Since

$$\int_0^\infty \int_{\mathbb{R}^n} |u|^2(s,y)dsdy \;\le\; C,$$

it follows that

$$\int_0^\infty \int_{|y|\ge\delta\lambda} |u|^2(s,y)dsdy \quad \to 0 \quad \text{as } \lambda \to \infty.$$

Finally,

$$\|I^3_{\lambda,\delta}\|_{L^q_x} \;\le\; \int_0^{\lambda^2\delta} \left\| \int_{|y|\ge\delta\lambda} \partial_i G\left(1-\frac{s}{\lambda^2}, x-\frac{y}{\lambda}\right) u^i u(y,s) \right\|_{L^q_x}$$

$$\le\; C\int_0^{\lambda^2\delta} \left(1-\frac{s}{\lambda^2}\right)^{-1+\frac{1}{q}-\frac{1}{2}} \|u^i u(y,s)\|_{L^1(|y|\ge\lambda\delta)} \;\le\; C\int_0^{\lambda^2\delta}\int_{|y|\ge\delta\lambda} |u|^2(y,s)dyds$$

and we finish in the same way as before. □

Problem $(\mathcal{P}_3)$. In this case, by slightly modifying the proof above, we get the following proposition.

PROPOSITION 5.5. *Let w be the solution of $(\mathcal{P}_3)$ with $f = u$ and $q \ge 1$. Then*

$$t^{\frac{1}{2}+\frac{n}{2}(1-\frac{1}{q})} \left\| w(t) + \left(\int_0^\infty \int_{\mathbb{R}^n} u^i u^j(\sigma,y)dyd\sigma\right) \partial_i G(t) * \nabla\partial_j E_n \right\|_q \to 0 \quad \textit{as } t \to \infty,$$

provided that $u_0 \in L^{2r}(\mathbb{R}^n)$ for some $q \ge r > \frac{nq}{q+n}$.

As a consequence of these results, we get the following theorem.

THEOREM 5.6. *Let u be a strong solution of* (NS) *with data* $u_0 \in L^1(\mathbb{R}^n, 1+|x|) \cap L^n(\mathbb{R}^n)$ *of small* L^n *norm and* $q \geq 1$. *Then*

$$t^{\frac{1}{2}+\frac{n}{2}(1-\frac{1}{q})}\|u(t) - MG(t) + m^i\partial_i G(t) + R(t)\|_q \to 0$$

as $t \to \infty$, *where*

$$R(t) = \left(\int_0^\infty \int_{\mathbb{R}^n} u^i u(\sigma, y) dy d\sigma\right) \partial_i G(t) + \left(\int_0^\infty \int_{\mathbb{R}^n} u^i u^j(\sigma, y) dy d\sigma\right) \partial_i G(t) * \nabla\partial_j E_n,$$

provided that $u_0 \in L^{2r}(\mathbb{R}^n)$ *for some* $q \geq r > \frac{nq}{q+n}$.

Remark 5.1. If, instead of the problems $(\mathcal{P}_i)$ corresponding to solutions of (NS), we consider those corresponding to solutions v of $(\mathcal{L}_n)$ with $f = G(t) * u_0$ replaced by $f = u$, the analogues of Propositions 5.4 and 5.5 also hold. Therefore, if $u_0 \in L^1(1+|x|; \mathbb{R}^n) \cap L^n(\mathbb{R}^n)$, $q \geq 1$, and $u_0 \in L^{2r}(\mathbb{R}^n)$ for some $q \geq r \geq \frac{nq}{q+n}$, the solution v of $(\mathcal{L}_n)$ satisfies

$$t^{\frac{1}{2}+\frac{n}{2}(1-\frac{1}{q})}\|v(t) - MG(t) - m^i\partial_i G(t) + R(t)\|_q \to 0$$

as $t \to \infty$, where

$$R(t) = \left(\int_0^\infty \int_{\mathbb{R}^n} (G(t) * u_0)^i (G(t) * u_0)(\sigma, y) dy d\sigma\right) \partial_i G(t)$$

$$+ \left(\int_0^\infty \int_{\mathbb{R}^n} (G(t) * u_0)^i (G(t) * u_0)^j (\sigma, y) dy d\sigma\right) \partial_i G(t) * \nabla\partial_j E_n.$$

Acknowledgments. The author would like to thank E. Zuazua for suggesting to her this question.

REFERENCES

[1] W. BORCHERS AND T. MIYAKAWA, *L^2 decay for Navier–Stokes flows in unbounded domains with applications to exterior stationary flows*, Arch. Rational Mech. Anal., 118 (1992), pp. 273–295.

[2] H. BEIRAO DA VEIGA, *Existence and asymptotic behavior for strong solutions of the Navier–Stokes equations in the whole space*, Indiana Univ. Math. J., 36 (1986), pp. 149–166.

[3] L. CAFARELLI, R. KOHN, AND L. NIRENBERG, *Partial regularity of suitable weak solutions of the Navier–Stokes equations*, Comm. Pure Appl. Math., 35 (1982), pp. 771–831.

[4] R. COIFMAN, P. L. LIONS, Y. MEYER, AND S. SEMMES, *Compacité par compensation et espaces de Hardy*, C. R. Acad. Sci. Paris Sér. I , 309 (1989), pp. 945–949.

[5] R. COIFMAN AND Y. MEYER, *Au-delà des opérateurs pseudo-différentiels*, Astérisque, 57 (1978), pp. 77–142.

[6] A. CARPIO, *Comportement asymptotique des solutions des équations du tourbillon en dimensions* 2 *et* 3, C. R. Acad. Sci. Paris Sér. I, 316 (1993), pp. 1289–1294.

[7] J. DUOANDIKOETXEA AND E. ZUAZUA, *Moments, Masses de Dirac et decomposition de fonctions*, C. R. Acad. Sci. Paris Sér. I, 315 (1992), pp. 693–698.

[8] E. B. FABES, B. F. JONES, AND N. M. RIVIERE, *The initial value problem for the Navier–Stokes equations with data in L^p*, Arch. Rational Mech. Anal., 45(1972), pp. 222–240.

[9] C. FEFFERMAN AND E. M. STEIN, *H^p spaces of several variables*, Acta Math., 129 (1972), pp. 137–193.

[10] Y. GIGA AND T. MIYAKAWA, *Solutions in L^r of the Navier–Stokes initial value problem*, Arch. Rational Mech. Anal., 89 (1985), pp. 267–281.

[11] Y. GIGA, T. MIYAKAWA, AND H. OSADA, *Two dimensional Navier–Stokes flow with measures as initial vorticity*, Arch. Rational Mech. Anal., 104 (1988), pp. 223–250.

[12] Y. Giga and T. Kambe, *Large time behavior of the vorticity of two dimensional viscous flow and its applications to vortex formation*, Comm. Math. Phys., 117 (1988), pp. 549–568.

[13] E. Hopf, *Uber die Anfangswertaufgabe fur die hydrodynamischen Grundgleichungen*, Math. Nachr., 4 (1950–1951), pp. 213–231.

[14] T. Kato, *Strong L^p solutions of the Navier–Stokes equations in $\mathbb{R}^n$, with applications to weak solutions*, Math. Z., 187 (1984), pp. 471–480.

[15] ———, *Strong L^p solutions of Navier–Stokes equations in Morrey Spaces*, Bol. Soc. Brasil Mat., 22 (1992), pp. 127–155.

[16] T. Kato and T. Fujita, *On the Navier–Stokes initial value problem* I, Arch. Rational Mech. Anal., 16 (1964), pp. 269–315.

[17] R. Kajikiya and T. Miyakawa, *On L^2 decay of weak solutions of the Navier–Stokes equations in $\mathbb{R}^n$*, Math. Z., 192 (1986), pp. 135–148.

[18] J. Leray, *Etude de diverses équations intégrales non linéaires et de quelques problèmes que pose l'Hydrodynamique*, J. Math. Pures Appl., 12 (1933), pp. 1–82.

[19] ———, *Sur le mouvement d'un liquide visqueux emplissant l'espace*, Acta Math., 63 (1934), pp. 193–248.

[20] M. E. Schonbek, *L^2 decay for weak solutions of the Navier–Stokes equations*, Arch. Rational Mech. Anal., 88 (1985), pp. 209–222.

[21] ———, *Large time behavior of solutions to the Navier–Stokes equations*, Comm. Partial Differential Equations, 11 (1986), pp. 733–763.

[22] ———, *Lower bounds of rates of decay for solutions to the Navier–Stokes equations*, J. Amer. Math. Soc., 4 (1991), pp. 423–449.

[23] J. Serrin, *The initial value problem for the Navier–Stokes equation*, in Nonlinear Problems, Proc. Symp. MRC, University of Wisconsin, Madison, 1963, pp. 66–98.

[24] M. E. Taylor, *Analysis on Morrey spaces and applications to Navier–Stokes and other evolution equations*, Comm. Partial Differential Equations, 17 (1992), pp. 1407–1456.

[25] W. Von Wahl, *Regularity questions for Navier–Stokes equations, Approximation methods for Navier–Stokes problems*, Lecture Notes in Math., 771 (1977), pp. 538–542.

[26] M. Wiegner, *Decay results for weak solutions of the Navier–Stokes equations on $\mathbb{R}^n$*, J. London Math. Soc., 35 (1987), pp. 303–313.

[27] E. Zuazua, *Weakly nonlinear large time behaviour in scalar convection-diffusion equations*, Integral and Differential Equations, 6 (1993), pp. 1481–1492.

SIAM J. MATH. ANAL.
Vol. 27, No. 2, pp. 476–485, March 1996

THE DEGENERACY OF A FAST-DIFFUSION EQUATION AND STABILITY*

YUAN-WEI QI†

Abstract. It is known that for any nonnegative initial value $u_0(x)$, the solution $u(x,t)$ of the initial-boundary value problem $u_t = \operatorname{div}(|\nabla u|^{m-1}\nabla u)$ in a bounded domain $\Omega \subset R^N$ with $u|_{\partial\Omega} = 0$, where $0 < m < 1$, becomes degenerate in finite time $T > 0$, i.e., it tends to zero as $t \to T$. Therefore, it is important to know the spatial pattern of $u(x,t)$ as $t \to T^-$. In this paper we study this problem and prove that the spatial pattern is characterized by solutions which are in the form of separation of variables and have the same extinction time T.

Key words. fast-diffusion equation, finite-time extinction, stability, convergence

AMS subject classifications. 34A10, 35B05, 35B40, 35K55

1. Introduction.

In this paper we study the initial-boundary value problem

$$\begin{aligned} &u_t(x,t) = \operatorname{div}(|\nabla u|^{m-1}\nabla u), && x \in \Omega,\ t > 0,\\ &u(x,0) = u_0(x) \geq 0, && x \in \Omega,\\ &u = 0, && x \in \partial\Omega, \end{aligned} \tag{1}$$

where $0 < m < 1$ and Ω is a bounded domain with smooth boundary. Equation (1) arises from a number of different applications includes non-Newtonian flow fluids [2], [14] and nonlinear filtration [12]. The case is referred as “fast” diffusion since it exhibits behaviour similar to that of the well-known equation

$$u_t = \triangle u^m, \quad 0 < m < 1 \tag{2}$$

(cf. [7]). We refer the interested reader to [5], [13], and the references therein for results on Dirichlet initial-boundary value problem of (2).

Indeed, it was shown in [7] that, like (2), any solution of (1) decays to zero in finite time. Hence it is in strong contrast to the case $m > 1$, where the solution decays to zero as time goes to infinity, like an inverse power of t [1]. However, the spatial pattern of solution $u(x,t)$ of (1) when t approaches the extinction time $T > 0$ was not studied in [7].

The main purpose of this paper is to study the behaviour of solution u when t approaches the extinction time T. Motivated by linear theory, we seek a function $z(x)$, positive for $x \in \Omega$, and a function $R(t)$ such that $z(x)R(t)$ is a separable solution of (1) which becomes extinct at finite time $T > 0$. We expect that for any solution $u(x,t)$ which has the same extinction time $T > 0$, $u(x,t)/R(t) \to z(x)$ as $t \to T$ in suitable function spaces. Indeed, as shown in Theorem 2 below, the conclusion is true provided z is the only positive solution of problem (I) below.

This work was motivated by [7] (in bounded domain) and [11] (in the whole space R^N) on (1). Other works on (2) such as [5] (in bounded domain) and [4] (in the whole space R^N) also influenced the contents and techniques of this paper.

We assume throughout this paper the following.

Assumptions. (i) $u_0(x) \geq 0$ is sufficiently smooth and is zero on $\partial\Omega$.

* Received by the editors December 9, 1993; accepted for publication (in revised form) July 26, 1994.

† Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.

(ii) For some $\alpha > 0$, $u(x,t)$ is a $C^{1+\alpha}$ solution of (1) which is positive on $\Omega \times (0,T)$ and fulfills the initial and boundary value conditions.

(iii) $u(x,T) = 0$ for all $x \in \overline{\Omega}$.

(iv) $u_t \in C(0,T; L^2(\Omega)) \cap C(0,T; W_0^{-1,1+m}(\Omega))$

We note in passing that some of the assumptions can be weakened and the results of this paper still hold. However, since our objective is to find the asymptotic profile of extinction, we will not elaborate on the weakest conditions for all the manipulations to be legitimate. Nevertheless, the estimates in [6]–[8] justify the above assumptions as appropriate.

Consider the homogeneity of equation (1) and the boundary condition. It is natural to seek a separable solution. Indeed, such a solution which becomes extinct at finite time $T > 0$ takes the form $u(x,t) = (1-m)(T-t)^{1/(1-m)} z(x)$. Then z is a solution of the following boundary value problem

$$\text{(I)} \begin{cases} \operatorname{div}((| \nabla z|^{m-1} \nabla z) + \lambda z = 0, & x \in \Omega, \\ z > 0, & x \in \Omega, \\ z = 0, & x \in \partial\Omega, \end{cases}$$

where $\lambda = 1/(1-m)^m$.

The existence of (I) can be studied by using a variational argument and a Pohozaev type identity. The special case where Ω is a ball and z is radial symmetric was studied in [15]. In spite of the fact that the study of (I) can be done by simply modifying the argument on the semilinear equation $\triangle u + f(u) = 0$, we cannot find the result for general domain Ω in the literature. Therefore, we show the following result on the existence of (I) in the next section.

THEOREM 1. *Let* $0 < m < 1$.

(i) *If* $N \le 2$ *or* $m > (N-2)/(N+2)$ *for* $N \ge 3$, *then there exists a positive solution of* (I).

(ii) *If* $N = 1$ *or* $N \ge 2$ *and* $\Omega = B_r(x_0)$, *a ball with center* x_0 *and radius* $r > 0$, *then the positive radial solution is unique.*

(iii) *If* $m \le (N-2)/(N+2)$, $N \ge 3$, *and* Ω *is star shaped, then* (I) *has no positive classical solution.*

(iv) *If* $m \le (N-2)/(N+2)$, $N \ge 3$, *and* Ω *is an annulus, then there exists a positive classical solution of* (I).

Remark. By a *classical solution* of (I) we mean the one which belongs to $C^{1+\alpha}$, for some $\alpha > 0$. In this case the differential equation is satisfied pointwise in the classical sense.

Once the existence on (I) is settled, the question is whether the separable solutions characterize the spatial pattern of the general solution when $t \to T$. The following result answers this question and will be proved in §3.

THEOREM 2. *Let* $0 < m < 1$. $v(x,t) = u(x,t)/(T-t)^{1/(1-m)}$.

(i) *Suppose* $N \le 2$ *or* $m > (N-2)/(N+2)$ *and* $N \ge 3$. *Then there exists an increasing sequence of times* $t_n \to T^-$ *and a positive classical solution* z *of* (I) *such that* $v(\cdot, t_n) \to z$ *in* $W_0^{1,1+m}(\Omega)$.

(ii) *In case* (i), *if there is a unique solution to* (I), *then* $v(\cdot,t) \to z(\cdot)$ *as* $t \to T^-$ *in* $W_0^{1,1+m}(\Omega)$.

2. The steady state. In this section we study (I) and prove Theorem 1.

Proof of Theorem 1. (i) The standard way to prove existence is to consider the following minimization problem.

Problem A. Minimize $\int_\Omega |\bigtriangledown h|^{1+m}\,dx$ over the class of functions in $W_0^{1,1+m}(\Omega)$ satisfying $\int_\Omega h^2\,dx = k$, where $k > 0$ is a constant.

Since the embedding $W_0^{1,1+m}(\Omega) \hookrightarrow L^2(\Omega)$ is compact in case (i), the Palais–Smale (PS) condition holds. Hence the minimum is achieved and is a $C^{1+\alpha}(\overline{\Omega})$ solution of $\operatorname{div}(|\bigtriangledown u|^{m-1}\bigtriangledown u) + \mu u = 0$ (see [9]), where $\mu > 0$. A scaling gives a solution of (I). This completes the proof of (i).

(ii) Uniqueness in the case where $N = 1$ and $\Omega = (0,1)$ follows from the fact that if z is a solution of (I), then the solution to the initial value problem

$$(|Q'|^{m-1}Q')' = -\lambda Q, \quad Q(0) = 0, \quad Q'(0) = \eta^{2/(1+m)}z'(0)$$

is given by $Q^\eta(x) = \eta z(\eta^\alpha x)$ with $\alpha = (1-m)/(1+m)$. Hence there is only one value, $\eta = 1$, for which $Q(1) = 0$. Likewise, the uniqueness is true in the case where Ω is a ball and z is radial symmetric.

(iii) The proof of this part is based on a Pohozaev identity. We note that although z may not belong to $C^2(\Omega)$, $\operatorname{div}(\bigtriangledown z|^{m-1}\bigtriangledown z)$ is in $C^{1+\alpha}(\overline{\Omega})$ and therefore the following formal manipulation is rigorous.

Multiplying (I) by u and integrating by parts we have

$$-\int_\Omega |\bigtriangledown z|^{m+1}\,dx + \int_\Omega z^2\,dx = 0. \tag{3}$$

Likewise, multiplying (I) by $x\cdot\bigtriangledown u$ we obtain

$$-\frac{N-1-m}{1+m}\int_\Omega |\bigtriangledown z|^{m+1}\,dx + \frac{N}{2}\int_\Omega z^2\,dx = -\int_{\partial\Omega} |\bigtriangledown z|^{m+1}(x\cdot\nu)\,ds, \tag{4}$$

where ν is the outward normal of $\partial\Omega$. Subtracting (3) from (4) we find

$$\left(\frac{N}{2} - \frac{N-1-m}{1+m}\right)\int_\Omega |\bigtriangledown z|^{m+1}\,dx = \int_{\partial\Omega} |\bigtriangledown z|^{m+1}(x\cdot\nu)\,ds. \tag{5}$$

But by our assumption on m, $N/2-(N-1-m)/(1+m) \le 0$. Therefore, the left-hand side is nonpositive, whereas the right-hand side is positive for z nontrivial because Ω is star shaped and $\bigtriangledown z \neq 0$ on $\partial\Omega$. This proves (iii).

(iv) When Ω is an annulus we can consider the radial solutions $z(x) = z(r)$, where $r = |x|$. In this case (I) takes the form

$$\text{(IR)}\begin{cases} \frac{1}{r^{N-1}}(r^{N-1}|z'|^{m-1}z')' + \lambda z = 0, & R_1 < r < R_0, \\ z > 0, & R_1 < r < R_0, \\ z = 0, & r = R_1 \text{ and } r = R_0. \end{cases}$$

Let

$$\xi = \left[\frac{N-1-m}{m} r^{(N-1-m)/m}\right]^{-1}, \quad y(\xi) = z(r). \tag{6}$$

Then (IR) takes the form

$$\begin{cases} (|y'|^{m-1}y')' + \rho(\xi)y = 0, & \xi_0 < \xi < \xi_1, \\ y > 0, & \xi_0 < \xi < \xi_1, \\ y = 0, & \xi = \xi_0 \text{ and } \xi = \xi_1, \end{cases}$$

where

$$\rho(\xi)=\left(\frac{N-1-m}{m}\xi\right)^{-k},\qquad k=\frac{N(1+m)}{N-1-m}, \tag{7}$$

$$\xi_i=\left(\frac{N-1-m}{m}R_i^{(N-1-m)/m}\right)^{-1},\qquad i=0,1. \tag{8}$$

The existence of a solution to (I) can be proved by using a shooting argument. For this purpose we shall examine the family of solutions of the initial value problem

$$\text{(II)}\begin{cases} (|y'|^{m-1}y')'+\rho(\xi)y=0, & \xi<\xi_1,\\ y(\xi_1)=0, & y'(\xi)=-b<0.\end{cases}$$

Here ξ_1 is a positive number that will be kept fixed throughout.

It is clear that if y is positive in some interval (β,ξ_1) with $\beta\geq 0$, then

$$y(\xi)\leq b(\xi_1-\xi)\qquad \text{in } (\beta,\xi_1). \tag{9}$$

Consequently, if $\beta>0$ the solution can be extended to the left of β. Denote

$$\xi_0(b)=\inf\{\xi_0>0: u(\xi;b)>0\quad \text{in } (\xi_0,\xi_1)\}. \tag{10}$$

It is a standard ordinary differential equation (ODE) result that the function $b\to\xi_0(b)$ is continuously differentiable in the neighborhood of every $b>0$ such that $\xi_0(b)>0$. This is because $y'(\xi_0(b);b)>0$. Then we can use the energy functionals

$$J_1(\xi)=\frac{|y'(\xi)|^{1+m}}{1+m}+\frac{1}{2}\rho(\xi)y^2, \tag{11}$$

$$J_2(\xi)=\frac{|y'(\xi)|^{1+m}}{(1+m)\rho(\xi)}+\frac{1}{2}y^2 \tag{12}$$

and the Sturm–Liouville comparison theorem to prove the following technical result.

THEOREM 3. (i) *If* $b>0$ *and* $y(\cdot,b)$ *is defined and positive in* $(0,\xi_1)$ *then* $\lim_{\xi\to 0^+}y(\xi;b)=0$.

(ii) $\lim_{b\to 0}\xi_0(b)=0$.

(iii) $\lim_{b\to\infty}\xi_0(b)=\infty$

The proof of the result is technical and lengthy. We omit it here. It will be proved in [16]. The interested reader may also consult [3], where the semilinear case $m=1$ was treated. The argument there may be modified to work for the case $0<m<1$.

The direct consequence of the above is the existence of (I) for the annulus case. Thus the proof of Theorem 1 is complete.

3. The main results. We will prove Theorem 2 through a series of lemmas.

LEMMA 1. *Let* $m>0$ *if* $N\leq 2$ *or* $m>(N-2)/(N+2)$ *if* $N\geq 3$. *Then there exists a positive constant* $B=B(m,N,\Omega)$ *such that every solution* $u(x,t)$ *of* (1) *with finite extinction time* $T>0$ *satisfies*

$$(T-t)^{\frac{2}{1-m}}\leq B\int_\Omega u^2. \tag{13}$$

Proof. Multiplying equation (1) by u and integrating by parts, we obtain the identity

$$\frac{d}{dt}\int_\Omega u^2\,dx = -2\int_\Omega |\nabla u|^{m+1}\,dx. \tag{14}$$

The Sobolev embedding lemma guarantees that there exists a constant $c = c(m, N, \Omega)$ such that for any $h \in W_0^{1,1+m}(\Omega)$,

$$\left(\int_\Omega h^2\,dx\right)^{1/2} \le c\left(\int_\Omega |\nabla h|^{m+1}\,dx\right)^{1/(1+m)}. \tag{15}$$

We obtain a differential inequality by substituting (15) into (14), which after integration yields

$$\left(\int_\Omega u(x,t')^2\,dx\right)^{(1-m)/2} - \left(\int_\Omega u(x,t)^2\,dx\right)^{(1-m)/2} \le -B(t-t')$$

for $t \le t' < T$, where $B = (1-m)/c^{1+m}$. Letting $t' \to T$ and multiplying by -1, we get the lemma.

LEMMA 2. *Let u and m be as in Lemma* 1. *Then the solution u satisfies*

$$\int_\Omega u^2(x,t)\,dx \le (1-t/T)^{(1-m)/2}\int_\Omega u^2(x,0)\,dx. \tag{16}$$

Proof. First, we observe that

$$\begin{aligned}\frac{d}{dt}\int_\Omega |\nabla u|^{m+1}\,dx &= (1+m)\int_\Omega |\nabla u|^{m-1}\nabla u\cdot(\nabla u)_t\,dx\\ &= -(1+m)\int_\Omega \operatorname{div}(|\nabla u|^{m-1}\nabla u)u_t dx \qquad (17)\\ &= -(1+m)\int_\Omega |\operatorname{div}(|\nabla u|^{m-1}\nabla u)|^2 dx, \qquad (18)\end{aligned}$$

$$\int_\Omega |\nabla u|^{m+1}\,dx = -\int_\Omega \operatorname{div}(|\nabla u|^{m-1}\nabla u)u dx. \tag{19}$$

We used the condition $u = u_t = 0$ on $\partial\Omega$ in the above manipulation. Applying the Cauchy–Schwarz inequality, we find

$$\left(\int_\Omega |\nabla u|^{m+1}\right)^2 \le \int_\Omega |\operatorname{div}(|\nabla u|^{m-1}\nabla u)|^2 \int_\Omega u^2. \tag{20}$$

Combining (14), (17), and (20), we have

$$-\frac{\frac{d}{dt}\int_\Omega u^2\,dx}{2\int_\Omega u^2\,dx} \le -\frac{\frac{d}{dt}\int_\Omega |\nabla u|^{m+1}\,dx}{(1+m)\int_\Omega |\nabla u|^{m+1}\,dx}. \tag{21}$$

Integrating the above inequality, we have

$$\frac{\int_\Omega u^2(x,t)\,dx}{\int_\Omega u^2(x,s)\,dx} \ge \frac{(1+m)}{2}\frac{\int_\Omega |\nabla u(x,t)|^{m+1}\,dx}{\int_\Omega |\nabla u(x,s)|^{m+1}\,dx} \tag{22}$$

for $0 \le s < t \le T$. Set

$$J(t) = \left(\int_\Omega u^2(x,t)\,dx\right)^{(1-m)/2}. \tag{23}$$

Then (22) is just $J'(s) \le J'(t)$ for all $s \le t$, and therefore $J'' \ge 0$ and J is convex. Hence

$$\frac{J(T)-J(t)}{T-t} \ge \frac{J(t)-J(0)}{t}, \tag{24}$$

which is just (16).

Remark. Lemmas 1 and 2 with $t = 0$ provide estimates of the extinction time T in terms of the initial value. Note also that inequality (22) in the proof of Lemma 2 gives the following corollary.

COROLLARY 1. *Let*

$$G(h) = \int_\Omega |\nabla h(x,t)|^{m+1}\,dx \Big/ \left(\int_\Omega h^2(x,t)\,dx\right)^{(1+m)/2}. \tag{25}$$

Then the function $G(u(\cdot,t))$ is a nonincreasing function of t.

The functional G can be regarded as a generalization of the Rayleigh–Ritz quotient when $m \ne 1$. If $N \le 2$ or if $N \ge 3$ and $m > (N-2)/(N+2)$, then $G(h)$ has a positive minimum over $W_0^{1,1+m}(\Omega)$. If $N \ge 3$ and $m \le (N-2)/(N+2)$, the infimum of G is zero on $W_0^{1,1+m}$ but the minimum does not exist.

Proof of Theorem 2. Let

$$w(x,s) = \frac{T^\lambda}{(T-t)^\lambda}u(x,t),$$

where $s = -\log(T-t) + \log T$. In order to prove (i) of the theorem, we must show that there exists a sequence of times $s_n \to \infty$ such that $w(\cdot,s_n) \to z(\cdot)$ in $W_0^{1,1+m}(\Omega)$, where z is a solution of (I). To prove (ii), we have to show that $w(\cdot,s) \to z(\cdot)$.

The function $w(x,s)$ satisfies

$$w_s(x,t) = \operatorname{div}(|\nabla w|^{m-1}\nabla w) + \lambda w \tag{26}$$

in $\Omega \times (0,\infty)$, with $w = 0$ on $\partial\Omega$ and $w(x,0) = u(x,0)$, $x \in \Omega$.

We define the functional

$$I(h) = \int_\Omega \left(\frac{1}{1+m}|\nabla h|^{m+1}\,dx - \frac{\lambda}{2}h^2\right)\,dx \tag{27}$$

and $f(s) = I(w(\cdot,s))$. It is easy to verify that

$$f'(s) = -\int_\Omega w_s(x,s)^2\,dx \le 0. \tag{28}$$

Lemma 2 shows that $\int_\Omega w(x,s)^2\,dx$ is bounded for all $s > 0$. Hence $f(s)$ is bounded. Consequently, $\int_\Omega |\nabla w(x,s)|^{m+1}$ is uniformly bounded. Thus $\lim_{s\to\infty} f(s)$ exists and there exists a sequence of times $s_n \to \infty$ such that $f'(s_n) \to 0$ by (28).

We proceed to show that for a subsequence of s_n, again labelled s_n, $w(\cdot,s_n) \to z(\cdot)$ in $W_0^{1,1+m}(\Omega)$. Since $w(\cdot,s_n)$ is a bounded sequence in $W_0^{1,1+m}$, there exists a subsequence, again labelled by $w(\cdot, s_n)$, such that $w(\cdot, s_n)$ converges weakly to $z(\cdot)$ in $W_0^{1,1+m}(\Omega)$. In addition, from the compact-embedding theorem, we have

$$\int_\Omega w(x, s_n)^2\,dx \to \int_\Omega z(x)^2\,dx, \tag{29}$$

since $m > (n-2)/(n+2)$ if $N \geq 3$ or $m > 0$ if $N \leq 2$. In fact,

$$\int_\Omega w(x, s)^2\,dx \to \int_\Omega z(x)^2\,dx, \tag{30}$$

since $\int_\Omega w(x,s)^2\,dx$ is nonincreasing by Lemma 2.

Next we show that $z(x)$ is a classical solution of (I) and $w(x, s_n) \to z(x)$ in $W_0^{1,1+m}(\Omega)$. To prove that z is a classical solution of (I), we note that since

$$w(\cdot, s_n) \to v(\cdot), \quad w_s(\cdot, s_n) \to 0 \qquad \text{in } L^2(\Omega),$$

$\operatorname{div}(|\nabla u|^{m-1}\nabla u)(\cdot, s_n)$ is convergent in $L^2(\Omega)$. Moreover, when we multiply (1) by $\eta \in C_c^\infty(\Omega)$ and integrate by parts, we find

$$-\int_\Omega |\nabla w|^{m-1}\nabla w \cdot \nabla w\eta\,dx + \lambda\int_\Omega w\eta\,dx = \int_\Omega w_s\eta\,dx.$$

Let $s_n \to \infty$. It then follows that

$$-\int_\Omega |\nabla z|^{m-1}\nabla z \cdot \nabla\eta\,dx + \lambda\int_\Omega z\eta\,dx = 0.$$

Therefore z is a weak solution of (I) in $W_0^{1,1+m}(\Omega)$. The standard bootstrap technique (see [9]) yields that z is a classical solution. Here again, the condition $N \leq 2$ or $m > (n-2)/(n+2)$ if $N \geq 3$ is used.

We now show that $w(\cdot, s_n) \to z$ in $W_0^{1,1+m}(\Omega)$. As is clear from (26) and the fact that $w(\cdot, s_n) \to z(\cdot)$ in $L^2(\Omega)$ and $w_s(\cdot, s_n) \to 0$,

$$\operatorname{div}(|\nabla w|^{m-1}\nabla w)(\cdot, s_n) \to \operatorname{div}(|\nabla z|^{m-1}\nabla z)(\cdot) \qquad \text{in } L^2(\Omega).$$

On the other hand, we find, by using Hölder's inequality,

(31)

$$\begin{aligned}
&\int_\Omega |\nabla w - \nabla z|^{1+m}dx \\
&\leq \left[\int_\Omega (|\nabla w| + |\nabla z|)^{1+m}\right]^p \left[\int_\Omega (|\nabla w| + |\nabla z|)^{m-1}|\nabla(w-z)|^2\right]^q \\
&\leq C\left[\int_\Omega |\nabla w|^{1+m} + |\nabla z|^{1+m}\right]^p \left[\int_\Omega (|\nabla w|^{m-1}\nabla w - |\nabla z|^{m-1}\nabla z)(\nabla(w-z))\right]^q \\
&\leq C\left[\int_\Omega |\nabla w|^{1+m} + |\nabla z|^{1+m}\right]^p \left[\int_\Omega \operatorname{div}(|\nabla w|^{m-1}\nabla w - |\nabla z|^{m-1}\nabla z)(z-w)\right]^q \\
&\longrightarrow 0 \qquad \text{as } s_n \to \infty,
\end{aligned}$$

where $p = \frac{1-m}{2}$, $q = \frac{1+m}{2}$. Thus $w(\cdot, s_n) \to z$ in $W_0^{1,1+m}(\Omega)$.

It is clear that since $w(x, s_n) \geq 0$, $z(x) \geq 0$. Moreover, since $\int_\Omega w^2(x,s)dx \geq C > 0$ for some $C > 0$ by Lemma 1 and $w(\cdot, s_n) \to z(\cdot)$ in $L^2(\Omega)$, we know that z is not identically zero. It then follows from a Harnack-type inequality (see Proposition 1 below) that z is strictly positive in Ω. Thus $z > 0$.

PROPOSITION 1. *Let z be a nonnegative solution of* (I) *which is not identically zero. Suppose $B_{4R}(y) \subset \Omega$ and $1 \leq p < N/(N-1-m)$; then*

$$R^{-n/p}||z||_{L^p(B_{2R}(y))} \leq C \inf_{B_R(y)} z, \tag{32}$$

where $C = C(N, p, R, ||z||_\infty)$.

Proof. The argument is essentially the same as that of [17] and Theorems 8.17 and 8.18 in [10]. Nevertheless, those works concern more general problems and the estimates obtained there are not sharp. Therefore, we present a proof which gives the result of Proposition 1.

Let $\phi = \eta^{1+m}\overline{z}^\beta$, $\overline{z} = z + \epsilon$, where $\beta < 0$ and $\epsilon > 0$.

$$\nabla\phi = (1+m)\eta^m \nabla \eta \overline{z}^\beta + \beta\overline{z}^{\beta-1} \nabla z \eta^{1+m}. \tag{33}$$

Multiplying (I) by ϕ and integrating by parts, we find

$$\begin{aligned} &\int_\Omega \eta^{1+m}\overline{z}^{\beta-1}|\nabla z|^{(1+m)}dx \\ &\quad + (1+m)\int_\Omega \eta^m \nabla\eta\cdot\nabla z|\nabla z|^{m-1}\overline{z}^\beta\, dx - \int_\Omega \eta^{1+m}\overline{z}^\beta z dx = 0. \end{aligned} \tag{34}$$

We can estimate, for any $0 < \epsilon \leq 1$,

$$\begin{aligned} |\eta^m \nabla\eta\cdot\nabla z||\nabla z|^{m-1}\overline{z}^\beta &\leq |\nabla z|^m|\nabla\eta|\eta^m\overline{z}^\beta \\ &\leq \frac{m\epsilon}{1+m}|\nabla z|^{1+m}\eta^{1+m}\overline{z}^{\beta-1} + \frac{\epsilon^{-m}}{1+m}|\nabla\eta|^{1+m}\overline{z}^{\beta+m}. \end{aligned} \tag{35}$$

By setting $\epsilon = \min\{1, |\beta|/3\}$, we obtain from (34) and (35)

$$\begin{aligned} \int_\Omega \eta^{1+m}\overline{z}^{\beta-1}|\nabla z|^{1+m} &\leq C(|\beta|)\int_\Omega \overline{z}^{\beta+m}(\overline{z}^{1-m}\eta^{1+m} + |\nabla\eta|^{1+m})dx \\ &\leq C(|\beta|)(||\overline{z}||_\infty^{1-m}+1)\int_\Omega \overline{z}^{\beta+m}(\eta^{1+m} + |\nabla\eta|^{1+m})dx \\ &\leq C(|\beta|, ||\overline{z}||_\infty)\int_\Omega \overline{z}^{\beta+m}(\eta^{1+m} + |\nabla\eta|^{1+m})dx, \end{aligned} \tag{36}$$

where $C(|\beta|)$ is bounded provided $|\beta|$ is bounded away from zero and $C(|\beta|, ||\overline{z}||_\infty) = C(|\beta|)(||\overline{z}||_\infty^{1-m}+1)$. Let $\delta = \beta + 1$,

$$w = \begin{cases} \overline{z}^{\frac{\beta+m}{1+m}} & \text{if } \beta \neq -m, \\ \log\overline{z} & \text{if } \beta = -m. \end{cases}$$

We may rewrite (36) as

$$\int_\Omega |\eta\nabla w|^{1+m}dx \leq \delta^{1+m}C(|\beta|, ||\overline{z}||_\infty)\int_\Omega w^{1+m}(\eta^{1+m} + |\nabla\eta|^{1+m})dx. \tag{37}$$

Then, we find from the Sobolev inequality

$$||\eta w||_{N(1+m)/(N-1-m)}^{1+m} \leq C\int_\Omega (|\eta\nabla w|^{1+m} + |w\nabla\eta|^{1+m})dx. \tag{38}$$

Hence, substituting (37) into (38), we get

$$||\eta w||_{N(1+m)/(N-1-m)} \leq C(1+\delta)||(\eta + |\bigtriangledown \eta|)w||_{1+m}. \tag{39}$$

We now specify the cutoff function η more precisely. Let $0 < r_1 < r_2$ be two positive numbers. Set $\eta = 1$ in B_{r_1}, $\eta = 0$ in $\Omega \setminus B_{r_2}$ with $|\bigtriangledown \eta| \leq 2/(r_2 - r_1)$. Writing $\sigma = N/(N-1-m)$, we have

$$||w||_{L^{\sigma(1+m)}(B_{r_1})} \leq \frac{C(1+\delta)}{r_2 - r_1}||w||_{L^{(1+m)}(B_{r_2})}. \tag{40}$$

Thus we can start an iteration procedure and follow exactly the argument of Theorems 8.17 and 8.18 in [10] to derive (32).

We now prove (ii). Suppose the contrary, that $w(\cdot, s)$ do not tend to $R(\cdot)$, where $R(x)$ is the unique positive classical solution. We note that R is a minimizer of Problem A, where

$$K = \int_\Omega R^2(x)\,dx = \lim_{s\to\infty}\int_\Omega w^2(x,s)\,dx. \tag{41}$$

Since $||w(\cdot,s)||_{1,1+m} \leq M$ for some constant $M > 0$, we can extract a sequence of times S_n such that $w(\cdot, S_n)$ tends weakly in $W_0^{1,1+m}(\Omega)$ to a function $Q \neq R$. Moreover, $K = \int_\Omega Q^2(x)\,dx$ by (21) and (41). Since $\int_\Omega |\bigtriangledown h|^{1+m}\,dx$ is lower semicontinuous with respect to weak convergence, we also have

$$\int_\Omega |\bigtriangledown Q|^{1+m}\,dx \leq \liminf_{n\to\infty}\int_\Omega |\bigtriangledown w(x,S_n)|^{1+m}\,dx. \tag{42}$$

But from the proof of (i), we see that

$$f(s) \downarrow \frac{1}{1+m}\int_\Omega |\bigtriangledown R|^{1+m}\,dx - \frac{\lambda}{2}\int_\Omega |\bigtriangledown R|^2\,dx \tag{43}$$

as $s \to \infty$. Hence

$$\int_\Omega |\bigtriangledown Q|^{1+m}\,dx \leq \int_\Omega |\bigtriangledown R|^{1+m}\,dx. \tag{44}$$

Therefore, Q is also a minimizer for Problem A, which is a contradiction. Thus $w(\cdot, s)$ converge to $R(\cdot)$ in $W_0^{1,1+m}(\Omega)$ as $s \to \infty$. This completes the proof of (ii).

REFERENCES

[1] D. G. Aronson and L. A. Peletier, *Large time behaviour of solutions of the porous media equation in bounded domains*, J. Differential Equations, 39 (1981), pp. 378–412.

[2] G. Astarita and G. Marrucci, *Principles of non-Newtonian fluid mechanics*, McGraw–Hill, New York, 1974.

[3] C. Bandle, C. V. Coffman, and M. Marcus, *Nonlinear elliptic problems in annular domains*, J. Differential Equations, 69 (1987), pp. 322–345.

[4] P. Benilan and M. G. Crandall, *The continuous dependence on ϕ of solution of $u_t - \triangle\phi(u) = 0$*, Indiana Univ. Math. J., 30 (1981), pp. 161–177.

[5] J. G. Berryman and C. J. Holland, *Stability of the separable solution for fast diffusion*, Arch. Rational Mech. Anal., 74 (1980), pp. 379–388.

[6] Y. Z. Chen, *Holder continuity of the gradients of solutions of nonlinear degenerate parabolic systems*, Acta Math. Sinica (N.S.), 2 (1986), pp. 309–331.

[7] Y. Z. Chen and E. Di Benedetto, *On the local behaviour of solutions of singular parabolic equations*, Arch. Rational Mech. Anal., 103 (1988), pp. 319–345.

[8] ———, *Boundary estimates for solutions of nonlinear degenerate parabolic systems*, J. Reine Angew. Math., 395 (1989), pp. 102–131.

[9] E. Di Benedetto, $C^{1+\alpha}$ *local regularity of weak solutions of degenerate elliptic equations*, Nonlinear Anal., 7 (1983), pp. 827–850.

[10] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[11] M. A. Herrero and J. L. Vazquez, *On the propagation properties of a nonlinear degenerate parabolic equation*, Comm. Partial Differential Equations, 7 (1982), pp. 1381–1402.

[12] A. S. Kalashnikov, *On a nonlinear equation appearing in the theory of non-stationary filtration*, Trudy Sem. Petrovsk., 4 (1978).

[13] Y. C. Kwong, *Asymptotic behaviour of a plasma type equation with finite extinction*, Arch. Rational Mech. Anal., 104 (1988), pp. 277–294.

[14] L. K. Martinson and K. B. Pavlov, *Unsteady shear flows of a conducting fluid with a rheological power law*, Magnit. Gidrodinamika, 2 (1971), pp. 50–58.

[15] W.-M. Ni and J. Serrin, *Existence and non-existence theorems for quasi-linear partial differential equations: The anormalous case*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl., 77 (1986), pp. 231–257.

[16] Y. W. Qi, *The existence of solutions to* $div(|\bigtriangledown u|^{m-1}\bigtriangledown u) + f(u) = 0$ *in annulus*, preprint, 1993.

[17] J. Serrin, *Local behaviour of solutions of quasilinear elliptic equations*, Acta Math., 111 (1964), pp. 247–302.

SIAM J. MATH. ANAL.
Vol. 27, No. 2, pp. 486–514, March 1996

009

ASYMPTOTIC ANALYSIS OF THE BOUNDARY LAYER FOR THE REISSNER–MINDLIN PLATE MODEL*

DOUGLAS N. ARNOLD† AND RICHARD S. FALK‡

Abstract. We investigate the structure of the solution of the Reissner–Mindlin plate equations in its dependence on the plate thickness in the cases of soft and hard clamped, soft and hard simply supported, and traction free boundary conditions. For the transverse displacement, rotation, and shear stress, we develop asymptotic expansions in powers of the plate thickness. These expansions are uniform up to the boundary for the transverse displacement, but for the other variables there is a boundary layer, which is stronger for the soft simply supported and traction-free plate and weaker for the soft clamped plate than for the hard clamped and hard simply supported plate. We give rigorous error bounds for the errors in the expansions in Sobolev norms. As an application, we derive new regularity results for the solutions and new estimates for the difference between the Reissner–Mindlin solution and the solution to the corresponding biharmonic model.

Key words. Reissner, Mindlin, plate, boundary layer

AMS subject classifications. 73K10, 73K25

1. Introduction.

The Reissner–Mindlin model for the bending of an isotropic elastic plate in equilibrium determines ω, the transverse displacement of the midplane, and ϕ, the rotation of fibers normal to the midplane, as the solution of the partial differential equations

$$-t^3 \,\mathbf{div}\, \mathcal{C}\,\mathcal{E}(\boldsymbol{\phi}) - \lambda t\,(\mathbf{grad}\,\omega - \boldsymbol{\phi}) = \boldsymbol{F},$$
$$-\lambda t \,\mathrm{div}\,(\mathbf{grad}\,\omega - \boldsymbol{\phi}) = G.$$

Here $\boldsymbol{F}$ is the applied couple per unit area, G is the applied transverse load density per unit area, t is the plate thickness, $\lambda = Ek/2(1+\nu)$ with E the Young's modulus, ν the Poisson ratio, and k the shear correction factor, $\mathcal{E}(\boldsymbol{\phi})$ is the symmetric part of the gradient of $\boldsymbol{\phi}$, and the fourth-order tensor $\mathcal{C}$ is defined by

$$\mathcal{C}\mathcal{T} = D\left[(1-\nu)\mathcal{T} + \nu\,\mathrm{tr}(\mathcal{T})\mathcal{I}\right], \quad D = \frac{E}{12(1-\nu^2)},$$

for any 2×2 matrix $\mathcal{T}$ ($\mathcal{I}$ denotes the 2×2 identity matrix). These equations are satisfied on the plane region Ω occupied by the midsection of the plate. In this paper, we investigate the dependence on the plate thickness of solutions to some boundary value problems associated to these equations.

[*] Received by the editors March 10, 1993; accepted for publication (in revised form) August 2, 1994.

[†] Department of Mathematics, Pennsylvania State University University Park, PA 16802. The research of this author was supported by National Science Foundation grants DMS-86-01489, DMS-8902433, and DMS-9205300.

[‡] Department of Mathematics, Rutgers University, New Brunswick, NJ 08903. The research of this author was supported by National Science Foundation grants DMS-87-03354, DMS-8902120, and DMS-9106051.

We consider various homogeneous boundary conditions of physical interest:

$$\boldsymbol{\phi}\cdot\boldsymbol{n}=\boldsymbol{\phi}\cdot\boldsymbol{s}=\omega=0 \qquad \text{(hard clamped)}, \tag{1.1}$$

$$\boldsymbol{\phi}\cdot\boldsymbol{n}=M_{\boldsymbol{s}}(\boldsymbol{\phi})=\omega=0 \qquad \text{(soft clamped)}, \tag{1.2}$$

$$M_{\boldsymbol{n}}(\boldsymbol{\phi})=\boldsymbol{\phi}\cdot\boldsymbol{s}=\omega=0 \qquad \text{(hard simply supported)}, \tag{1.3}$$

$$M_{\boldsymbol{n}}(\boldsymbol{\phi})=M_{\boldsymbol{s}}(\boldsymbol{\phi})=\omega=0 \qquad \text{(soft simply supported)}, \tag{1.4}$$

$$M_{\boldsymbol{n}}(\boldsymbol{\phi})=M_{\boldsymbol{s}}(\boldsymbol{\phi})=\partial\omega/\partial n-\boldsymbol{\phi}\cdot\boldsymbol{n}=0 \qquad \text{(free)}, \tag{1.5}$$

in which $\boldsymbol{n}$ and $\boldsymbol{s}$ denote the unit normal and counterclockwise tangent vectors, respectively, and $M_{\boldsymbol{n}}(\boldsymbol{\phi}) := \boldsymbol{n}\cdot \mathcal{C}\,\mathcal{E}(\boldsymbol{\phi})\boldsymbol{n}$, $M_{\boldsymbol{s}}(\boldsymbol{\phi}) := \boldsymbol{s}\cdot \mathcal{C}\,\mathcal{E}(\boldsymbol{\phi})\boldsymbol{n}$. Each of the first four boundary value problems admits a unique solution $\omega \in H^1(\Omega)$, $\boldsymbol{\phi}\in \boldsymbol{H}^1(\Omega)$ for any $\boldsymbol{F}\in \boldsymbol{L}^2(\Omega)$ and $G \in L^2(\Omega)$. The existence theory for the free plate is slightly more complicated and will be discussed in §6.

We do not treat the Reissner–Mindlin model in its full generality. In addition to the assumption of homogeneous boundary conditions, we shall assume that there is no applied couple, so $\boldsymbol{F}\equiv 0$, and that the constitutive parameters E, ν, and k are independent of t. It seems clear that the techniques developed here apply to more general situations as well.

We also suppose that $G = gt^3$, where the function g does not depend on t. This is a convenient normalization, which leads to $\boldsymbol{\phi}$ and ω having a nonzero limit as t tends to zero. Given that the first differential equation and the boundary conditions are taken to be homogeneous, this normalization is not restrictive. If G were to be proportional to some other function $h(t)$, we could make the change of dependent variables $\bar{\boldsymbol{\phi}} = t^3\boldsymbol{\phi}/h(t)$, $\bar{\omega} = t^3\omega/h(t)$ and the new variables would satisfy the Reissner–Mindlin equations with load proportional to t^3.

With these assumptions, the Reissner–Mindlin equations become

$$-\operatorname{\mathbf{div}} \mathcal{C}\,\mathcal{E}(\boldsymbol{\phi}) - \lambda t^{-2}(\operatorname{\mathbf{grad}}\omega - \boldsymbol{\phi}) = 0, \tag{1.6}$$

$$-\lambda t^{-2}\operatorname{div}(\operatorname{\mathbf{grad}}\omega - \boldsymbol{\phi}) = g. \tag{1.7}$$

After a similar normalization of the load, the biharmonic model for plate bending may be written

$$D\,\Delta^2\omega_0 = g \quad \text{in } \Omega,$$

and so its solution ω_0 is independent of the plate thickness. In contrast, the solution of the Reissner–Mindlin model exhibits a complex dependence on the plate thickness, which we investigate in the present paper. In previous work [1], we gave an analysis of the boundary layer for the Reissner–Mindlin model of hard clamped and hard simply supported plates. There are many additional complications in the case of more general boundary conditions, and so the analysis of [1] is not easily extended to the soft simply supported and free plates, for example. In this paper, we analyze the boundary layer for all the boundary conditions mentioned above in a unified fashion. While the approach here is more complete, it is also simpler than that of [1] in a number of ways. Thus the present paper essentially supersedes that one. We shall show that the boundary layer is strongest for the soft simply supported and free plate, somewhat weaker for the clamped and hard simply supported plate, and weakest for the soft clamped plate. In addition, we shall demonstrate that for the soft clamped and hard simply supported plates, there is no boundary layer near a flat portion of the boundary.

We shall develop asymptotic expansions with respect to t for ω and ϕ (as well as for other quantities associated with the solution such as the shear strain). The expansions take the forms

$$\omega \sim \omega_0 + t\omega_1 + t^2\omega_2 + \cdots ,$$
$$\phi \sim \phi_0 + \chi\boldsymbol{\Phi}_0 + t(\phi_1 + \chi\boldsymbol{\Phi}_1) + t^2(\phi_2 + \chi\boldsymbol{\Phi}_2) + \cdots ,$$

where the interior expansion functions ω_i and ϕ_i are independent of t and the boundary correctors $\boldsymbol{\Phi}_i$ depend on t only through the quantity ρ/t, ρ being the distance of a point of Ω from the boundary. More specifically,

$$\boldsymbol{\Phi}_i = \hat{\boldsymbol{\Phi}}_i(\rho/t, \theta),$$

where θ is a coordinate which roughly gives arc length parallel to the boundary (see §2), and the function $\hat{\boldsymbol{\Phi}}_i(\eta, \theta)$ has the form of a polynomial with respect to η with coefficients depending smoothly on θ times $\exp(-\sqrt{12k}\eta)$. Thus $\boldsymbol{\Phi}_i$ represents a boundary-layer function, which essentially lives in a strip of width t around the boundary. Finally, χ is a cutoff function which is independent of t and identically equal to unity in a neighborhood of $\partial\Omega$.

After some preliminary material in §2, we construct the terms of the asymptotic expansions in §3 (for all of the boundary conditions except for those of the free plate, which are treated in §6). Then, in the following two sections, we justify the expansions rigorously in the case of the soft simply supported plate, proving a priori bounds for the terms of the expansions in §4 and performing the error analysis in §5. This analysis can be adapted easily to the cases of hard simply supported and hard and soft clamped plates and somewhat less easily to the case of the free plate. The necessary modifications are discussed in §6. To make it easier for the reader to follow some of the computations performed in the derivation and analysis of the asymptotic expansions, we have included in an appendix a summary of the main formulas we have used. In the remainder of this introduction, we summarize some of the principal results.

For each of the boundary conditions, ω_0 is the solution of the biharmonic equation

$$D\Delta^2\omega_0 = g$$

determined by appropriate boundary conditions, namely

$$\omega_0 = \frac{\partial\omega_0}{\partial n} = 0$$

for the hard and soft clamped plates,

$$\omega_0 = (1-\nu)\frac{\partial^2\omega_0}{\partial n^2} + \nu\Delta\omega_0 = 0$$

for the hard and soft simply supported plates, and

$$(1-\nu)\frac{\partial^2\omega_0}{\partial n^2} + \nu\Delta\omega_0 = \frac{\partial\Delta\omega_0}{\partial n} + (1-\nu)\frac{\partial}{\partial s}\left(\frac{\partial^2\omega_0}{\partial s\partial n} - \kappa\frac{\partial\omega_0}{\partial s}\right) = 0$$

for the free plate. In the last expression, κ denotes the curvature of the boundary.

The next term in the expansion of the transverse displacement, ω_1, vanishes for the hard and soft clamped plates and the hard simply supported plate but not for the soft simply supported or free plates. In these cases, it is the solution of the homogeneous biharmonic problem

$$\Delta^2\omega_1 = 0 \quad \text{in } \Omega$$

with the inhomogeneous boundary conditions

$$\omega_1 = 0, \quad (1-\nu)\frac{\partial^2\omega_1}{\partial n^2} + \nu\,\Delta\,\omega_1 = -\frac{(1-\nu)}{\sqrt{3k}}\,\frac{\partial^3\omega_0}{\partial s^2\partial n}$$

for the soft simply supported plate and

$$(1-\nu)\frac{\partial^2\omega_1}{\partial n^2} + \nu\,\Delta\,\omega_1 = \frac{1}{\sqrt{3k}}\,\frac{\partial\,\Delta\,\omega_0}{\partial n},$$

$$\frac{\partial\,\Delta\,\omega_1}{\partial n} + (1-\nu)\frac{\partial}{\partial s}\left(\frac{\partial^2\omega_1}{\partial s\partial n} - \kappa\frac{\partial\omega_1}{\partial s}\right) = -\frac{(1-\nu)}{\sqrt{3k}}\,\frac{\partial}{\partial s}\left(\kappa\left[\frac{\partial^2\omega_0}{\partial s\partial n} - \kappa\frac{\partial\omega_0}{\partial s}\right]\right)$$

for the free plate.

Note that the expansions for the soft and hard simply supported plates differ already in the term ω_1. For the soft and hard clamped plates the terms ω_0, ω_1, and ω_2 all agree, but $\omega_3 = 0$ for the soft clamped plate and is generally nonzero for the hard clamped plate.

Turning to the expansion of ϕ, we find that in all five cases that $\phi_0 = \mathbf{grad}\,\omega_0$ and $\phi_1 = \mathbf{grad}\,\omega_1$ while $\phi_2 - \mathbf{grad}\,\omega_2 = \lambda^{-1}D\,\Delta\,\omega_0$, which is never zero (except in the trivial case $g \equiv 0$). For the boundary correctors, we find that $\boldsymbol{\Phi}_0$ vanishes in all five cases. For the soft simply supported and free plates,

$$\hat{\boldsymbol{\Phi}}_1(\eta,\theta) = -\frac{1}{\sqrt{3k}}\exp(-\sqrt{12k}\eta)\left(\frac{\partial^2\omega_0}{\partial s\partial n} - \kappa\frac{\partial\omega_0}{\partial s}\right)(0,\theta)\boldsymbol{s}.$$

For the hard clamped and hard simply supported plate, $\boldsymbol{\Phi}_1$ vanishes as well as $\boldsymbol{\Phi}_0$ and we have

$$\hat{\boldsymbol{\Phi}}_2(\eta,\theta) = -\frac{1}{6k(1-\nu)}\exp(-\sqrt{12k}\eta)\frac{\partial}{\partial s}\,\Delta\,\omega_0(0,\theta)\boldsymbol{s}$$

in both cases. For the soft clamped plate $\boldsymbol{\Phi}_0$, $\boldsymbol{\Phi}_1$, and $\boldsymbol{\Phi}_2$ all vanish. In all five cases, the first nonzero boundary corrector is purely tangential. Table 1 summarizes the terms in the asymptotic expansions of ω and ϕ which vanish.

TABLE 1
Vanishing terms in the asymptotic expansions.

soft simply supported	—	—	$\boldsymbol{\Phi}_0$	$\boldsymbol{\Phi}_1 \cdot \boldsymbol{n}$
free	—	—	$\boldsymbol{\Phi}_0$	$\boldsymbol{\Phi}_1 \cdot \boldsymbol{n}$
hard clamped	ω_1	ϕ_1	$\boldsymbol{\Phi}_0$, $\boldsymbol{\Phi}_1$	$\boldsymbol{\Phi}_2 \cdot \boldsymbol{n}$
hard simply supported	ω_1	ϕ_1	$\boldsymbol{\Phi}_0$, $\boldsymbol{\Phi}_1$	$\boldsymbol{\Phi}_2 \cdot \boldsymbol{n}$
soft clamped	ω_1, ω_3	ϕ_1, ϕ_3	$\boldsymbol{\Phi}_0$, $\boldsymbol{\Phi}_1$, $\boldsymbol{\Phi}_2$	$\boldsymbol{\Phi}_3 \cdot \boldsymbol{n}$

Using symbolic computation, we have computed exact solutions to the Reissner–Mindlin system on circular and semiinfinite plates for particular choices of the load function g, and have explicitly computed the asymptotic expansions of ω and ϕ through terms of order 6. These computations verify the sharpness of the results in this paper in that no terms of the expansions vanish except those given in the table. These results have been reported in [2].

As an application of our asymptotic analysis, we can determine the asymptotic behavior of Sobolev norms of solutions of the Reissner–Mindlin system. Supposing that g is sufficiently smooth, we have the following estimates, valid for both the soft simply supported and free plate, in which the constant C depends on g, Ω, and the elastic constants but is independent of t. Here $\|\cdot\|_s$ and $|\cdot|_s$ denote the norms in the Sobolev spaces $H^s(\Omega)$ and $H^s(\partial\Omega)$ (see §2).

The transverse displacement ω and all of its derivatives are bounded uniformly in t, that is,

$$\|\omega\|_s \le C, \quad s \in \mathbb{R},$$

but the regularity of the rotation $\boldsymbol{\phi}$ is limited by the boundary layer. For example, for the soft simply supported and free plates, we have

$$\|\boldsymbol{\phi}\|_s \le Ct^{\min(0,3/2-s)}, \quad s \in \mathbb{R},$$

so derivatives of order greater than 1 will generally tend to infinity in L^2 as $t \to 0$.

The quantity $\boldsymbol{\zeta} := t^{-2}(\mathbf{grad}\,\omega - \boldsymbol{\phi})$, which is proportional to the shear strain, is often of interest. From the above expansions, we get

$$\boldsymbol{\zeta} \sim -t^{-1}\chi\boldsymbol{\Phi}_1 + (\mathbf{grad}\,\omega_2 - \boldsymbol{\phi}_2 - \chi\boldsymbol{\Phi}_2) + \cdots,$$

so it has a stronger boundary layer. Indeed, for the soft simply supported and free plates, $\boldsymbol{\zeta}$ is not uniformly bounded in L^2, or even in H^s for $s > -1/2$:

$$\|\boldsymbol{\zeta}\|_s \le Ct^{\min(0,-1/2-s)}, \quad s \in \mathbb{R}.$$

The corresponding estimates for the hard clamped and hard simply supported plates are

$$\|\boldsymbol{\phi}\|_s \le Ct^{\min(0,5/2-s)}, \quad s \in \mathbb{R}, \qquad \|\boldsymbol{\zeta}\|_s \le Ct^{\min(0,1/2-s)}, \quad s \in \mathbb{R},$$

and for the soft clamped plate

$$\|\boldsymbol{\phi}\|_s \le Ct^{\min(0,7/2-s)}, \quad s \in \mathbb{R}, \qquad \|\boldsymbol{\zeta}\|_s \le Ct^{\min(0,3/2-s)}, \quad s \in \mathbb{R}.$$

Of course, the boundary layer does not limit the regularity of $\boldsymbol{\phi}$ or $\boldsymbol{\zeta}$ at a positive distance from $\partial\Omega$ nor does it affect the smoothness of their restrictions to $\partial\Omega$. Thus

$$\|\boldsymbol{\phi}\|_{H^s(\Omega_c)} + |\boldsymbol{\phi}|_s + \|\boldsymbol{\zeta}\|_{H^s(\Omega_c)} + |\boldsymbol{\zeta}|_s \le C, \quad s \in \mathbb{R},$$

for any compact subdomain Ω_c of Ω.

In the limit as $t \to 0$, the variables ω and $\boldsymbol{\phi}$ tend in L^2 to the leading terms of their asymptotic expansions. The number of derivatives which converge and the rate of convergence may be determined by examining the first neglected interior and boundary terms of the expansions. For any $s \in \mathbb{R}$, we get for the soft simply supported and free plate

$$\|\omega - \omega_0\|_s \le Ct, \qquad \|\boldsymbol{\phi} - \boldsymbol{\phi}_0\|_s \le Ct^{\min(1,3/2-s)}.$$

Note that the rate of convergence for $\boldsymbol{\phi}$ depends on the Sobolev norm under consideration. For each of the variables, taking more terms from the expansion increases the rate of convergence and taking sufficiently many terms in the expansions gives approximation of any desired algebraic order of convergence in t in any desired Sobolev space (provided g is sufficiently regular). For example,

$$\|\omega - \omega_0 - t\omega_1\|_s \le Ct^2, \qquad \|\boldsymbol{\phi} - \boldsymbol{\phi}_0 - t(\boldsymbol{\phi}_1 + \chi\boldsymbol{\Phi}_1)\|_s \le Ct^{\min(2,5/2-s)}.$$

For the hard clamped and hard simply supported plates, the analogous results are

$$\|\omega - \omega_0\|_s \le Ct^2, \qquad \|\boldsymbol{\phi} - \boldsymbol{\phi}_0\|_s \le Ct^{\min(2,5/2-s)},$$
$$\|\omega - \omega_0 - t^2\omega_2\|_s \le Ct^3, \qquad \|\boldsymbol{\phi} - \boldsymbol{\phi}_0 - t^2(\boldsymbol{\phi}_2 + \chi\boldsymbol{\Phi}_2)\|_s \le Ct^{\min(3,7/2-s)}.$$

For the soft clamped plate,

$$\|\omega-\omega_0\|_s \le Ct^2, \qquad \|\phi-\phi_0\|_s \le Ct^{\min(2,7/2-s)},$$
$$\|\omega-\omega_0-t^2\omega_2\|_s \le Ct^3, \qquad \|\phi-\phi_0-t^2\phi_2\|_s \le Ct^{\min(3,7/2-s)}.$$

It is also possible to use our asymptotic expansions to derive estimates in function spaces other than H^s. The technique for doing this is described in [1]. Further references for the Reissner–Mindlin model and its boundary-layer behavior can also be found there. Many of the results in this paper were described without proof in [2], where explicit illustrations of the theory are constructed.

2. Notation and preliminaries. The letter C denotes a generic constant, not necessarily the same in each occurrence. We assume that Ω is a smooth, bounded, and simply connected domain in $\mathbb{R}^2$. The $L^2(\Omega)$ and $L^2(\partial\Omega)$ inner products are denoted by $(\cdot\,,\,\cdot)$ and $\langle\cdot\,,\,\cdot\rangle$, respectively. We also use the usual L^2-based Sobolev spaces $H^s(\Omega)$ and $H^s(\partial\Omega)$, real $s\ge 0$, with norms denoted by $\|\cdot\|_s$ and $|\cdot|_s$. When the domain argument is omitted, L^2 and H^s refer to $L^2(\Omega)$ and $H^s(\Omega)$. The space $\mathring{H}^s=\mathring{H}^s(\Omega)$ is the closure of C_0^∞ in H^s. The interpolation inequality

$$\|g\|_{s+v}^u \le C\|g\|_s^{u-v}\|g\|_{s+u}^v, \quad s\ge 0,\ u\ge v\ge 0, \tag{2.1}$$

holds. If $g\in L^2$ and $\Delta^{-1}g$ denotes the unique function in $H^2\cap\mathring{H}^1$ whose Laplacian is equal to g, then

$$C^{-1}\|\,\Delta^{-1}g\|_{s+2} \le \|g\|_s \le C\|\,\Delta^{-1}g\|_{s+2}, \quad s\ge 0,$$

where the constant C may depend on s and Ω but not on g. In other words, $g\mapsto \|\,\Delta^{-1}g\|_{s+2}$ defines an equivalent norm on H^s for $s\ge 0$. We also define some negatively indexed norms which maintain this equivalence:

$$\|g\|_s := \|\,\Delta^{-1}g\|_{s+2}, \quad -2\le s<0.$$

For $s=-1$, this is equivalent to the norm in the dual space of $\mathring{H}^1$. For $s=-2$, it is equivalent to the norm in the dual space of $H^2\cap\mathring{H}^1$. With this definition, (2.1) holds for $s\ge -2$. We shall make frequent use of this fact to bound sums of the form $\sum_{i=0}^n t^i\|g\|_{s+i}$ by a multiple of the sum of the first and last terms.

We also require the quotient space $H^s/\mathbb{R}$. An element $p\in H^s/\mathbb{R}$ is a coset consisting of all functions in H^s differing from a fixed function by a constant. The quotient norm is given by

$$\|p\|_{s/\mathbb{R}} = \min_{q\in p}\|q\|_s.$$

(In fact, $\|p\|_{s/\mathbb{R}}=\|\bar{p}\|_s$, where $\bar{p}$ is the unique function in the coset p having mean value zero.)

We use boldface type to denote 2-vector-valued functions, operators whose values are vector-valued functions, and spaces of vector-valued functions. Script type is used in a similar way for 2 × 2-matrix objects. Thus, for example, $\operatorname{div}\boldsymbol{\psi}\in L^2$ for $\boldsymbol{\psi}\in\boldsymbol{H}^1$, while $\mathbf{div}\,\mathcal{T}\in\boldsymbol{L}^2$ for $\mathcal{T}\in\mathcal{H}^1$. Finally, we use various standard differential operators:

$$\mathbf{grad}\,r=\begin{pmatrix}\partial r/\partial x\\ \partial r/\partial y\end{pmatrix}, \qquad \operatorname{div}\boldsymbol{\psi}=\frac{\partial\psi_1}{\partial x}+\frac{\partial\psi_2}{\partial y},$$

$$\mathbf{div}\begin{pmatrix}t_{11} & t_{12}\\ t_{21} & t_{22}\end{pmatrix}=\begin{pmatrix}\partial t_{11}/\partial x+\partial t_{12}/\partial y\\ \partial t_{21}/\partial x+\partial t_{22}/\partial y\end{pmatrix},$$

$$\mathbf{curl}\, p = \begin{pmatrix} -\partial p/\partial y \\ \partial p/\partial x \end{pmatrix}, \qquad \operatorname{rot} \psi = \frac{\partial \psi_1}{\partial y} - \frac{\partial \psi_2}{\partial x}.$$

Note that these differential operators annihilate constants and consequently induce operators on the quotient space $H^s/\mathbb{R}$ for each s. We denote the induced operator in the same way as the original. Thus, for example, if $p \in H^1/\mathbb{R}$, $\mathbf{curl}\, p$ denotes the element of $\boldsymbol{L}^2$ obtained by applying the curl to any element in the coset p.

We record here for later reference the identity

$$\sum_{i=0}^{n}\sum_{j=0}^{i} f(i-j,j) = \sum_{i=0}^{n}\sum_{j=0}^{n-i} f(i,j). \tag{2.2}$$

To describe the boundary layer, we define the usual boundary-fitted coordinates in a neighborhood of the boundary. Let ρ_0 be a positive number less than the minimum radius of curvature of $\partial\Omega$ and define

$$\Omega_0 = \{\, \boldsymbol{z} - \rho \boldsymbol{n}_{\boldsymbol{z}} \mid \boldsymbol{z} \in \partial\Omega, 0 < \rho < \rho_0 \,\},$$

where $\boldsymbol{n}_{\boldsymbol{z}}$ is the outward unit normal to Ω at $\boldsymbol{z}$. Let $\boldsymbol{z}(\theta) = (X(\theta), Y(\theta))$, $\theta \in [0, L)$, be a parametrization of $\partial\Omega$ by arclength which we extend L-periodically to $\theta \in \mathbb{R}$. The correspondence

$$(\rho, \theta) \mapsto \boldsymbol{z} - \rho \boldsymbol{n}_{\boldsymbol{z}} = (X(\theta) - \rho Y'(\theta), Y(\theta) + \rho X'(\theta))$$

is a diffeomorphism of $(0, \rho_0) \times \mathbb{R}/L$ on Ω_0. Let $\kappa(\theta)$ denote the curvature of $\partial\Omega$ at $\boldsymbol{z}(\theta)$ and set

$$\sigma(\rho, \theta) := \frac{1}{1 - \kappa(\theta)\rho}.$$

The unit vector fields of the outward normal and the counterclockwise tangent extend from $\partial\Omega$ to Ω_0 as functions of θ, independent of ρ, and satisfy

$$\boldsymbol{n} = -\,\mathbf{grad}\, \rho = -\sigma(\rho, \theta)^{-1}\, \mathbf{curl}\, \theta, \qquad \boldsymbol{s} = \sigma(\rho, \theta)^{-1}\, \mathbf{grad}\, \theta = -\,\mathbf{curl}\, \rho.$$

We shall also use the stretched variable $\hat{\rho} = \rho/t$. When required for clarity, we use hats to denote the change of variables to $(\hat{\rho}, \theta)$ coordinates, that is,

$$\hat{f}(\hat{\rho}, \theta) := f(x, y).$$

3. An asymptotic expansion of the solution. We now develop asymptotic expansions of $\boldsymbol{\phi}$ and ω with respect to the plate thickness. Such expansions normally consist of two parts, an interior expansion and a boundary-layer expansion. Now it follows easily from (1.6) and (1.7) that the transverse displacement ω satisfies the biharmonic equation

$$D\,\Delta^2 \omega = g - \lambda^{-1} D t^2\, \Delta\, g, \tag{3.1}$$

which indicates that ω admits no boundary layer and hence can be described by an interior expansion alone. However, the rotation vector $\boldsymbol{\phi}$ satisfies the singular perturbation equation given in (1.6) and hence can be expected to include a boundary layer. Thus we shall seek expansions of the form

$$\omega \sim \sum_{i=0}^{\infty} t^i \omega_i, \qquad \boldsymbol{\phi} \sim \sum_{i=0}^{\infty} t^i \boldsymbol{\phi}_i + \sum_{i=1}^{\infty} t^i \boldsymbol{\Phi}_i,$$

where ω_i and $\boldsymbol{\phi}_i$ are smooth functions independent of t, while $\boldsymbol{\Phi}_i(x,y) = \hat{\boldsymbol{\Phi}}_i(\hat{\rho},\theta)$ with $\hat{\boldsymbol{\Phi}}$ a smooth function on $[0,\infty) \times \partial\Omega$. We have suppressed the term $\boldsymbol{\Phi}_0$ since it turns out to be zero in all cases. In order that the expansion for $\boldsymbol{\phi}$ is defined everywhere in Ω even though $\boldsymbol{\Phi}_i$ is defined only on Ω_0, we introduce a smooth cutoff function χ which is a function of ρ alone, independent of θ and t, and identically one for $0 \le \rho \le \rho_0/3$, identically zero for $\rho > 2\rho_0/3$.

In this section, we give precise definitions of all the functions $\boldsymbol{\phi}_i$, ω_i, and $\boldsymbol{\Phi}_i$. In §5 (Theorems 5.1–5.3), we shall prove the validity of the expansions. More precisely, we shall show that by choosing n large enough we can make the corresponding remainder terms

$$\omega_n^E := \omega - \sum_{i=0}^{n} t^i \omega_i, \qquad \boldsymbol{\phi}_n^E := \boldsymbol{\phi} - \sum_{i=0}^{n} t^i \boldsymbol{\phi}_i - \chi \sum_{i=1}^{n} t^i \boldsymbol{\Phi}_i$$

smaller than any desired power of t in any Sobolev norm.

Taking the divergence of (1.6) and using (1.7), we see that $\operatorname{div}\boldsymbol{\phi}$ satisfies Poisson's equation:

$$D\,\Delta \operatorname{div}\boldsymbol{\phi} = g.$$

This suggests an alternate form for the asymptotic expansion of $\boldsymbol{\phi}$ in which the terms of the boundary-layer expansion are divergence free and hence can be written as the curls of scalar functions. Inserting some convenient factors, the alternate expansion is

$$\boldsymbol{\phi} \sim \sum_{i=0}^{\infty} t^i \boldsymbol{\phi}_i - \lambda^{-1}\chi t^2 \sum_{i=0}^{\infty} t^i \,\mathbf{curl}\, P_i$$

with $P_i(x,y) = \hat{P}_i(\hat{\rho},\theta)$ with $\hat{P}_i : [0,\infty) \times \partial\Omega \to \mathbb{R}$ smooth. Now

$$\mathbf{curl}\, P_i = \frac{\partial P_i}{\partial \rho}\,\mathbf{curl}\,\rho + \frac{\partial P_i}{\partial \theta}\,\mathbf{curl}\,\theta = -t^{-1}\frac{\partial \hat{P}_i}{\partial \hat{\rho}}\boldsymbol{s} - \sigma(\rho,\theta)\frac{\partial \hat{P}_i}{\partial \theta}\boldsymbol{n}.$$

Formally inserting the Taylor expansion

$$\sigma(\rho,\theta) = \sum_{j=0}^{\infty} [\kappa(\theta)\rho]^j = \sum_{j=0}^{\infty} [\kappa(\theta)t\hat{\rho}]^j$$

and equating the two forms of the boundary-layer expansion, we get that

$$-\lambda^{-1} t^2 \sum_{i=0}^{\infty} t^i \,\mathbf{curl}\, P_i = \sum_{i=1}^{\infty} t^i \boldsymbol{\Phi}_i. \tag{3.2}$$

This gives the relation between $\boldsymbol{\Phi}_i$ and P_i:

$$\boldsymbol{\Phi}_i = \lambda^{-1}\left\{ \frac{\partial \hat{P}_{i-1}}{\partial \hat{\rho}}\boldsymbol{s} + \sum_{j=0}^{i-2} [\kappa(\theta)\hat{\rho}]^j \frac{\partial \hat{P}_{i-j-2}}{\partial \theta}\boldsymbol{n} \right\}, \quad i \ge 1. \tag{3.3}$$

We now proceed to the definitions of w_i, $\boldsymbol{\phi}_i$, and P_i (with $\boldsymbol{\Phi}_i$ determined from P_i by (3.3)).

In order to motivate the definitions of the expansion functions, we shall reason formally. Let $\boldsymbol{\phi}^I$ denote $\sum_{i=0}^{\infty} t^i \boldsymbol{\phi}_i$ and let p^B denote $\sum_{i=0}^{\infty} t^i P_i$ (these definitions are only formal, since the sums need not be convergent). We want pairs $(\boldsymbol{\phi}^I, \omega)$ to solve the Reissner–Mindlin differential equations and $-\lambda^{-1}t^2(\mathbf{curl}\, p^B, 0)$ to solve the corresponding homogeneous differential equations, so that the pair $(\boldsymbol{\phi}, \omega)$, which (when $\chi \equiv 1$) is formally their sum, will satisfy the inhomogeneous equations. Inserting the expansions for ϕ^I and ω into the Reissner–Mindlin equations and equating like powers of t gives the equations

$$\lambda(\phi_i - \mathbf{grad}\,\omega_i) = \mathbf{div}\, C\,\mathcal{E}(\phi_{i-2}), \tag{3.4}$$

$$\lambda\,\mathrm{div}(\phi_i - \mathbf{grad}\,\omega_i) = \delta_{i2}g, \tag{3.5}$$

where δ_{ij} is the Kronecker symbol. These equations are to hold for $i = 0, 1, \ldots$ with the convention that $\phi_j = 0$ for $j < 0$. From (3.4) and (3.5), we easily deduce that ω_i satisfies the biharmonic problem

$$D\,\Delta^2\,\omega_i = \delta_{i0}g - \delta_{i2}\lambda^{-1}D\,\Delta\,g, \tag{3.6}$$

as is to be expected in view of (3.1). It follows from (3.4) that

$$\phi_i = \mathbf{grad}\sum_{k=0}^{[i/2]}(\lambda^{-1}D)^k\,\Delta^k\,\omega_{i-2k}$$

or, in light of (3.6),

$$\phi_i = \mathbf{grad}\, z_i, \tag{3.7}$$

with

$$z_i = \omega_i + \lambda^{-1}D\,\Delta\,\omega_{i-2} + \delta_{i4}\lambda^{-2}Dg. \tag{3.8}$$

To obtain differential equations satisfied by the boundary-layer functions, we note that $(\mathbf{curl}\,P, 0)$ solves the homogeneous Reissner–Mindlin system if and only if P solves the differential equation

$$-t^2\lambda^{-1}D\frac{1-\nu}{2}\,\Delta\,P + P = 0. \tag{3.9}$$

In (ρ, θ) coordinates, we have (on Ω_0)

$$\begin{aligned}\Delta\,P &= |\,\mathbf{grad}\,\rho|^2\frac{\partial^2 P}{\partial\rho^2} + \Delta\,\rho\frac{\partial P}{\partial\rho} + |\,\mathbf{grad}\,\theta|^2\frac{\partial^2 P}{\partial\theta^2} + \Delta\,\theta\frac{\partial P}{\partial\theta}\\ &= \frac{\partial^2 P}{\partial\rho^2} - \kappa(\theta)\sigma(\rho,\theta)\frac{\partial P}{\partial\rho} + \sigma(\rho,\theta)^2\frac{\partial^2 P}{\partial\theta^2} + \rho\kappa'(\theta)\sigma(\rho,\theta)^3\frac{\partial P}{\partial\theta}\\ &= \frac{\partial^2 P}{\partial\rho^2} + \sum_{j=0}^{\infty}\rho^j\left(a_1^j\frac{\partial P}{\partial\rho} + a_2^j\frac{\partial^2 P}{\partial\theta^2} + a_3^j\frac{\partial P}{\partial\theta}\right).\end{aligned}$$

In the last step, we have (formally) replaced each coefficient with its Taylor series in ρ. It is easy to check that

$$a_1^j = -[\kappa(\theta)]^{j+1}, \qquad a_2^j = (j+1)[\kappa(\theta)]^j, \qquad a_3^j = \frac{j(j+1)}{2}[\kappa(\theta)]^{j-1}\kappa'(\theta). \tag{3.10}$$

Switching to the stretched variable $\hat{\rho}$, this becomes

$$\Delta\,P = t^{-2}\frac{\partial^2 P}{\partial\hat{\rho}^2} + \sum_{j=0}^{\infty}(t\hat{\rho})^j\left(a_1^j t^{-1}\frac{\partial P}{\partial\hat{\rho}} + a_2^j\frac{\partial^2 P}{\partial\theta^2} + a_3^j\frac{\partial P}{\partial\theta}\right).$$

Thus if we write (3.9) in $(\hat{\rho},\theta)$ variables, insert $\sum_{i=0}^{\infty} t^i P_i$ for P, and equate like powers of t, we get

$$\begin{aligned}(3.11)\qquad -\lambda^{-1}D\frac{1-\nu}{2}\frac{\partial^2 P_i}{\partial \hat{\rho}^2} + P_i &= \hat{F}_i(\hat{\rho},\theta)\\ &:= \lambda^{-1}D\frac{1-\nu}{2}\sum_{j=0}^{i-1}\hat{\rho}^j\left(a_1^j\frac{\partial P_{i-j-1}}{\partial\hat{\rho}} + a_2^j\frac{\partial^2 P_{i-j-2}}{\partial\theta^2} + a_3^j\frac{\partial P_{i-j-2}}{\partial\theta}\right), \quad i = 0,1,\dots,\end{aligned}$$

where again P_j is to be interpreted as 0 for $j < 0$. Note that (3.11) is an ordinary differential equation for the function P_i in the independent variable $\hat{\rho}$ in which θ enters as a parameter. We shall only consider solutions which satisfy the decay condition

$$(3.12)\qquad \lim_{\hat{\rho}\to\infty} P_i = 0.$$

This will ensure that each P_i decays exponentially with $\hat{\rho}$ and is therefore negligible outside of Ω_0.

The differential equations (3.6) and (3.11), together with appropriate boundary conditions, will be used to define the functions ω_i and P_i. Then the $\boldsymbol{\phi}_i$ are given by (3.7) and the $\boldsymbol{\Phi}_i$ by (3.3).

We now derive the boundary conditions and, for each of the boundary value problems we consider, show that the ω_i and P_i are uniquely determined. The boundary conditions for ω_i and P_i will be obtained from the boundary conditions for the Reissner–Mindlin system by inserting the asymptotic expansions and equating like powers of the thickness, and then using (3.7) to eliminate the $\boldsymbol{\phi}_i$.

The hard clamped plate. The boundary condition $\omega = 0$ leads, of course, to

$$(3.13)\qquad \omega_i = 0 \quad \text{on } \partial\Omega.$$

The boundary conditions $\boldsymbol{\phi}\cdot\boldsymbol{n} = 0$ and $\boldsymbol{\phi}\cdot\boldsymbol{s} = 0$ give $\boldsymbol{\phi}_i\cdot\boldsymbol{n} + \boldsymbol{\Phi}_i\cdot\boldsymbol{n} = 0$ and $\boldsymbol{\phi}_i\cdot\boldsymbol{s} + \boldsymbol{\Phi}_i\cdot\boldsymbol{s} = 0$. Using (3.3), these become

$$(3.14)\qquad \boldsymbol{\phi}_i\cdot\boldsymbol{n} = -\lambda^{-1}\frac{\partial P_{i-2}}{\partial\theta} \quad \text{on } \partial\Omega$$

and

$$(3.15)\qquad \boldsymbol{\phi}_i\cdot\boldsymbol{s} = -\lambda^{-1}\frac{\partial P_{i-1}}{\partial\hat{\rho}} \quad \text{on } \partial\Omega.$$

In view of (3.7), (3.14) can be expressed equivalently as

$$(3.16)\qquad \frac{\partial\omega_i}{\partial n} = -\lambda^{-1}\frac{\partial P_{i-2}}{\partial\theta} - \lambda^{-1}D\frac{\partial\,\Delta\,\omega_{i-2}}{\partial n} - \delta_{i4}\lambda^{-2}D\frac{\partial g}{\partial n} \quad \text{on } \partial\Omega,$$

and, using (3.13) and (3.7), we can write (3.15) as

$$(3.17)\qquad -\frac{\partial P_i}{\partial\hat{\rho}} = D\frac{\partial\,\Delta\,w_{i-1}}{\partial s} + \delta_{i3}\lambda^{-1}D\frac{\partial g}{\partial s} \quad \text{on } \partial\Omega.$$

We now show that all the ω_i and P_i are uniquely determined by (3.6), (3.11), (3.13)–(3.15), and (3.12). Indeed, from (3.11), (3.12), and (3.17), we immediately infer that $P_0 = 0$. We can then uniquely determine ω_i for $i = 0,1,2$ from (3.6), (3.13), and (3.16). These being known, P_i, $i = 1$, 2, 3 are uniquely determined, from which we can in turn compute ω_i for $i = 3$, 4, 5 and so forth. Note that ω_0 is determined from the usual boundary value problem for a clamped Kirchhoff plate. Also, (3.6), (3.13), and (3.16) all have vanishing right-hand sides for $i = 1$, so ω_1 and therefore $\boldsymbol{\phi}_1$ vanish.

The soft clamped plate. In this case, the boundary conditions (3.13) and (3.14) apply, but instead of $\boldsymbol{\phi} \cdot \boldsymbol{s} = 0$ we must enforce $M_{\boldsymbol{s}}\boldsymbol{\phi} = 0$ on $\partial\Omega$. Using (3.3) and the fact that

$$M_{\boldsymbol{s}}\boldsymbol{\Phi} = \frac{D(1-\nu)}{2}\left(-t^{-1}\frac{\partial \boldsymbol{\Phi}}{\partial \hat{\rho}} \cdot \boldsymbol{s} + \frac{\partial \boldsymbol{\Phi}}{\partial \theta} \cdot \boldsymbol{n}\right),$$

we get

(3.18)

$$\begin{aligned}\frac{\partial^2 P_i}{\partial \hat{\rho}^2} &= -\kappa\frac{\partial P_{i-1}}{\partial \hat{\rho}} + \frac{\partial^2 P_{i-2}}{\partial \theta^2} + \frac{2\lambda}{D(1-\nu)} M_{\boldsymbol{s}}\boldsymbol{\phi}_i \\ &= -\kappa\frac{\partial P_{i-1}}{\partial \hat{\rho}} + \frac{\partial^2 P_{i-2}}{\partial \theta^2} + 2\left(\frac{\partial^2}{\partial s \partial n} - \kappa\frac{\partial}{\partial s}\right)(\lambda\omega_i - D\,\Delta\,\omega_{i-2} + \delta_{i4}\lambda^{-1}Dg).\end{aligned}$$

We conclude that ω_0 is uniquely determined and that ω_1 again vanishes. Now since $\partial^2\omega_0/\partial s\partial n = \partial\omega_0/\partial s = 0$, we infer that P_0 and P_1 vanish as well and therefore that ω_3 does also. The other terms can be computed as follows: first ω_2, then P_2 and P_3, then ω_4 and ω_5, then P_4 and P_5, etc. It is interesting to note that ω_0, ω_1, ω_2, and P_0 are the same for the hard and soft clamped plates but ω_3 and P_1 are not (they vanish for the latter but not for the former).

We now show that for the soft clamped plate all the P_i vanish for any values of θ such that $\kappa(\theta) = 0$. Thus there is no boundary layer near a flat portion of the boundary. (This property holds as well for the hard simply supported plate but not for the other boundary conditions we consider.) To prove it in the case of the soft clamped plate, we note first that by (3.7), $\boldsymbol{\phi}_i$ is a gradient for all i. Using this fact, one computes that

$$M_{\boldsymbol{s}}\boldsymbol{\phi}_i = D(1-\nu)\left(\frac{\partial\boldsymbol{\phi}_i \cdot \boldsymbol{n}}{\partial s} - \kappa\boldsymbol{\phi}_i \cdot \boldsymbol{s}\right).$$

In view of (3.14), we have

$$M_{\boldsymbol{s}}\boldsymbol{\phi}_i = -\lambda^{-1}D(1-\nu)\frac{\partial^2 P_{i-2}}{\partial\theta^2}$$

wherever $\kappa = 0$. Our claim then follows from the defining equations for the P_i and induction.

The hard simply supported plate. For the hard simply supported plate, the boundary conditions are (3.13), (3.15), and, arising from the condition $M_{\boldsymbol{n}}\boldsymbol{\phi} = 0$,

$$M_{\boldsymbol{n}}\boldsymbol{\phi}_i = \lambda^{-1}D(1-\nu)\left(\kappa\frac{\partial P_{i-2}}{\partial\theta} + \frac{\partial^2 P_{i-1}}{\partial\theta\partial\hat{\rho}}\right) \quad \text{on } \partial\Omega, \tag{3.19}$$

where we have used (3.3) and the fact that

$$M_{\boldsymbol{n}}\boldsymbol{\Phi} = D\left(-t^{-1}\frac{\partial\boldsymbol{\Phi}}{\partial\hat{\rho}} \cdot \boldsymbol{n} + \nu\frac{\partial\boldsymbol{\Phi}}{\partial\theta} \cdot \boldsymbol{s}\right).$$

Using (3.7) and (3.8), we may rewrite this as

$$
\begin{aligned}
(3.20) \qquad D\left[(1-\nu)\frac{\partial^2}{\partial n^2}+\nu\,\Delta\right]&\left(\omega_i+\lambda^{-1}D\,\Delta\,\omega_{i-2}+\delta_{i4}\lambda^{-2}Dg\right)\\
&=\lambda^{-1}D(1-\nu)\left(\kappa\frac{\partial P_{i-2}}{\partial\theta}+\frac{\partial^2 P_{i-1}}{\partial\theta\partial\hat\rho}\right).
\end{aligned}
$$

Since (3.17) holds, we again have $P_0=0$. Using (3.13) and (3.20), we see that ω_0 is determined from the usual boundary value problem for a simply supported Kirchhoff plate and that ω_1 vanishes. We can then continue by computing P_1 and P_2, then ω_2 and ω_3, etc.

Now since ϕ_i is a gradient, one can verify that

$$
\mathbf{div}\,C\,\mathcal{E}(\phi_i)\cdot\mathbf{s}=\frac{\partial M_{\mathbf{n}}\phi_i}{\partial s}+D(1-\nu)\frac{\partial}{\partial s}\left[\frac{\partial\phi_i\cdot\mathbf{s}}{\partial s}-\kappa\phi_i\cdot\mathbf{n}\right].
$$

If we assume that κ vanishes on a nondegenerate interval, then, combining this equation with (3.19) and (3.15), we can express $\mathbf{div}\,C\,\mathcal{E}(\phi_i)\cdot\mathbf{s}$ in terms of P_{i-2} and P_{i-1} for θ in this interval. Now (3.15) and (3.4) combine to give

$$
\frac{\partial P_i}{\partial\hat\rho}=-\,\mathbf{div}\,C\,\mathcal{E}(\phi_i)\cdot\mathbf{s}.
$$

Thus, on an interval where κ vanishes, $\partial P_i/\partial\hat\rho$ may be expressed in terms of P_{i-1} and P_{i-2} on that interval. A simple induction allows us to conclude that all the P_i vanish for such θ.

The soft simply supported plate. In this case, the boundary conditions are (3.13), (3.18), and (3.19). We can compute ω_0 and ϕ_0 from the same equations as for the hard simply supported case. Then P_0 can be computed (it need not vanish), and then ω_1 (which also need not vanish), P_1, etc.

4. A priori estimates for the soft simply supported plate. We now consider in detail the case of the soft simply supported plate. An easy computation shows that

$$
\hat P_0(\hat\rho,\theta)=D(1-\nu)\widehat{\frac{\partial^2\omega_0}{\partial s\partial n}}(0,\theta)e^{-c\hat\rho},\qquad \text{where } c=\sqrt{\frac{2\lambda}{D(1-\nu)}}=\sqrt{12k}.
$$

We may show in general that $\hat P_i$ are polynomials in $\hat\rho$ times the decaying exponential $e^{-c\hat\rho}$. The specific form is given in the following theorem.

THEOREM 4.1. *For $i\in\mathbb{N}$,*

$$
\hat P_i(\hat\rho,\theta)=e^{-c\hat\rho}\sum_{k=0}^{i}\sum_{j=0}^{i}\sum_{l=0}^{i-j}\alpha_{ijkl}(\theta)\hat\rho^k\frac{\partial^l}{\partial\theta^l}\widehat{M_{\mathbf{s}}\phi_j}(0,\theta),
$$

where the α_{ijkl} are smooth functions of θ which depend only upon the domain Ω.

Proof. Let us say that a function is of type (m,n) if it is a sum of terms of the form

$$
\alpha(\theta)\hat\rho^k\frac{\partial^l}{\partial\theta^l}\widehat{M_{\mathbf{s}}\phi_j}(0,\theta)
$$

with $k,j,l\in\mathbb{N}$ satisfying $k\le m$, $j+l\le n$, and α a smooth function of θ depending only on Ω. We wish to show that P_i is of type (i,i) for $i\in\mathbb{N}$. We shall use induction on i. The result is known for $i=0$. If we assume its validity for $0,1,\dots,i-1$, we easily check that F_i defined in (3.11) is of type $(i-1, i)$ and that the right-hand side of (3.18) is of type $(0, i)$ (there is no ρ dependence since this is on the boundary). It is then easy to see that the unique solution P_i of (3.11), (3.12), and (3.18) must be of type (i, i), as desired. □

Using this formula, we now turn to the derivation of a priori estimates for the interior expansions and boundary correctors. The following estimates are obtained immediately from the form of P_i.

THEOREM 4.2. *For any $i \in \mathbb{N}$ and $s \in \mathbb{R}$ there exists a constant C depending only on Ω, E, ν, k, s, and i, such that*

$$|P_i|_s + \left|\frac{\partial P_i}{\partial \hat{\rho}}\right|_s \le C \sum_{j=0}^{i} |M_s \phi_j|_{s+i-j}.$$

Using this result, we next obtain bounds for the terms in the interior expansions of ϕ and ω.

THEOREM 4.3. *For all real $s \ge 0$ and $i \in \mathbb{N}$, there exists a constant C such that*

$$\|\omega_i\|_{s+2} + \|\phi_i\|_{s+1} \le C \|g\|_{s+i-2}.$$

Proof. Let B_2 denote the boundary differential operator

$$B_2 \omega = D[(1-\nu)\frac{\partial^2 \omega}{\partial n^2} + \nu \, \Delta \, \omega].$$

It easily follows from (3.8), (3.6), (3.13), and (3.20) that

$$D \, \Delta^2 z_i = \delta_{i0} g \quad \text{in } \Omega, \qquad z_i = \lambda^{-1} D \, \Delta \, \omega_{i-2} + \delta_{i4} \lambda^{-2} D g \quad \text{on } \partial\Omega,$$
$$B_2 z_i = \lambda^{-1} D(1-\nu) \left(\kappa \frac{\partial P_{i-2}}{\partial \theta} + \frac{\partial^2 P_{i-1}}{\partial \theta \partial \hat{\rho}} \right) \quad \text{on } \partial\Omega.$$

Applying standard estimates for the biharmonic, we obtain for $s \ge 0$ that

$$\begin{aligned} &\|z_i\|_{s+2} \\ &\le C \left(\delta_{i0} \|g\|_{s-2} + |\, \Delta \, \omega_{i-2}|_{s+3/2} + \delta_{i4} |g|_{s+3/2} + |\frac{\partial P_{i-2}}{\partial \theta}|_{s-1/2} + \left| \frac{\partial^2 P_{i-1}}{\partial \theta \partial \hat{\rho}} \right|_{s-1/2} \right) \\ &\le C \left(\delta_{i0} \|g\|_{s-2} + \|\omega_{i-2}\|_{s+4} + \delta_{i4} \|g\|_{s+2} + |P_{i-2}|_{s+1/2} + \left| \frac{\partial P_{i-1}}{\partial \hat{\rho}} \right|_{s+1/2} \right) \\ &\le C \left(\delta_{i0} \|g\|_{s-2} + \|\omega_{i-2}\|_{s+4} + \delta_{i4} \|g\|_{s+2} + \sum_{j=0}^{i-1} |M_s \phi_j|_{s+i-j-1/2} \right) \\ &\le C \left(\delta_{i0} \|g\|_{s-2} + \|\omega_{i-2}\|_{s+4} + \delta_{i4} \|g\|_{s+2} + \sum_{j=0}^{i-1} \|\phi_j\|_{s+i-j+1} \right). \end{aligned}$$

Now $\phi_i = \mathbf{grad}\, z_i$ and by the definition of z_i and the triangle inequality, it easily follows that

$$\|\omega_i\|_{s+2} \le C \left(\|z_i\|_{s+2} + \|\omega_{i-2}\|_{s+4} + \delta_{i4} \|g\|_{s+2} \right).$$

Combining these results, we obtain

$$\|\omega_i\|_{s+2} + \|\phi_i\|_{s+1} \le C\left(\delta_{i0}\|g\|_{s-2} + \|\omega_{i-2}\|_{s+4} + \delta_{i4}\|g\|_{s+2} + \sum_{j=0}^{i-1}\|\phi_j\|_{s+i-j+1}\right).$$

The result for $i = 0$ follows directly and the result for $i \ge 1$ is then obtained by induction. □

COROLLARY 4.4. *For real $s \ge -1/2$ and $i \in \mathbb{N}$, there exists a constant C such that*

$$|P_i|_s + \left|\frac{\partial P_i}{\partial \hat{\rho}}\right|_s + |M_s\phi_i|_s \le C\|g\|_{s+i-3/2}.$$

Proof. Since

$$M_s\phi_j = M_s(\mathbf{grad}\, z_j) = D(1-\nu)\left(\frac{\partial^2}{\partial s\partial n} - \kappa\frac{\partial}{\partial s}\right)z_j$$

on $\partial\Omega$, we have

$$\begin{aligned}|M_s\phi_j|_{s+i-j} &\le C\left(\left|\frac{\partial z_j}{\partial n}\right|_{s+i-j+1} + |z_j|_{s+i-j+1}\right) \le C\|z_j\|_{s+i-j+5/2}\\ &\le C\left(\|\omega_j\|_{s+i-j+5/2} + \|\omega_{j-2}\|_{s+i-j+9/2} + \delta_{j4}\|g\|_{s+i-j+5/2}\right)\\ &\le C\|g\|_{s+i-3/2}, \quad j = 0, 1, \ldots, i.\end{aligned}$$

The result follows from this estimate and Theorem 4.2.

We next consider the derivation of interior norm estimates for the boundary correctors. To get these results, we make use of the following elementary lemma.

LEMMA 4.5. *Suppose $a > 0$, $b \ge 1$, and $p(x)$ is a polynomial of degree $\le n$ with positive coefficients. Then there exists a constant $K_n(a)$ depending only on n and a such that*

$$\int_b^\infty p(x)e^{-ax}\,dx \le K_n(a)e^{-ab}p(b).$$

Proof. It clearly suffices to prove the result for $p(x) = x^n$. In this case, it reduces to showing that

$$\int_0^\infty (1 + x/b)^n e^{-ax}\,dx \le K_n(a) \quad \text{for all } b \ge 1,$$

which is obvious. □

Next recall that

$$\Omega_0 = \{\, \boldsymbol{z} - \rho\boldsymbol{n}_{\boldsymbol{z}} \mid \boldsymbol{z} \in \partial\Omega, \quad 0 < \rho < \rho_0 \,\}$$

and set

$$\Omega_1 = \{\, \boldsymbol{z} - \rho\boldsymbol{n}_{\boldsymbol{z}} \mid \boldsymbol{z} \in \partial\Omega, \quad \rho_0/3 < \rho < \rho_0 \,\},$$

so $\chi \equiv 1$ on $\Omega_0 \setminus \Omega_1$ and $\chi \equiv 0$ on a neighborhood of $\Omega \setminus \Omega_0$. The following result is similar to results previously derived in [1] (cf. Theorems 4.1 and 4.5).

LEMMA 4.6. *Suppose $k, l, n, s \in \mathbb{N}$,*

$$\hat{P}(\hat{\rho}, \theta) = \hat{\alpha}(\theta)\exp(-c\hat{\rho})p(\hat{\rho}),$$

and

$$\hat{f}(\hat{\rho},\theta) = \hat{\rho}^k \frac{\partial^{l+n}}{\partial\hat{\rho}^l \partial\theta^n} \hat{P}(\hat{\rho},\theta),$$

where α is a smooth function depending on $\partial\Omega$ and p is a polynomial. Then there exists a constant C depending only on Ω, p, k, l, n, and s such that

$$\|f\|_{s,\Omega_0} \le Ct^{1/2-s} \sum_{m=0}^{s} t^m |\alpha|_{m+n}.$$

Moreover, for any $j \ge 0$, there exists a constant C' depending on C and j such that

$$\|f\|_{s,\Omega_1} \le C't^{1/2+j-s} \sum_{m=0}^{s} t^m |\alpha|_{m+n}.$$

We now obtain bounds on the P_i in Ω.

THEOREM 4.7. *For any $i, j, k, l, n, s \in \mathbb{N}$, there is a constant C such that*

$$\|\hat{\rho}^k \frac{\partial^{l+n}}{\partial\hat{\rho}^l \partial\theta^n} P_i\|_{s,\Omega_0} \le C(t^{1/2-s}\|g\|_{n+i-3/2} + t^{1/2}\|g\|_{n+s+i-3/2}), \tag{4.1}$$

$$\|\hat{\rho}^k \frac{\partial^{l+n}}{\partial\hat{\rho}^l \partial\theta^n} P_i\|_{s,\Omega_1} \le Ct^j(t^{1/2-s}\|g\|_{n+i-3/2} + t^{1/2}\|g\|_{n+s+i-3/2}). \tag{4.2}$$

Proof. From Lemma 4.6 and Theorem 4.1, we get

$$\|\hat{\rho}^k \frac{\partial^{l+n}}{\partial\hat{\rho}^l \partial\theta^n} P_i\|_{s,\Omega_0} \le Ct^{1/2-s} \sum_{m=0}^{s} t^m \sum_{j=0}^{i} |M_s\phi_j|_{n+m+i-j}.$$

By Corollary 4.4, this is bounded by

$$Ct^{1/2-s} \sum_{m=0}^{s} t^m \|g\|_{n+m+i-3/2} \le C(t^{1/2-s}\|g\|_{n+i-3/2} + t^{1/2}\|g\|_{n+s+i-3/2}).$$

Similarly,

$$\begin{aligned} \|\hat{\rho}^k \frac{\partial^{l+n}}{\partial\hat{\rho}^l \partial\theta^n} P_i\|_{s,\Omega_1} &\le C_j t^{1/2+j-s} \sum_{m=0}^{s} t^m \sum_{j=0}^{i} |M_s\phi_j|_{n+m+i-j} \\ &\le C_j t^{1/2+j-s} \sum_{m=0}^{s} t^m \|g\|_{n+m+i-3/2} \\ &\le C_j t^j (t^{1/2-s}\|g\|_{n+i-3/2} + t^{1/2}\|g\|_{n+s+i-3/2}). \qquad \square \end{aligned}$$

Using (3.3), we easily obtain the following, where $\phi_n^B = \sum_{i=0}^n t^i \Phi_i$ and $p_n^B = \sum_{i=0}^n t^i P_i$.

COROLLARY 4.8. *For any $s, n, j \in \mathbb{N}$, there is a constant C such that*

$$\|p_n^B\|_{s,\Omega_0} \le C(t^{1/2-s}\|g\|_{-3/2} + t^{n+1/2}\|g\|_{n+s-3/2}), \tag{4.3}$$

$$\|\boldsymbol{\Phi}_n\|_{s,\Omega_0} \le C(t^{1/2-s}\|g\|_{n-5/2} + t^{1/2}\|g\|_{n+s-5/2}), \tag{4.4}$$

$$\|\boldsymbol{\phi}_n^B\|_{s,\Omega_0} \le C(t^{3/2-s}\|g\|_{-3/2} + t^{n+1/2}\|g\|_{n+s-5/2}), \tag{4.5}$$

$$\|p_n^B\|_{s,\Omega_1} \le Ct^j(t^{1/2-s}\|g\|_{-3/2} + t^{n+1/2}\|g\|_{n+s-3/2}), \tag{4.6}$$

$$\|\boldsymbol{\Phi}_n\|_{s,\Omega_1} \le Ct^j(t^{1/2-s}\|g\|_{n-5/2} + t^{1/2}\|g\|_{n+s-5/2}), \tag{4.7}$$

$$\|\boldsymbol{\phi}_n^B\|_{s,\Omega_1} \le Ct^j(t^{3/2-s}\|g\|_{-3/2} + t^{n+1/2}\|g\|_{n+s-5/2}). \tag{4.8}$$

5. Error estimates for the soft simply supported plate. In this section, we shall derive estimates for differences between the solution components of the Reissner–Mindlin equations and finite sums of the asymptotic expansions. We shall not bound these differences directly but rather first bound their images under a differential operator and then apply a priori estimates for the operator. The differential operator we employ is not the Reissner–Mindlin operator but rather a singularly perturbed Stokes-like operator which arises in an equivalent formulation of the Reissner–Mindlin equations due to Brezzi and Fortin [3].

The Brezzi–Fortin formulation begins with the Helmholtz decomposition of the transverse shear stress vector

$$\lambda t^{-2}(\mathbf{grad}\,\omega - \boldsymbol{\phi}) = \mathbf{grad}\,r + \mathbf{curl}\,p, \quad r \in \mathring{H}^1,\ p \in H^1/\mathbb{R}. \tag{5.1}$$

Then it is easy to see that r may be determined by the Poisson equation

$$-\Delta r = g \tag{5.2}$$

together with the homogeneous Dirichlet boundary condition, and then $\boldsymbol{\phi}$ and p may be determined from the perturbed Stokes-like system

$$-\mathbf{div}\,\mathcal{C}\,\mathcal{E}(\boldsymbol{\phi}) - \mathbf{curl}\,p = \mathbf{grad}\,r, \tag{5.3}$$

$$-\operatorname{rot}\boldsymbol{\phi} + \lambda^{-1}t^2\,\Delta p = 0 \tag{5.4}$$

together with the boundary conditions

$$M_{\boldsymbol{n}}\boldsymbol{\phi} = 0, \qquad M_{\boldsymbol{s}}\boldsymbol{\phi} = 0, \qquad \boldsymbol{\phi}\cdot\boldsymbol{s} + \lambda^{-1}t^2\frac{\partial p}{\partial n} = 0. \tag{5.5}$$

Note that p is only determined modulo $\mathbb{R}$, i.e., up to an additive constant. Finally, ω satisfies

$$-\Delta\omega = -\operatorname{div}\boldsymbol{\phi} - \lambda^{-1}t^2\,\Delta r \tag{5.6}$$

and vanishes on the boundary.

The weak formulation of (5.3), (5.4), and the boundary conditions (5.5) seeks $\boldsymbol{\phi} \in \boldsymbol{H}^1$, $p \in H^1/\mathbb{R}$ such that

$$a(\boldsymbol{\phi}, \boldsymbol{\psi}) - (\mathbf{curl}\,p, \boldsymbol{\psi}) = (\mathbf{grad}\,r, \boldsymbol{\psi}) \quad \text{for all } \boldsymbol{\psi} \in \boldsymbol{H}^1, \tag{5.7}$$

$$-(\boldsymbol{\phi} + \lambda^{-1}t^2\,\mathbf{curl}\,p, \mathbf{curl}\,q) = 0 \quad \text{for all } q \in H^1/\mathbb{R}, \tag{5.8}$$

where

$$a(\boldsymbol{\phi}, \boldsymbol{\psi}) = (\mathcal{C}\,\mathcal{E}(\boldsymbol{\phi}), \mathcal{E}(\boldsymbol{\psi})).$$

To continue, we need to define an asymptotic approximation to p. From (3.5), we see that $\lambda(\phi_{i+2} - \mathbf{grad}\,\omega_{i+2}) + \delta_{i0}\,\mathbf{grad}\,r$ is divergence free. Therefore, we can determine a function p_i, unique modulo $\mathbb{R}$, by

$$\mathbf{curl}\,p_i = -\lambda(\phi_{i+2} - \mathbf{grad}\,\omega_{i+2}) - \delta_{i0}\,\mathbf{grad}\,r = -\,\mathbf{div}\,C\,\mathcal{E}(\phi_i) - \delta_{i0}\,\mathbf{grad}\,r. \tag{5.9}$$

It follows immediately from Theorem 4.3 and regularity for the Dirichlet problem that

$$\|p_n\|_{s/\mathbb{R}} \le C\|g\|_{s+n-2}, \quad s \in \mathbb{R},\ s \ge 0. \tag{5.10}$$

Note that, by (3.7), $\mathbf{curl}\,p_i$ is a gradient, so p_i is harmonic for all i. We may now write our asymptotic expansion of p:

$$p \sim \sum_{i=0}^{\infty} t^i p_i + \chi \sum_{i=0}^{\infty} t^i P_i.$$

Let us now introduce some notation for the finite interior and boundary expansion sums. Set

$$\omega_n^I = \sum_{i=0}^{n} t^i w_i, \quad \phi_n^I = \sum_{i=0}^{n} t^i \phi_i, \quad p_n^I = \sum_{i=0}^{n} t^i p_i, \quad \phi_n^B = \sum_{i=0}^{n} t^i \boldsymbol{\Phi}_i, \quad p_n^B = \sum_{i=0}^{n} t^i P_i,$$

and

$$\omega_n^E = \omega - \omega_n^I, \qquad \phi_n^E = \phi - \phi_n^I - \chi\phi_n^B, \qquad p_n^E = p - p_n^I - \chi p_{n-1}^B.$$

Note that we deliberately choose one less term in the boundary-layer expansion for p than for the other terms.

The following three theorems give estimates in Sobolev norms of general index for the differences between ϕ, p, and ω and their finite asymptotic approximations. Note that the rates of convergence for ϕ and p decline as the index of the Sobolev norm increases, but this is not true for ω. This reflects the presence of a boundary layer for the first two variables but not the third.

THEOREM 5.1. *For any $n \in \mathbb{N}$, there exists a constant C independent of t such that*

$$\|\phi_n^E\|_1 + \|p_n^E\|_{0/\mathbb{R}} + t\|\,\mathbf{curl}\,p_n^E\|_0 \le C(t^{n+1/2}\|g\|_{n-3/2} + t^{n+3/2}\|g\|_{n-1/2}).$$

THEOREM 5.2. *For any $n, s \in \mathbb{N}$, $s \ge 2$, there exists a constant C independent of t such that*

$$\|\phi_n^E\|_s + t\|p_n^E\|_{s/\mathbb{R}} \le C(t^{n+3/2-s}\|g\|_{n-3/2} + t^{n+1}\|g\|_{n+s-2}).$$

THEOREM 5.3. *For any $n, s \in \mathbb{N}$, $s \ge 2$, there exists a constant C independent of t such that*

$$\|\omega_n^E\|_2 \le C(t^{n+1}\|g\|_{n-1} + t^{n+5/2}\|g\|_{n+1/2}),$$
$$\|\omega_n^E\|_s \le C(t^{n+1}\|g\|_{n+s-3} + t^{n+s}\|g\|_{n+2s-4}), \quad s \ge 3.$$

The proofs depend on a number of estimates and equations which we collect here and prove at the end of the section. These results show that the formal equations (3.2) and (3.9) and the moment boundary conditions are indeed satisfied, at least to high order, by the finite boundary-layer expansions.

LEMMA 5.4. *For any $n, s \in \mathbb{N}$, there exists a constant C for which*

$$\|\chi\phi_n^B + \lambda^{-1}t^2\,\mathbf{curl}(\chi p_{n-1}^B)\|_s \le C(t^{n+3/2-s}\|g\|_{n-3/2} + t^{n+3/2}\|g\|_{n+s-3/2}), \tag{5.11}$$

$$\|\operatorname{div}(\chi\phi_n^B)\|_s \le C(t^{n+1/2-s}\|g\|_{n-3/2} + t^{n+1/2}\|g\|_{n+s-3/2}), \tag{5.12}$$

$$\|D\frac{1-\nu}{2}\operatorname{rot}(\chi\phi_n^B) - p_{n-1}^B\|_s \le C(t^{n+1/2-s}\|g\|_{n-3/2} + t^{n+1/2}\|g\|_{s+n-3/2}), \tag{5.13}$$

$$D\frac{1-\nu}{2}\operatorname{rot}(\chi\phi_n^B) - p_{n-1}^B + t^n\left(\lambda^{-1}\frac{D(1-\nu)}{2}\frac{\partial^2\hat{P}_n}{\partial\hat{\rho}^2} - \hat{P}_n\right) = 0 \quad \text{on } \partial\Omega, \tag{5.14}$$

$$M_{\boldsymbol{n}}(\phi_n^I + \phi_n^B) = D\operatorname{div}\phi_n^B \quad \text{on } \partial\Omega, \tag{5.15}$$

$$M_{\boldsymbol{s}}(\phi_n^I + \phi_n^B) = \lambda^{-1}\frac{D(1-\nu)}{2}t^n\frac{\partial^2\hat{P}_n}{\partial\hat{\rho}^2} \quad \text{on } \partial\Omega, \tag{5.16}$$

$$M_{\boldsymbol{s}}(\phi_n^I + \phi_n^B) + D\frac{1-\nu}{2}\operatorname{rot}(\chi\phi_n^B) - p_{n-1}^B = t^n\hat{P}_n \quad \text{on } \partial\Omega. \tag{5.17}$$

In the interest of brevity, we introduce the following notation for the quantity on the right-hand side of the estimate in Theorem 5.1:

$$\Lambda = t^{n+1/2}\|g\|_{n-3/2} + t^{n+3/2}\|g\|_{n-1/2}.$$

Proof of Theorem 5.1. It follows immediately from (5.9) that

$$-\mathbf{div}\,C\,\mathcal{E}(\phi_n^I) - \mathbf{curl}\,p_n^I = \mathbf{grad}\,r, \quad n = 0, 1, \ldots .$$

Therefore,

$$-a(\phi_n^I, \psi) + (\mathbf{curl}\,p_n^I, \psi) = -(\mathbf{grad}\,r, \psi) - \langle M_{\boldsymbol{n}}\phi_n^I, \psi\cdot\boldsymbol{n}\rangle - \langle M_{\boldsymbol{s}}\phi_n^I, \psi\cdot\boldsymbol{s}\rangle, \quad \psi \in \boldsymbol{H}^1. \tag{5.18}$$

Using the identity

$$\begin{aligned} a(\phi, \psi) = D\frac{1-\nu}{2}(\operatorname{rot}\phi, \operatorname{rot}\psi) + D(\operatorname{div}\phi, \operatorname{div}\psi) \\ + \langle M_{\boldsymbol{n}}\phi - D\operatorname{div}\phi, \psi\cdot\boldsymbol{n}\rangle + \left\langle M_{\boldsymbol{s}} + D\frac{1-\nu}{2}\operatorname{rot}\phi, \psi\cdot\boldsymbol{s}\right\rangle, \end{aligned}$$

we get

$$\begin{aligned} &-a(\chi\phi_n^B, \psi) + (\mathbf{curl}\,\chi p_{n-1}^B, \psi) \\ &= -(D\frac{1-\nu}{2}\operatorname{rot}(\chi\phi_n^B) - \chi p_{n-1}^B, \operatorname{rot}\psi) - D(\operatorname{div}(\chi\phi_n^B), \operatorname{div}\psi) \\ &- \langle M_{\boldsymbol{n}}\phi_n^B - D\operatorname{div}\phi_n^B, \psi\cdot\boldsymbol{n}\rangle - \left\langle M_{\boldsymbol{s}}\phi_n^B + D\frac{1-\nu}{2}\operatorname{rot}\phi_n^B - p_{n-1}^B, \psi\cdot\boldsymbol{s}\right\rangle, \quad \psi \in \boldsymbol{H}^1. \end{aligned} \tag{5.19}$$

Adding (5.7), (5.18), and (5.19) and using (5.15) and (5.17) gives the error equation corresponding to (5.7):

$$\begin{aligned} a(\phi_n^E, \psi) - (\mathbf{curl}\,p_n^E, \psi) = -\left(D\frac{1-\nu}{2}\operatorname{rot}(\chi\phi_n^B) - \chi p_{n-1}^B, \operatorname{rot}\psi\right) \\ - D\left(\operatorname{div}(\chi\phi_n^B), \operatorname{div}\psi\right) - t^n\langle\hat{P}_n, \psi\cdot\boldsymbol{s}\rangle, \quad \psi \in \boldsymbol{H}^1. \end{aligned} \tag{5.20}$$

Turning to the second equation, we have from (5.9) that

$$\phi_n^I = \mathbf{grad}\,\omega_n^I - \lambda^{-1}t^2\,\mathbf{grad}\,r - \lambda^{-1}t^2\,\mathbf{curl}\,p_{n-2}^I.$$

Combining this with (5.1), we obtain

$$\begin{aligned}\phi_n^E = \mathbf{grad}(\omega - \omega_n^I) - \lambda^{-1}t^2\,\mathbf{curl}\,p_n^E&\\ - \lambda^{-1}(t^{n+1}\,\mathbf{curl}\,p_{n-1} + t^{n+2}\,\mathbf{curl}\,p_n) - [\chi\phi_n^B + \lambda^{-1}t^2\,\mathbf{curl}(\chi p_{n-1}^B)]&.\end{aligned}$$

Multiplying by $\mathbf{curl}\,q$ for $q \in H^1$, using the orthogonality of gradients and curls, and rearranging, gives the error equation corresponding to (5.8):

$$\begin{aligned}(\phi_n^E + \lambda^{-1}t^2\,\mathbf{curl}\,p_n^E, \mathbf{curl}\,q) = -\lambda^{-1}(t^{n+1}\,\mathbf{curl}\,p_{n-1} + t^{n+2}\,\mathbf{curl}\,p_n, \mathbf{curl}\,q)&\\ - \left(\chi\phi_n^B + \lambda^{-1}t^2\,\mathbf{curl}(\chi p_{n-1}^B), \mathbf{curl}\,q\right), \quad q \in H^1&. \qquad (5.21)\end{aligned}$$

The desired error estimate will be obtained from (5.20) and (5.21) using a few choices of the test functions ψ and q. For this we will need to bound various terms arising on the right-hand sides.

Our first choice of test functions is $\psi = \phi_n^E$ in (5.20) and $q = p_n^E - t^n\chi P_n$ in (5.21). (The more obvious test function $q = p_n^E$ could also be used here, but not for the case of the free plate, since there we will need q to vanish on the boundary.) Adding these equations and rearranging terms, we get

$$a(\phi_n^E, \phi_n^E) + \lambda^{-1}t^2\|\,\mathbf{curl}\,p_n^E\|_0^2 = T_1 + T_2 + \cdots + T_9,$$

where

$$\begin{aligned}
T_1 &= -\left(D\frac{1-\nu}{2}\,\mathrm{rot}(\chi\phi_n^B) - \chi p_{n-1}^B, \mathrm{rot}\,\phi_n^E\right),\\
T_2 &= -D\left(\mathrm{div}(\chi\phi_n^B), \mathrm{div}\,\phi_n^E\right),\\
T_3 &= -t^n\langle \hat{P}_n, \phi_n^E\cdot s\rangle,\\
T_4 &= -\lambda^{-1}\left(t^{n+1}\,\mathbf{curl}\,p_{n-1} + t^{n+2}\,\mathbf{curl}\,p_n, \mathbf{curl}\,p_n^E\right),\\
T_5 &= \lambda^{-1}t^n\left(t^{n+1}\,\mathbf{curl}\,p_{n-1} + t^{n+2}\,\mathbf{curl}\,p_n, \mathbf{curl}(\chi\hat{P}_n)\right),\\
T_6 &= -\left(\chi\phi_n^B + \lambda^{-1}t^2\,\mathbf{curl}(\chi p_{n-1}^B), \mathbf{curl}\,p_n^E\right),\\
T_7 &= t^n\left(\chi\phi_n^B + \lambda^{-1}t^2\,\mathbf{curl}(\chi p_{n-1}^B), \mathbf{curl}(\chi\hat{P}_n)\right),\\
T_8 &= t^n\left(\phi_n^E, \mathbf{curl}(\chi\hat{P}_n)\right) = t^n(\mathrm{rot}\,\phi_n^E, \chi\hat{P}_n) + t^n\langle \hat{P}_n, \phi_n^E\cdot s\rangle,\\
T_9 &= \lambda^{-1}t^{n+2}\left(\mathbf{curl}\,p_n^E, \mathbf{curl}(\chi\hat{P}_n)\right).
\end{aligned}$$

By (5.13) and (5.12), we get

$$|T_1| + |T_2| \le C\Lambda\|\phi_n^E\|_{1/\mathbb{R}}. \qquad (5.22)$$

Next, from (4.1),

$$|T_3 + T_8| = |t^n(\mathrm{rot}\,\phi_n^E, \chi\hat{P}_n)| \le C\Lambda\|\phi_n^E\|_{1/\mathbb{R}}. \qquad (5.23)$$

Since the p_n are harmonic, using (5.10) we obtain for any q that

$$\begin{aligned}|(t^{n+1}\,\mathbf{curl} p_{n-1}+t^{n+2}\,\mathbf{curl}\,p_n,\mathbf{curl}\,q)| &= \left|\left\langle t^{n+1}\frac{\partial p_{n-1}}{\partial n}+t^{n+2}\frac{\partial p_n}{\partial n},\bar{q}\right\rangle\right| \\ &\le |\bar{q}|_0\left|t^{n+1}\frac{\partial p_{n-1}}{\partial n}+t^{n+2}\frac{\partial p_n}{\partial n}\right|_0 \le |\bar{q}|_0\left(t^{n+1}\|p_{n-1}\|_{3/2}+t^{n+2}\|p_n\|_{3/2}\right) \\ &\le C\|\bar{q}\|_0^{1/2}(t\|\,\mathbf{curl}\,q\|_0)^{1/2}\Lambda \le C\Lambda^2+\delta(\|q\|_{0/\mathbb{R}}^2+t^2\|\,\mathbf{curl}\,q\|_0^2),\end{aligned}$$

where $\bar{q}$ is the difference between q and its mean value and δ can be any positive number and will be chosen later. Applying this twice and using (4.1), we get

$$|T_4| \le C\Lambda^2+\delta(\|p_n^E\|_{0/\mathbb{R}}^2+t^2\|\,\mathbf{curl}\,p_n^E\|_0^2), \qquad |T_5| \le C\Lambda^2. \tag{5.24}$$

Finally, by (5.11),

$$|T_6| \le Ct\Lambda\|\,\mathbf{curl}\,p_n^E\|_0, \qquad |T_7| \le C\Lambda^2, \tag{5.25}$$

and, using (4.1),

$$|T_9| \le Ct\Lambda\|\,\mathbf{curl}\,p_n^E\|_0. \tag{5.26}$$

Combining (5.22)–(5.26) gives

$$a(\phi_n^E,\phi_n^E)+\lambda^{-1}t^2\|\,\mathbf{curl}\,p_n^E\|_0^2 \le C_\epsilon\Lambda^2+\epsilon(\|\phi_n^E\|_{1/\mathbb{R}}^2+\|p_n^E\|_{0/\mathbb{R}}^2+t^2\|\,\mathbf{curl}\,p_n^E\|_0^2), \tag{5.27}$$

where $\epsilon > 0$ is arbitrary and $C_\epsilon > 0$ depends on ϵ.

To get control over the L^2 norm of p_n^E, we use another test function in (5.20). Namely, we select $\psi \in \overset{\circ}{\boldsymbol{H}}{}^1$ with $\operatorname{rot}\psi = \bar{p}_n^E$ and $\|\psi\|_1 \le C\|p_n^E\|_{0/\mathbb{R}}$ (this is always possible). Then

$$\|p_n^E\|_{0/\mathbb{R}}^2 = \|\bar{p}_n^E\|_0^2 = (p_n^E,\operatorname{rot}\psi) = (\mathbf{curl}\,p_n^E,\psi) = a(\phi_n^E,\psi)-[a(\phi_n^E,\psi)-(\mathbf{curl}\,p_n^E,\psi)].$$

Using (5.20) and noting that ψ vanishes on $\partial\Omega$, we may write the term in brackets as

$$\left(D\frac{1-\nu}{2}\operatorname{rot}(\chi\phi_n^B)-\chi p_{n-1}^B,\mathbf{curl}\,\psi\right)+D\left(\operatorname{div}(\chi\phi_n^B),\operatorname{div}\psi\right).$$

Using (5.13), (5.12), and Schwarz's inequality, we easily conclude

$$\|p_n^E\|_{0/\mathbb{R}}^2 \le C\Lambda^2+C_1\|\phi_n^E\|_1^2. \tag{5.28}$$

The above estimates give us control over $a(\phi_n^E,\phi_n^E)$, $\|p_n^E\|_0/\mathbb{R}$, and $t\|\,\mathbf{curl}\,p_n^E\|_0$. The theorem would follow easily were $\psi \mapsto a(\psi,\psi)^{1/2}$ equivalent to the $\boldsymbol{H}^1$ norm. But this is not so, since $a(\psi,\psi)$ vanishes for ψ in the three-dimensional space

$$\boldsymbol{R} := \{\,(a-by,c+bx)\mid a,b,c\in\mathbb{R}\,\}$$

of plane rigid motions. However, $\psi \mapsto a(\psi,\psi)^{1/2}+\|\boldsymbol{P}\psi\|_0$ is equivalent to the $\boldsymbol{H}^1$ norm, with $\boldsymbol{P}$ the $\boldsymbol{L}^2$-projection onto $\boldsymbol{R}$. Therefore, we choose q in (5.21) of mean value zero such that $\mathbf{curl}\,q = \boldsymbol{P}\phi_n^E$, which is possible since the functions in $\boldsymbol{R}$ are divergence free. Then

$$\|\boldsymbol{P}\phi_n^E\|_0^2 = (\phi_n^E,\mathbf{curl}\,q) = (\phi_n^E+\lambda^{-1}t^2\,\mathbf{curl}\,p_n^E,\mathbf{curl}\,q)-\lambda^{-1}t^2(\mathbf{curl}\,p_n^E,\boldsymbol{P}\phi_n^E).$$

Using (5.21), (5.11), (5.10), and Schwarz's inequality, we conclude

$$\|\boldsymbol{P}\phi_n^E\|_0^2 \le C\Lambda^2+C_2t^2\|\,\mathbf{curl}\,p_n^E\|_0^2. \tag{5.29}$$

It is a fairly easy matter to conclude the proof from (5.27)–(5.29). Adding $1/(2C_2)$ times (5.29) to (5.27), we get after simple manipulations

$$\|\phi_n^E\|_1^2 + t^2\|\,\mathbf{curl}\,p_n^E\|_0^2 \le C\Lambda^2 + C_3\epsilon(\|\phi_n^E\|_1^2 + \|p_n^E\|_{0/\mathbb{R}}^2 + t^2\|\,\mathbf{curl}\,p_n^E\|_0^2),$$

for some constant C_3. Then adding $1/(2C_1)$ times (5.28) to this equation and similarly manipulating, we obtain

$$\|\phi_n^E\|_1^2 + \|p_n^E\|_{0/\mathbb{R}}^2 + t^2\|\,\mathbf{curl}\,p_n^E\|_0^2 \le C\Lambda^2 + C_4\epsilon(\|\phi_n^E\|_1^2 + \|p_n^E\|_{0/\mathbb{R}}^2 + t^2\|\,\mathbf{curl}\,p_n^E\|_0^2).$$

Finally, choosing ϵ sufficiently small we obtain the theorem. □

Proof of Theorem 5.2. By standard regularity results for plane elasticity,

$$\|\phi_n^E\|_s \le C\left(\|\,\mathbf{div}\,\mathcal{C}\,\mathcal{E}(\phi_n^E)\|_{s-2} + |M_{\boldsymbol{n}}\phi_n^E|_{s-3/2} + |M_{\boldsymbol{s}}\phi_n^E|_{s-3/2} + \|\boldsymbol{P}\phi_n^E\|_0\right).$$

From (5.20),

$$-\,\mathbf{div}\,\mathcal{C}\,\mathcal{E}(\phi_n^E) = \mathbf{curl}\,p_n^E - \mathbf{curl}\left[D\frac{1-\nu}{2}\operatorname{rot}(\chi\phi_n^B) - \chi p_{n-1}^B\right] + D\,\mathbf{grad}\operatorname{div}(\chi\phi_n^B).$$

Then applying (5.13), (5.12), (5.15), (5.16), (4.1), Theorem 5.1, and the trace theorem, we get

$$\|\phi_n^E\|_s \le C\left(\|p_n^E\|_{s-1/\mathbb{R}} + t^{n+3/2-s}\|g\|_{n-3/2} + t^{n+1/2}\|g\|_{s+n-5/2}\right).$$

Next, using regularity for the Neumann problem for the Laplacian, we know that

$$\|p_n^E\|_{s/\mathbb{R}} \le C\left(\|\,\Delta p_n^E\|_{s-2} + \left|\frac{\partial p_n^E}{\partial n}\right|_{s-3/2}\right).$$

By (5.21),

$$\Delta p_n^E = \lambda t^{-2}\left\{\operatorname{rot}\phi_n^E + \operatorname{rot}[\chi\phi_n^B + \lambda^{-1}t^2\,\mathbf{curl}(\chi p_{n-1}^B)]\right\}$$

and

$$\frac{\partial p_n^E}{\partial n} =$$

$$-\lambda t^{-2}\left\{\phi_n^E\cdot\boldsymbol{s} + \lambda^{-1}\left(t^{n+1}\frac{\partial p_{n-1}}{\partial n} + t^{n+2}\frac{\partial p_n}{\partial n}\right) + [\chi\phi_n^B + \lambda^{-1}t^2\,\mathbf{curl}(\chi p_{n-1}^B)]\cdot\boldsymbol{s}\right\}.$$

Applying (5.11), (5.10), the trace theorem, and (2.1), we obtain

$$\|p_n^E\|_{s/\mathbb{R}} \le C(t^{-2}\|\phi_n^E\|_{s-1} + t^{n+1/2-s}\|g\|_{n-3/2} + t^n\|g\|_{s+n-2}).$$

Combining these bounds, we have

$$\|\phi_n^E\|_s + t\|p_n^E\|_{s/\mathbb{R}} \le C(t^{-1}\|\phi_n^E\|_{s-1} + \|p_n^E\|_{s-1/\mathbb{R}} + t^{n+1/2-s}\|g\|_{n-3/2} + t^n\|g\|_{s+n-2}).$$

For $s = 2$, the theorem follows from this relation and Theorem 5.1, and for $s > 2$, it follows by induction on s. □

Proof of Theorem 5.3. From (5.6) and (3.5), we get

$$\Delta(\omega - \omega_n^I) = \operatorname{div}(\phi - \phi_n^I) = \operatorname{div}\left(\phi_{n+s-1}^E + \chi\phi_{n+s-1}^B + \sum_{j=n+1}^{n+s-1} t^j\phi_j\right).$$

The theorem follows by elliptic regularity for Poisson's problem, Lemma 4.3, (5.12), and Theorems 5.1 and 5.2. □

We conclude this section with the proof of Lemma 5.4.

Proof of Lemma 5.4. From (3.3) and (2.2), we can express ϕ_n^B in terms of the P_i:

$$\begin{aligned}\lambda\phi_n^B &= \sum_{i=1}^{n} t^i \frac{\partial \hat{P}_{i-1}}{\partial \hat{\rho}}\boldsymbol{s} + \sum_{i=2}^{n}\sum_{j=0}^{n-i} t^i (\kappa\hat{\rho}t)^j \frac{\partial \hat{P}_{i-2}}{\partial \theta}\boldsymbol{n} \\ &= \sum_{i=1}^{n} t^i \frac{\partial \hat{P}_{i-1}}{\partial \hat{\rho}}\boldsymbol{s} + \sum_{i=2}^{n} t^i \sigma[1-(\kappa\hat{\rho}t)^{n-i+1}] \frac{\partial \hat{P}_{i-2}}{\partial \theta}\boldsymbol{n}.\end{aligned} \tag{5.30}$$

Applying the identity

$$M_{\boldsymbol{n}}\boldsymbol{\psi} - D\operatorname{div}\boldsymbol{\psi} = -D(1-\nu)\left[\frac{\partial(\hat{\boldsymbol{\psi}}\cdot\boldsymbol{s})}{\partial\theta} + \kappa(\hat{\boldsymbol{\psi}}\cdot\boldsymbol{n})\right] \quad \text{on } \partial\Omega$$

and (3.19), we get that

$$M_{\boldsymbol{n}}\phi_n^B - D\operatorname{div}\phi_n^B = -\lambda^{-1}D(1-\nu)\sum_{i=0}^{n} t^i\left[\frac{\partial^2 \hat{P}_{i-1}}{\partial\theta\partial\hat{\rho}} + \kappa\frac{\partial \hat{P}_{i-2}}{\partial\theta}\right] = -M_{\boldsymbol{n}}\phi_n^I,$$

which proves (5.15).

Applying the identity

$$M_{\boldsymbol{s}}\boldsymbol{\psi} = \frac{D(1-\nu)}{2}\left[-t^{-1}\frac{\partial(\hat{\boldsymbol{\psi}}\cdot\boldsymbol{s})}{\partial\hat{\rho}} + \frac{\partial(\hat{\boldsymbol{\psi}}\cdot\boldsymbol{n})}{\partial\theta} - \kappa(\hat{\boldsymbol{\psi}}\cdot\boldsymbol{s})\right]$$

to (5.30) and using (3.18), we get

$$\begin{aligned}\lambda M_{\boldsymbol{s}}\phi_n^B &= \frac{D(1-\nu)}{2}\sum_{i=0}^{n} t^i\left[-t^{-1}\frac{\partial^2\hat{P}_{i-1}}{\partial\hat{\rho}^2} + \frac{\partial^2\hat{P}_{i-2}}{\partial\theta^2} - \kappa\frac{\partial\hat{P}_{i-1}}{\partial\hat{\rho}}\right] \\ &= \frac{D(1-\nu)}{2}\sum_{i=0}^{n} t^i\left(-t^{-1}\frac{\partial^2\hat{P}_{i-1}}{\partial\hat{\rho}^2} + \frac{\partial^2\hat{P}_{i}}{\partial\hat{\rho}^2}\right) - \lambda M_{\boldsymbol{s}}\phi_n^I \\ &= \frac{D(1-\nu)}{2}t^n\frac{\partial^2\hat{P}_n}{\partial\hat{\rho}^2} - \lambda M_{\boldsymbol{s}}\phi_n^I,\end{aligned}$$

which proves (5.16).

Using (5.30) and the expansion

$$t^2\,\mathbf{curl}\,p_{n-1}^B = -\sum_{i=0}^{n-1} t^{i+2}\left(t^{-1}\frac{\partial\hat{P}_i}{\partial\hat{\rho}}\boldsymbol{s} + \sigma\frac{\partial\hat{P}_i}{\partial\theta}\boldsymbol{n}\right) = -\sum_{i=1}^{n} t^i\frac{\partial\hat{P}_{i-1}}{\partial\hat{\rho}}\boldsymbol{s} - \sum_{i=2}^{n+1} t^i\sigma\frac{\partial\hat{P}_{i-2}}{\partial\theta}\boldsymbol{n},$$

we get

$$\lambda\phi_n^B + t^2\,\mathbf{curl}\,p_{n-1}^B = -t^{n+1}\sigma\sum_{i=0}^{n-1}(\kappa\hat{\rho})^{n-i-1}\frac{\partial\hat{P}_i}{\partial\theta}\boldsymbol{n}. \tag{5.31}$$

It now follows directly from Theorem 4.7 that

$$\|\lambda\phi_n^B + t^2\,\mathbf{curl}\,p_{n-1}^B\|_{s,\Omega_0} \le C(t^{n+3/2-s}\|g\|_{n-3/2} + t^{n+3/2}\|g\|_{n+s-3/2}).$$

Finally, using (4.6), we get

$$\begin{aligned}\|\lambda\chi\boldsymbol{\phi}_n^B + t^2\,\mathbf{curl}(\chi p_{n-1}^B)\|_s &\le \|\chi(\lambda\boldsymbol{\phi}_n^B + t^2\,\mathbf{curl}\,p_{n-1}^B)\|_s + t^2\|p_{n-1}^B\cdot\mathbf{curl}\,\chi\|_s\\ &\le C(\|\lambda\boldsymbol{\phi}_n^B + t^2\,\mathbf{curl}\,p_{n-1}^B\|_{s,\Omega_0} + t^2\|p_{n-1}^B\|_{s,\Omega_1})\\ &\le C(t^{n+3/2-s}\|g\|_{n-3/2} + t^{n+3/2}\|g\|_{n+s-3/2}),\end{aligned}$$

which proves (5.11).

Now for any $\boldsymbol{\psi}$,

$$\operatorname{div}\boldsymbol{\psi} = -\frac{\partial\boldsymbol{\psi}}{\partial\rho}\cdot\boldsymbol{n} + \sigma\frac{\partial\boldsymbol{\psi}}{\partial\theta}\cdot\boldsymbol{s} = \left(-t^{-1}\frac{\partial}{\partial\hat\rho} + \sigma\kappa\right)(\hat{\boldsymbol{\psi}}\cdot\boldsymbol{n}) + \sigma\frac{\partial(\hat{\boldsymbol{\psi}}\cdot\boldsymbol{s})}{\partial\theta}.$$

From (5.31), we then have

$$\operatorname{div}\boldsymbol{\phi}_n^B = \left(-t^{-1}\frac{\partial}{\partial\hat\rho} + \sigma\kappa\right)\left[-t^{n+1}\sigma\sum_{i=0}^{n-1}(\kappa\hat\rho)^{n-i-1}\frac{\partial\hat P_i}{\partial\theta}\right].$$

It then follows easily from Theorem 4.7 that

$$\|\operatorname{div}(\chi\boldsymbol{\phi}_n^B)\|_{s,\Omega_0} \le C(t^{n+1/2-s}\|g\|_{n-3/2} + t^{n+1/2}\|g\|_{n+s-3/2}).$$

To complete the proof of (5.12), we use (4.8).

Finally, we give the proof of (5.13) and (5.14). For $i = 0, 1, \dots, n$, we get by simple identities that

$$\operatorname{rot}\boldsymbol{\Phi}_i = \frac{\partial\boldsymbol{\Phi}_i}{\partial\rho}\cdot\boldsymbol{s} + \sigma\frac{\partial\boldsymbol{\Phi}_i}{\partial\theta}\cdot\boldsymbol{n} = \frac{\partial\boldsymbol{\Phi}_i}{\partial\rho}\cdot\boldsymbol{s} + \left\{\sum_{j=0}^{n-i}(\kappa\rho)^j + \sigma(\kappa\rho)^{n-i+1}\right\}\frac{\partial\boldsymbol{\Phi}_i}{\partial\theta}\cdot\boldsymbol{n}.$$

Hence,

$$\begin{aligned}\operatorname{rot}\boldsymbol{\phi}_n^B &= \sum_{i=0}^{n} t^i\operatorname{rot}\hat{\boldsymbol{\Phi}}_i = \sum_{i=0}^{n} t^i\left\{t^{-1}\frac{\partial\hat{\boldsymbol{\Phi}}_i}{\partial\hat\rho}\cdot\boldsymbol{s} + \left[\sum_{j=0}^{n-i}(\kappa\rho)^j + \sigma(\kappa\rho)^{n-i+1}\right]\frac{\partial\hat{\boldsymbol{\Phi}}_i}{\partial\theta}\cdot\boldsymbol{n}\right\}\\ &= \lambda^{-1}\sum_{i=0}^{n-1} t^i\frac{\partial^2\hat P_i}{\partial\hat\rho^2} + \sum_{i=0}^{n} t^i\left[\sum_{j=0}^{n-i}(\kappa\rho)^j + \sigma(\kappa\rho)^{n-i+1}\right]\frac{\partial\hat{\boldsymbol{\Phi}}_i}{\partial\theta}\cdot\boldsymbol{n},\end{aligned}$$

where we used (3.3) and reindexed the first sum in the last step. Turning to the double sum on the left-hand side, we use the identity (2.2) to obtain

$$\begin{aligned}&\sum_{i=0}^{n} t^i\left[\sum_{j=0}^{n-i}(\kappa\hat\rho t)^j + \sigma(\kappa\hat\rho t)^{n-i+1}\right]\frac{\partial\hat{\boldsymbol{\Phi}}_i}{\partial\theta}\cdot\boldsymbol{n}\\ &\quad= \sum_{i=0}^{n}\sum_{j=0}^{i} t^{i-j}(\kappa\hat\rho t)^j\frac{\partial\hat{\boldsymbol{\Phi}}_{i-j}}{\partial\theta}\cdot\boldsymbol{n} + t^{n+1}\sum_{i=0}^{n}\sigma(\kappa\hat\rho)^{n-i+1}\frac{\partial\hat{\boldsymbol{\Phi}}_i}{\partial\theta}\cdot\boldsymbol{n}\\ &\quad= \sum_{i=0}^{n} t^i\sum_{j=0}^{i}(\kappa\hat\rho)^j\frac{\partial\hat{\boldsymbol{\Phi}}_{i-j}}{\partial\theta}\cdot\boldsymbol{n} + t^{n+1}\sum_{i=0}^{n}\sigma(\kappa\hat\rho)^{n-i+1}\frac{\partial\hat{\boldsymbol{\Phi}}_i}{\partial\theta}\cdot\boldsymbol{n}.\end{aligned}$$

Using (3.3) and (2.2), we further obtain that

$$
\begin{aligned}
\sum_{j=0}^{i}(\kappa\hat\rho)^j \frac{\partial\hat{\boldsymbol{\Phi}}_{i-j}}{\partial\theta}\cdot\boldsymbol{n} &= \sum_{j=0}^{i}(\kappa\hat\rho)^j\left[\frac{\partial(\hat{\boldsymbol{\Phi}}_{i-j}\cdot\boldsymbol{n})}{\partial\theta} - \kappa\hat{\boldsymbol{\Phi}}_{i-j}\cdot\boldsymbol{s}\right]\\
&= \lambda^{-1}\sum_{j=0}^{i}(\kappa\hat\rho)^j\left\{\frac{\partial}{\partial\theta}\left[\sum_{l=0}^{i-j}(\kappa\hat\rho)^l\frac{\partial\hat P_{i-j-l-2}}{\partial\theta}\right] - \kappa\frac{\partial\hat P_{i-j-1}}{\partial\hat\rho}\right\}\\
&= \lambda^{-1}\sum_{j=0}^{i}\sum_{l=0}^{j}(\kappa\hat\rho)^{j-l}\frac{\partial}{\partial\theta}\left[(\kappa\hat\rho)^l\frac{\partial\hat P_{i-j-2}}{\partial\theta}\right] - \lambda^{-1}\sum_{j=0}^{i}\kappa(\kappa\hat\rho)^j\frac{\partial\hat P_{i-j-1}}{\partial\hat\rho}\\
&= \lambda^{-1}\sum_{j=0}^{i}\sum_{l=0}^{j}\left[(\kappa\hat\rho)^j\frac{\partial^2\hat P_{i-j-2}}{\partial\theta^2} + l\kappa^{j-1}\kappa'\hat\rho^j\frac{\partial\hat P_{i-j-2}}{\partial\theta}\right] - \lambda^{-1}\sum_{j=0}^{i}\kappa(\kappa\hat\rho)^j\frac{\partial\hat P_{i-j-1}}{\partial\hat\rho}\\
&= \lambda^{-1}\sum_{j=0}^{i}\left[(j+1)(\kappa\hat\rho)^j\frac{\partial^2\hat P_{i-j-2}}{\partial\theta^2} + \frac{j(j+1)}{2}\kappa^{j-1}\kappa'\hat\rho^j\frac{\partial\hat P_{i-j-2}}{\partial\theta} - \kappa(\kappa\hat\rho)^j\frac{\partial\hat P_{i-j-1}}{\partial\hat\rho}\right]\\
&= \lambda^{-1}\sum_{j=0}^{i}\hat\rho^j\left[a_2^j\frac{\partial^2\hat P_{i-j-2}}{\partial\theta^2} + a_3^j\frac{\partial\hat P_{i-j-2}}{\partial\theta} + a_1^j\frac{\partial\hat P_{i-j-1}}{\partial\hat\rho}\right]\\
&= -\lambda^{-1}\frac{\partial^2\hat P_i}{\partial\hat\rho^2} + \frac{2}{D(1-\nu)}\hat P_i,
\end{aligned}
$$

where the a_i^j are defined in (3.10) and we used (3.11) in the last step. Collecting these results, we have

$$
\begin{aligned}
&\operatorname{rot}\phi_n^B\\
&= \lambda^{-1}\sum_{i=0}^{n-1}t^i\frac{\partial^2\hat P_i}{\partial\hat\rho^2} - \sum_{i=0}^{n}t^i\left[\lambda^{-1}\frac{\partial^2\hat P_i}{\partial\hat\rho^2} - \frac{2}{D(1-\nu)}\hat P_i\right] + t^{n+1}\sum_{i=0}^{n}\sigma(\kappa\hat\rho)^{n-i+1}\frac{\partial\hat{\boldsymbol{\Phi}}_i}{\partial\theta}\cdot\boldsymbol{n}\\
&= -\lambda^{-1}t^n\frac{\partial^2\hat P_n}{\partial\hat\rho^2} + \frac{2}{D(1-\nu)}p_n^B + t^{n+1}\sum_{i=0}^{n}\sigma(\kappa\hat\rho)^{n-i+1}\frac{\partial\hat{\boldsymbol{\Phi}}_i}{\partial\theta}\cdot\boldsymbol{n},
\end{aligned}
$$

and so

$$
D\frac{1-\nu}{2}\operatorname{rot}\phi_n^B - p_{n-1}^B = t^n\left[-D\frac{1-\nu}{2}\lambda^{-1}\frac{\partial^2\hat P_n}{\partial\hat\rho^2} + \hat P_n\right] + t^{n+1}\sum_{i=0}^{n}\sigma(\kappa\hat\rho)^{n-i+1}\frac{\partial\hat{\boldsymbol{\Phi}}_i}{\partial\theta}.
$$

Equation (5.14) follows directly. Using Theorem 4.7 and (4.4), we then obtain

$$
\|D\frac{1-\nu}{2}\operatorname{rot}\phi_n^B - p_{n-1}^B\|_{s,\Omega_0} \le C(t^{n+1/2-s}\|g\|_{n-3/2} + t^{n+1/2}\|g\|_{s+n-3/2}).
$$

Finally, using (4.8), we obtain

$$
\begin{aligned}
&\|D\frac{1-\nu}{2}\operatorname{rot}(\chi\phi_n^B) - \chi p_{n-1}^B\|_s\\
&\qquad\le \|\chi\left(D\frac{1-\nu}{2}\operatorname{rot}\phi_n^B - p_{n-1}^B\right)\|_s + \|D\frac{1-\nu}{2}\phi_n^B\cdot\mathbf{curl}\,\chi\|_s\\
&\qquad\le C(\|D\frac{1-\nu}{2}\operatorname{rot}\phi_n^B - p_{n-1}^B\|_{s,\Omega_0} + \|\phi_n^B\|_{s,\Omega_1})\\
&\qquad\le C(t^{n+1/2-s}\|g\|_{n-3/2} + t^{n+1/2}\|g\|_{s+n-3/2}).
\end{aligned}
$$

This completes the verification of (5.13). □

6. Other boundary conditions. In this section, we discuss the modifications to the foregoing analysis necessary to handle the remaining four other boundary conditions discussed in the introduction: the hard clamped plate, the soft clamped plate, the hard simply supported plate, and the free plate. We shall see that Theorems 5.1–5.3 remain true as stated in all cases.

For the *hard clamped* and *hard simply supported* plates, these were proved in [1]. (The method of proof was somewhat different and required slightly more regularity to obtain the estimates for ω. However, the present method of proof can easily be adapted to correct this.) Since $\phi_1 = 0$ for these boundary conditions, it follows from (5.9) that $p_1 = 0$ as well. Exploiting this, one may slightly improve the regularity requirements for the estimates of $\boldsymbol{\phi}_1^E$ and p_1^E. See [1] for the precise result.

The analysis for the soft clamped plate is very close to that presented here. The space $\boldsymbol{H}^1$ in which $\boldsymbol{\phi}$ is sought is replaced by the subspace of $\boldsymbol{H}^1$ consisting of functions whose normal component vanishes on the boundary. Because of this, a few terms which we estimated in §5 are zero, so the analysis is slightly simpler. A more essential difference between the soft clamped and soft simply supported plates is that the boundary layer for the former is much weaker. In fact, the boundary layer for the soft clamped plate is weaker than for any of the other four boundary conditions we consider. Specifically, as shown in §3, the boundary-layer expansion functions P_0 and P_1 and consequently $\boldsymbol{\Phi}_0$, $\boldsymbol{\Phi}_1$, and $\boldsymbol{\Phi}_2$ all vanish. Moreover, the interior expansion functions ω_i, ϕ_i, and p_i, $i = 1$ and 3, vanish as well. Consequently, $\boldsymbol{\phi}_0^E = \boldsymbol{\phi}_1^E = \boldsymbol{\phi}_2^E + t^2\phi_2$ and $p_0^E = p_1^E = p_2^E + t^2 p_2$. Thus, for example, we see from Theorem 5.1 that $\phi - \phi_0$ is $O(t^2)$ in H^1 and $p - p_0$ is $O(t^2)$ in L^2. (These quantities are only order $O(t^{1/2})$ for the soft simply supported plate and the free plate and $O(t^{3/2})$ for the hard clamped and hard simply supported plates.)

It remains to consider the case of the free plate. First we summarize some basic existence results for the biharmonic and Reissner–Mindlin plate models with traction boundary conditions. Given functions $g \in L^2(\Omega)$, $f, h \in L^2(\partial\Omega)$, the variational problem to find $\omega \in H^2(\Omega)$ satisfying

$$(\mathcal{C}\,\mathcal{E}(\mathbf{grad}\,\omega), \mathcal{E}(\mathbf{grad}\,\mu)) = (g, \mu) - \langle f, \mu \rangle + \left\langle h, \frac{\partial \mu}{\partial n} \right\rangle \quad \text{for all } \mu \in H^2(\Omega)$$

has a solution if and only if the given data is compatible in the sense that

$$(g, \mu) - \langle f, \mu \rangle + \left\langle h, \frac{\partial \mu}{\partial n} \right\rangle = 0 \quad \text{for all } \mu \in \mathbb{L},$$

where $\mathbb{L}$ denotes the three-dimensional space of linear polynomial functions on Ω. In this case, the solution is determined up to the addition of an arbitrary element of $\mathbb{L}$. Performing integration by parts, one obtains the identity

$$(\mathcal{C}\,\mathcal{E}(\mathbf{grad}\,\omega), \mathcal{E}(\mathbf{grad}\,\mu)) = (D\,\Delta^2\,\omega, \mu) - \langle B_3\omega, \mu \rangle + \left\langle B_2\omega, \frac{\partial \mu}{\partial n} \right\rangle, \quad \omega, \mu \in H^2,$$

where

$$B_2\omega := M_{\boldsymbol{n}}\,\mathbf{grad}\,\omega, \qquad B_3\omega := \frac{\partial}{\partial s} M_{\boldsymbol{s}}\,\mathbf{grad}\,\omega + [\mathbf{div}\,\mathcal{C}\,\mathcal{E}(\mathbf{grad}\,\omega)] \cdot \boldsymbol{n}.$$

From this we deduce the boundary value problem corresponding to the weak formulation just discussed:

$$D\,\Delta^2\omega = g \quad \text{in } \Omega, \qquad B_2\omega = h, \quad B_3\omega = f \quad \text{on } \partial\Omega.$$

Note that the traction-free biharmonic plate problem, i.e., the case when $f = h = 0$, has a solution if and only if the load function g is orthogonal to $\mathbb{L}$.

Analogously, the Reissner–Mindlin boundary value problem for a traction-free plate, given by equations (1.6) and (1.7) and the boundary conditions (1.5), has a solution if and only if the load g is compatible with the traction-free conditions, i.e., it is L^2-orthogonal to $\mathbb{L}$. The solution pair $(\omega, \boldsymbol{\phi})$ is then determined up to the addition of a pair in

$$\mathbb{L}_\nabla := \{\,(l, \mathbf{grad}\, l) \mid l \in \mathbb{L}\,\}.$$

We henceforth assume that g is compatible. We now proceed to the construction of the expansion functions ω_i, $\boldsymbol{\phi}_i$, P_i, and p_i in the case of the free plate. The boundary conditions we use are (3.18), (3.19), and, from the last equality in (1.5),

$$(\boldsymbol{\phi}_i - \mathbf{grad}\,\omega_i)\cdot \boldsymbol{n} = -\lambda^{-1}\frac{\partial \hat{P}_{i-2}}{\partial\theta} \tag{6.1}$$

or, in view of (3.4),

$$\mathbf{div}\, \mathcal{C}\,\mathcal{E}(\boldsymbol{\phi}_i)\cdot\boldsymbol{n} = -\frac{\partial \hat{P}_i}{\partial\theta}. \tag{6.2}$$

Now, from (3.18) and (6.2), we have

$$\begin{aligned}\frac{\partial}{\partial s}M_s\boldsymbol{\phi}_i + \mathbf{div}\,\mathcal{C}\,\mathcal{E}(\boldsymbol{\phi}_i)\cdot\boldsymbol{n} &= \lambda^{-1}D\frac{1-\nu}{2}\frac{\partial}{\partial\theta}\left(\frac{\partial^2\hat{P}_i}{\partial\hat{\rho}^2} + \kappa\frac{\partial\hat{P}_{i-1}}{\partial\hat{\rho}} - \frac{\partial^2\hat{P}_{i-2}}{\partial\theta^2}\right) - \frac{\partial\hat{P}_i}{\partial\theta}\\ &= \lambda^{-1}D(1-\nu)\frac{\partial}{\partial\theta}\left(\kappa\frac{\partial\hat{P}_{i-1}}{\partial\hat{\rho}} - \frac{\partial^2\hat{P}_{i-2}}{\partial\theta^2}\right),\end{aligned}$$

where we used (3.11) with $\hat{\rho} = 0$ in the last step. Using (3.7), we convert this to a boundary condition on ω_i:

$$B_3\omega_i = -B_3(\lambda^{-1}D\,\Delta\,\omega_{i-2} + \delta_{i4}\lambda^{-2}Dg) + \lambda^{-1}D(1-\nu)\frac{\partial}{\partial\theta}\left(\kappa\frac{\partial\hat{P}_{i-1}}{\partial\hat{\rho}} - \frac{\partial^2\hat{P}_{i-2}}{\partial\theta^2}\right). \tag{6.3}$$

The construction of expansion functions satisfying (3.6), (3.7), (3.11), (5.9), (3.18), (3.19), (6.1), and (3.12) proceeds as follows. First we define $\omega_i \in H^2$ from the biharmonic equation (3.6) together with the boundary conditions (3.20), which we may write as

$$B_2\omega_i = -B_2(\lambda^{-1}D\,\Delta\,\omega_{i-2} + \delta_{i4}\lambda^{-2}Dg) + \lambda^{-1}D(1-\nu)\left(\kappa\frac{\partial\hat{P}_{i-2}}{\partial\theta} + \frac{\partial^2\hat{P}_{i-1}}{\partial\theta\partial\hat{\rho}}\right), \tag{6.4}$$

and (6.3). Note that for $i = 0$ this is simply the biharmonic problem for a traction-free plate with load g, so ω_0 is determined up to addition of a linear function. As we shall show shortly, this problem always admits a solution, so that once P_j is known for $j < i$, ω_i is determined up to addition of a linear function. Then $\boldsymbol{\phi}_i$ is given by (3.7) and (3.8) as before, so the pair $(\omega_i, \boldsymbol{\phi}_i)$ is determined up to addition of an element of $\mathbb{L}_\nabla$. Note that $M_s\boldsymbol{\phi}_i$ is determined completely, and so we can uniquely determine P_i by the differential equation (3.11), the boundary condition (3.18), and the decay condition (3.12). Thus we compute, in order, ω_0, ϕ_0, P_0, ω_1, ϕ_1, P_1, ..., always with (ω_i, ϕ_i) determined up to addition of an element of $\mathbb{L}_\nabla$, and P_i determined completely.

To see that the biharmonic problems for the ω_i admit solutions, we must show that

$$(6.5) \qquad (\delta_{i0} g - \delta_{i2}\lambda^{-1} D \,\Delta\, g, \mu) - \langle f, \mu\rangle + \left\langle h, \frac{\partial \mu}{\partial n}\right\rangle = 0 \quad \text{for all } \mu \in \mathbb{L},$$

when f is given by the right-hand side of (6.3) and h by the right-hand side of (6.4). Setting $u = \lambda^{-1} D\,\Delta\,\omega_{i-2} + \delta_{i4}\lambda^{-2} D g$ and using the biharmonic equation satisfied by ω_{i-2} (which we can assume by induction), we get $D\,\Delta^2\, u = \delta_{i2}\lambda^{-1} D g$. Hence if $\mu \in \mathbb{L}$,

$$(\delta_{i0} g - \delta_{i2}\lambda^{-1} D g, \mu) = -\langle B_3 u, \mu\rangle + \left\langle B_2 u, \frac{\partial \mu}{\partial n}\right\rangle.$$

Thus, to complete the verification of (6.5), it suffices to show

$$\left\langle \frac{\partial}{\partial \theta}\left(\kappa \frac{\partial \hat{P}_{i-1}}{\partial \hat{\rho}}\right), \mu \right\rangle = \left\langle \frac{\partial^2 \hat{P}_{i-1}}{\partial \theta \partial \hat{\rho}}, \frac{\partial \mu}{\partial n}\right\rangle$$

and

$$\left\langle \frac{\partial \hat{P}_{i-2}}{\partial \theta^3}, \mu \right\rangle = -\left\langle \kappa \frac{\partial \hat{P}_{i-2}}{\partial \theta}, \frac{\partial \mu}{\partial n}\right\rangle,$$

for all $\mu \in \mathbb{L}$. These may be verified with elementary calculus, independent of the particular functions P_{i-1} and P_{i-2}.

We now define functions p_i and r, as was done in the beginning of §5 for the soft simply supported plate. From (3.7), (3.8), and (3.6), we see that $\operatorname{div} \mathbf{div}\, C\,\mathcal{E}(\phi_i) = \delta_{i0} g$. Hence, defining $r \in H^1/\mathbb{R}$ by

$$-\Delta\, r = g \quad \text{in } \Omega, \qquad \frac{\partial r}{\partial n} = 0 \quad \text{on } \partial\Omega,$$

we see that $\mathbf{div}\, C\,\mathcal{E}(\phi_i) + \delta_{i0}\, \mathbf{grad}\, r$ is divergence free. Hence we may again define a function $p_i \in H^1$, unique modulo $\mathbb{R}$, by (5.9). Now from (5.9) and (6.2), we see that $\partial(p_i + P_i)/\partial s = 0$. Therefore, we may normalize p_i so that

$$(6.6) \qquad p_i + P_i = 0 \quad \text{on } \partial\Omega.$$

This completes the construction of the expansion functions.

In §§4 and 5, we presented the analysis of the asymptotic expansions in such a way that they adapt with a minimum of effort to the case of the free plate. Due to the different boundary conditions, we need to use different negatively indexed Sobolev norms. Instead of the definition given in §2, we define $\|\cdot\|_s$ to be the norm in the dual space H^s. With this understanding, all of the results of §4 hold with essentially the same proofs. Of course, in the proof of Theorem 4.3, we use the traction problem for the biharmonic rather than the simply supported plate problem.

Turning to the error analysis in §5, we again use the Helmholtz decomposition as in (5.1), except that now $r \in H^1/\mathbb{R}$ and $p \in \overset{\circ}{H}{}^1$. We then recover the differential equations (5.2)–(5.4), and (5.6), now with the boundary conditions

$$\frac{\partial r}{\partial n} = M_n \phi = M_s \phi = p = \frac{\partial \omega}{\partial n} - \phi \cdot \boldsymbol{n} = 0.$$

These determine (r, ϕ, p, ω) up to an additive constant in r and addition of an element of $\mathbb{L}_\nabla$ to (ω, ϕ).

The norms on the left-hand sides of the estimates in Theorems 5.1–5.3 need to be modified in the obvious ways because of the indeterminancy in (ϕ, ω) and the determinancy of p. That is, the norms of ϕ_n^E are in the Sobolev spaces modulo $\mathbb{R}$, those on ω_n^E in the Sobolev spaces modulo $\mathbb{L}$, and those on p_n^E in the full Sobolev spaces. The proofs of these theorems carry over easily. In particular, Lemma 5.4 holds without change.

The main part of the proof of Theorem 5.1 involved the choice of test functions $\psi = \phi_n^E$ in (5.20) and $q = p_n^E - t^n \chi P_n$ in (5.21). Notice that this choice of q vanishes on the boundary because of (6.6) and so is an allowable test function. This part of the proof carries over to the free case without problem.

Two more choices of test functions complete the proof of the theorem. For the second one, we take $\psi \in \boldsymbol{H}^1$ with $\operatorname{rot} \psi = p_n^E$, which allows us to get control over the full L^2 norm of p_n^E. Finally, to control the infinitesimal rotation in ϕ_n^E, we choose a test function $q \in \mathring{H}^1$ in (5.21) with nonvanishing integral and use the fact that

$$\psi \to a(\psi, \psi)^{1/2} + \left| \int_\Omega q \operatorname{rot} \psi \right|$$

defines a norm equivalent to the usual norm in $\boldsymbol{H}^1/\mathbb{R}^2$.

The proof of Theorem 5.2 adapts easily. Naturally, we use a Dirichlet rather than a Neumann problem to obtain bounds on p_n^E, using that fact that $p_n^E = t^n P_n$ on $\partial\Omega$. Analogously, to prove Theorem 5.3, we use a Neumann problem for ω_n^E and the fact that

$$\frac{\partial \omega_n^E}{\partial n} = \phi_n^E \cdot \boldsymbol{n} = \phi_{n+s-1}^E \cdot \boldsymbol{n} + \sum_{n+1}^{n+s-1} t^j (\phi_j + \boldsymbol{\Phi}_j) \cdot \boldsymbol{n}.$$

Appendix. In this appendix, we collect some elementary formulas for the convenience of the reader.

It follows immediately from the definitions of rot and **curl** that

$$\operatorname{rot} \mathbf{curl}\, q = -\Delta q, \qquad \mathbf{curl}\, q \cdot \boldsymbol{n} = -\frac{\partial q}{\partial s}, \qquad \mathbf{curl}\, q \cdot \boldsymbol{s} = \frac{\partial q}{\partial n},$$

and

$$(\operatorname{rot} \psi, q) = (\psi, \mathbf{curl}\, q) - \langle \psi \cdot \boldsymbol{s}, q \rangle.$$

Simple computations show that

$$\begin{aligned} \mathbf{div}\, \mathcal{C}\,\mathcal{E}(\mathbf{grad}\, v) &= D\, \mathbf{grad}\, \Delta v, & \mathbf{div}\, \mathcal{C}\,\mathcal{E}(\mathbf{curl}\, p) &= D \frac{1-\nu}{2}\, \mathbf{curl}\, \Delta p, \\ \operatorname{div} \mathbf{div}\, \mathcal{C}\,\mathcal{E}(\phi) &= D\, \Delta \operatorname{div} \phi, & \Delta \phi &= \mathbf{grad} \operatorname{div} \phi - \mathbf{curl} \operatorname{rot} \phi, \end{aligned}$$

and on $\partial\Omega$,

$$
\begin{aligned}
M_{\boldsymbol{n}}\boldsymbol{\phi} &:= \boldsymbol{n}\cdot \mathcal{C}\,\mathcal{E}(\boldsymbol{\phi})\boldsymbol{n} = D\left(\frac{\partial\boldsymbol{\phi}}{\partial n}\cdot\boldsymbol{n} + \nu\frac{\partial\boldsymbol{\phi}}{\partial s}\cdot\boldsymbol{s}\right),\\
M_{\boldsymbol{s}}\boldsymbol{\phi} &:= \boldsymbol{s}\cdot \mathcal{C}\,\mathcal{E}(\boldsymbol{\phi})\boldsymbol{n} = \frac{D(1-\nu)}{2}\left(\frac{\partial\boldsymbol{\phi}}{\partial n}\cdot\boldsymbol{s} + \frac{\partial\boldsymbol{\phi}}{\partial s}\cdot\boldsymbol{n}\right),\\
M_{\boldsymbol{n}}(\mathbf{grad}\,v) &= D\left[(1-\nu)\frac{\partial^2 v}{\partial n^2} + \nu\,\Delta\,v\right],\\
M_{\boldsymbol{s}}(\mathbf{grad}\,v) &= D(1-\nu)\left[\frac{\partial^2 v}{\partial s\partial n} - \kappa\frac{\partial v}{\partial s}\right],\\
M_{\boldsymbol{n}}(\mathbf{curl}\,p) &= D(1-\nu)\left[-\frac{\partial^2 p}{\partial s\partial n} + \kappa\frac{\partial p}{\partial s}\right],\\
M_{\boldsymbol{s}}(\mathbf{curl}\,p) &= D\frac{(1-\nu)}{2}\left[\frac{\partial^2 p}{\partial n^2} - \frac{\partial^2 p}{\partial s^2} - \kappa\frac{\partial p}{\partial n}\right],
\end{aligned}
$$

and

$$
\frac{\partial\boldsymbol{n}}{\partial s} = \kappa\boldsymbol{s}, \qquad \frac{\partial\boldsymbol{s}}{\partial s} = -\kappa\boldsymbol{n},
$$

$$
\boldsymbol{n}\cdot\mathcal{H}(v)\boldsymbol{n} = \frac{\partial^2 v}{\partial n^2}, \qquad \boldsymbol{s}\cdot\mathcal{H}(v)\boldsymbol{n} = \frac{\partial^2 v}{\partial s\partial n} - \kappa\frac{\partial v}{\partial s}, \qquad \boldsymbol{s}\cdot\mathcal{H}(v)\boldsymbol{s} = \frac{\partial^2 v}{\partial s^2} + \kappa\frac{\partial v}{\partial n},
$$

where $\mathcal{H}(v)$ denotes the Hessian matrix of second partial derivatives of v.

REFERENCES

[1] D. N. Arnold and R. S. Falk, *The boundary layer for the Reissner-Mindlin plate model*, SIAM J. Math. Anal., 21 (1990), pp. 281–312.

[2] ———, *Edge effects in the Reissner–Mindlin plate theory*, in Analytical and Computational Models for Shells, A. K. Noor, T. Belytschko, and J. Simo, eds., American Society of Mechanical Engineers, New York, 1989, pp. 71–90.

[3] F. Brezzi and M. Fortin, *Numerical approximation of Mindlin–Reissner plates*, Math. Comp., 47 (1986), pp. 151–158.

SIAM J. MATH. ANAL.
Vol. 27, No. 2, pp. 515–527, March 1996

010

TWO PROBLEMS FROM DRAINING FLOWS INVOLVING THIRD-ORDER ORDINARY DIFFERENTIAL EQUATIONS*

F. BERNIS[†] AND L. A. PELETIER[‡]

Abstract. A mathematical analysis is given of two third-order ordinary differential equations which arise in models for flows of thin viscous films over solid surfaces. Questions about existence, uniqueness, and qualitative properties of solutions are discussed.

Key words. differential equations, nonlinear, fluid mechanics, coating flow

AMS subject classifications. 34B15, 34E05, 76D99

1. Introduction.

In a recent survey paper, Tuck and Schwartz [12] discussed a series of third-order ordinary differential equations (ODEs) arising in the study of the flow of a thin film of viscous fluid over a solid surface. When such a film drains down a vertical wall and the effects of surface tension and gravity as well as viscosity are taken into account, one is led to an equation of the form

$$\frac{d^3u}{dx^3} = f(u)$$

for the film profile $u(x)$ in a coordinate frame moving with the fluid.

In [12] different possible choices of the function f are given. For drainage down a dry surface with the x-axis pointing downwards, this function becomes

$$\text{(A)} \qquad f(u) = -1 + \frac{1}{u^2}.$$

This function is singular at $u = 0$, that is, at the tip of the film. If the surface is prewetted by a very thin film of thickness $\delta > 0$, the function f becomes

$$\text{(B)} \qquad f(u) = -1 + \frac{1+\delta+\delta^2}{u^2} - \frac{\delta+\delta^2}{u^3}.$$

Since u may now be expected to be bounded away from zero, the singularity at $u = 0$ is no longer relevant.

When the surface is dry, insight into the shape of the film close to the tip may be obtained by studying the limit of solutions of equation (B) as $\delta \to 0$. In suitably scaled coordinates this leads to equations involving the functions [1, 2, 11, 12]

$$\text{(C),(D)} \qquad f(u) = \frac{1}{u^2} \quad \text{and} \quad f(u) = \frac{1}{u^2} - \frac{1}{u^3}.$$

Equation (A) also occurs in different film flows, such as spin coating and spray coating [7, 8]. In addition to the asymptotic context given above, equation (C) is interesting in its own right in that it describes the spreading of certain oil drops on horizontal surfaces [9].

In [12] the authors formulate a series of well-posed mathematical problems arising from the study of these draining flows. In this paper we address two of them. The

* Received by the editors December 29, 1993; accepted for publication (in revised form) June 14, 1994.

[†] Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain.

[‡] Mathematical Institute, Leiden University, PB 9512, 2300 RA Leiden, The Netherlands.

first, problem (I), involves the simpler function (C) and occurs in the asymptotic analysis near the tip. The second, problem (II), involves the original function (A) and describes the draining of a film along a dry wall, which is uniform far upstream.

The objective of our analyses is to prove some basic properties of these problems, such as the existence of a solution and the domain on which it exists, its uniqueness and such qualitative properties as monotonicity or oscillatory behaviour, and asymptotics far up- and downstream.

A comparable analysis for equations (B) and (D) has been given by Troy [11].

Let us now state the first problem in detail. We look for a smooth function $u(x)$, defined for all $x \in (-\infty, \infty)$, which has the following properties:

$$\text{(I)} \qquad \begin{cases} u''' = \dfrac{1}{u^2}, \\ u(0) = 1 \quad \text{and} \quad u'(0) = 0, \\ u''(x) \to 0 \quad \text{as} \quad x \to -\infty. \end{cases}$$

We shall prove that such a solution indeed exists, is also unique, and has the following asymptotic behaviour at $-\infty$ and at $+\infty$:

$$u(x) \sim -x(3 \log |x|)^{1/3} \quad \text{as} \quad x \to -\infty, \tag{1.1}$$

$$u(x) \sim \frac{1}{2} K x^2 \quad \text{as} \quad x \to +\infty. \tag{1.2}$$

Here K is a positive constant. In fact, $K = \lim_{x \to +\infty} u''(x)$. The associated numerical values given in [12] are

$$u''(0) = 1.2836 \qquad \text{and} \qquad K = 2.1591.$$

It is interesting to note that the behaviour near $x = -\infty$ is the same as that found in [11] for equation (D).

These asymptotic estimates have been obtained before by formal methods [1, 5, 12].

We shall prove these results by transforming the third-order differential equation to the classical Emden–Fowler equation,

$$y'' + t^{-k} y^{\sigma} = 0, \quad t > 0,$$

with $k = 2$ and $\sigma = -\frac{1}{2}$. Studying this equation instead and making use of its specific properties, such as the convexity of its solutions, we obtain the required information about u.

It is interesting to mention here a somewhat related transformation, recently proposed in [4], which casts the equation into an autonomous system of two first-order equations.

The second problem we consider is

$$\text{(II)} \qquad \begin{cases} u''' = -1 + \dfrac{1}{u^2}, \\ u(x) \to 1 \quad \text{as} \quad x \to -\infty. \end{cases}$$

Clearly this problem has the trivial solution $u(x) = 1$.

Numerical studies [12] suggest that nontrivial solutions of (II) exist and are positive and oscillatory for all $x \in (-\infty, \infty)$, with increasing maxima and decreasing minima. The objective of our analysis of problem (II) is to confirm these observations and explore the character of the oscillations. Specifically, we shall prove the following sequence of results about nontrivial solutions u:

A. u exists and is positive on the whole line.

B. $u-1$ has infinitely many zeros; these zeros form an increasing sequence $\{a_n\}$, $-\infty < n < +\infty$ and $a_n \to -\infty$ as $n \to -\infty$, while $a_n \to +\infty$ as $n \to +\infty$.

C. u has a unique minimum in each interval (a_{2n}, a_{2n+1}), attained at a point b_{2n}. The sequence of minima $u(b_{2n})$ is decreasing and

$$u(b_{2n}) \to 0 \quad \text{as} \quad n \to +\infty.$$

D. u has a unique maximum in each interval (a_{2n+1}, a_{2n+2}), attained at a point b_{2n+1}. The sequence of maxima $u(b_{2n+1})$ is increasing and

$$u(b_{2n+1}) \to \infty \quad \text{as} \quad n \to +\infty.$$

E. The length of the intervals in which $u < 1$ tends to zero, while the length of the intervals in which $u > 1$ tends to infinity. More precisely,

$$a_{2n+1} - a_{2n} \to 0 \quad \text{as} \quad n \to +\infty,$$

$$a_{2n+2} - a_{2n+1} \to \infty \quad \text{as} \quad n \to +\infty.$$

F. The sequences

$$|u'(a_n)|, \quad |u'(c_n)|, \quad |u''(a_n)|, \quad |u''(b_n)|$$

in which the points c_n are the zeros of u'', are increasing and tend to infinity as $n \to +\infty$.

It is remarkable that property **F** does not discriminate between the intervals in which $u > 1$ and those in which $u < 1$.

2. Problem (I): Existence and uniqueness. We recall that problem (I) is

$$u''' = \frac{1}{u^2}, \tag{2.1}$$

$$u(0) = 1 \quad \text{and} \quad u'(0) = 0, \tag{2.2}$$

$$u''(x) \to 0 \quad \text{as} \quad x \to -\infty. \tag{2.3}$$

THEOREM 2.1. *There exists a unique solution $u = u(x)$ of problem* (2.1)–(2.3). *This solution is defined for all $x \in \mathbf{R}$.*

We begin with some preliminary observations. Suppose that $u(x)$ is a solution of (2.1)–(2.3). Then, because $u'''(x) > 0$ everywhere and $u''(-\infty) = 0$, we can inmediately conclude that

$$u(x) \geq 1 \quad \text{and} \quad u''(x) > 0 \quad \text{for all} \quad x \in \mathbf{R}, \tag{2.4}$$

$$u'(x) \begin{cases} < 0 & \text{if} \quad x < 0, \\ > 0 & \text{if} \quad x > 0, \end{cases} \tag{2.5}$$

and

$$u(x) \to +\infty \quad \text{as} \quad x \to -\infty.$$

Thus, since any solution is strictly decreasing on $(-\infty, 0)$, we may introduce u as an independent variable and, as in [6], introduce the function

$$y(u) = \{u'(x)\}^2 \tag{2.6}$$

as a dependent variable. (Notice that $u' = -\sqrt{y}$.) Carrying out the transformation, we obtain for y the problem

$$y'' + \frac{2}{u^2}\frac{1}{\sqrt{y}} = 0, \quad y > 0 \quad \text{for} \quad 1 < u < \infty, \tag{2.7}$$

$$y(1) = 0, \quad y'(u) \to 0 \quad \text{as} \quad u \to +\infty. \tag{2.8}$$

It will be sufficient to prove the existence and uniqueness of a solution of problem (2.7)–(2.8). We shall do this by means of a shooting argument, replacing the condition at $u = +\infty$ by a second condition at $u = 1$, so that we then have

$$y(1) = 0 \quad \text{and} \quad y'(1) = \alpha > 0. \tag{2.9}$$

We could also use Proposition 2.4 of [10], which establishes existence and uniqueness by a different argument.

It is readily seen that for any $\alpha > 0$, problem (2.7), (2.9) has a unique local solution $y(u, \alpha)$ which can be continued as long as $y > 0$. Because the graph of $y(u, \alpha)$ is concave, it follows that if y exists on the whole half line $[1, \infty)$, then

$$y'(u, \alpha) > 0 \text{ for all } u > 1 \quad \text{and} \quad \lim_{u \to \infty} y'(u, \alpha) \text{ exists.}$$

Moreover, we have the following monotonicity lemma.

LEMMA 2.2. *Suppose that $\alpha_1 < \alpha_2$. Then, as long as $y(u, \alpha_1)$ exists,*

$$y(u, \alpha_1) < y(u, \alpha_2) \quad \textit{and} \quad y'(u, \alpha_1) < y'(u, \alpha_2).$$

In addition, if $y(u, \alpha_1)$ exists for all $u \geq 1$, then

$$0 \leq y'(\infty, \alpha_1) < y'(\infty, \alpha_2) < \infty.$$

Proof. For convenience we shall write $y_i(u) = y(u, \alpha_i)$, $i = 1, 2$. It is enough to prove that $y_1'(u) < y_2'(u)$.

Since $\alpha_1 < \alpha_2$, the assertion is true near $u = 1$. Suppose it first fails to hold at some point $u_0 > 1$. Then

$$y_1'(u_0) = y_2'(u_0) \quad \text{and} \quad y_1' < y_2' \quad \text{on} \quad (1, u_0). \tag{2.10}$$

However, if we integrate the equations for y_1 and y_2 over $(1, u_0)$ and subtract, we obtain

$$y_1'(u_0) - y_2'(u_0) = \alpha_1 - \alpha_2 - 2\int_1^{u_0} \left(\frac{1}{\sqrt{y_1}} - \frac{1}{\sqrt{y_2}} \right) \frac{ds}{s^2} < 0 \tag{2.11}$$

because $y_1 < y_2$ on $(1, u_0)$. This contradicts (2.10).

The second assertion follows from (2.11) when we set $u_0 = \infty$. Observe that $y_i'(\infty)$ must be finite and nonnegative.

We now define two sets of initial slopes:

$$S_+ = \{\alpha > 0 \ : \ y \text{ exists on } [1, \infty) \text{ and } y'(\infty, \alpha) > 0\},$$
$$S_- = \{\alpha > 0 \ : \ y \text{ vanishes at some } u_0 \in [1, \infty)\}.$$

It is clear that $S_+ \cap S_- = \emptyset$ and by continuous dependence on the initial data, both sets are open. In the next two lemmas, we shall show that they are nonempty.

LEMMA 2.3. $S_+ \neq \emptyset$.

Proof. Suppose that $S_+ = \emptyset$. Then for any $\alpha > 0$ there exists a point u_α, which may be infinite, such that $y(\cdot, \alpha)$ is increasing on $(1, u_\alpha)$ and $y'(u_\alpha, \alpha) = 0$. If we integrate (2.7) over $(1, u_\alpha)$, we obtain

$$\alpha = 2 \int_1^{u_\alpha} \frac{ds}{s^2 \sqrt{y(s,\alpha)}}.$$

Define

$$\hat{y}(u) = \begin{cases} y(u,1) & \text{for } 1 < u < u_1, \\ y(u_1,1) & \text{for } u_1 \leq u < \infty. \end{cases}$$

Then, for $\alpha > 1$,

$$\alpha < 2 \int_1^{u_\alpha} \frac{ds}{s^2 \sqrt{\hat{y}(s)}} < 2 \int_1^{\infty} \frac{ds}{s^2 \sqrt{\hat{y}(s)}} < \infty$$

and we have a contradiction.

LEMMA 2.4. $S_- \neq \emptyset$.

Proof. Observe that by concavity,

$$y(u,\alpha) < \alpha(u-1) \quad \text{on} \quad (1, \sigma_\alpha),$$

where $[1, \sigma_\alpha)$ is the maximal interval of existence of $y(\cdot, \alpha)$. Hence

$$y''(u,\alpha) < -\frac{2}{\sqrt{\alpha}} \frac{1}{u^2 \sqrt{u-1}} \quad \text{on} \quad (1, \sigma_\alpha)$$

and so

$$y'(u,\alpha) < \alpha - \frac{2}{\sqrt{\alpha}} \int_1^u \frac{ds}{s^2 \sqrt{s-1}} \quad \text{for} \quad 1 < u < \sigma_\alpha.$$

Thus, if S_- were empty and so $y'(u,\alpha) > 0$ for all $u \geq 1$ and all $\alpha > 0$, then

$$\alpha^{3/2} \geq 2 \int_1^{\infty} \frac{ds}{s^2 \sqrt{s-1}} = \pi$$

for all $\alpha > 0$. This is clearly a contradiction.

COROLLARY 2.5. $(0, \pi^{2/3}] \subset S_-$.

From Lemmas 2.3 and 2.4, we conclude that there exists a shooting angle α_0 such that $y_0(u) = y(u, \alpha_0)$ has the properties

$$y_0(u) > 0 \quad \text{and} \quad y_0'(u) > 0 \quad \text{for all} \quad u > 1$$

and

$$\lim_{u \to +\infty} y_0'(u) = 0.$$

This is the desired solution of problem (2.7), (2.8), from which we deduce at once that problem (2.1)–(2.3) has a solution $u(x)$ defined for $x \in (-\infty, 0]$. But this solution $u(x)$ can be continued to the whole line $\mathbf{R}$ because, as explained above, $u(x) \geq 1$ and hence u''' remains bounded.

By the strict monotonicity of $y'(\infty, \alpha)$ with respect to α, established in Lemma 2.2, the solution $y_0(u)$ must be unique and hence $u(x)$ is also unique.

The proof of Theorem 2.1 is now complete.

3. Problem (I): Asymptotic behaviour. Most of this section is devoted to proving the asymptotic estimate (1.1), which by definition means that

$$\lim_{x\to-\infty} \frac{u(x)}{-x(3\log|x|)^{1/3}} = 1. \tag{3.1}$$

To study the behaviour of the solution $u(x)$ of problem (I) as $x \to -\infty$ we consider again the change of variables (2.6) and equation (2.7) and recall that $x \to -\infty$ for $u(x)$ corresponds to $u \to +\infty$ for $y(u)$.

We present a self-contained proof of Theorem 3.3 below, which could also be derived from [10, Thm. 3.6].

In what follows $y(u)$ stands for a solution of (2.7) such that $y'(u) \to 0$ as $u \to +\infty$.

LEMMA 3.1. *We have*

$$\limsup_{u\to\infty} \frac{y(u)}{(\log u)^{2/3}} \ge 3^{2/3} \equiv M.$$

Proof. Suppose to the contrary that

$$\limsup_{u\to\infty} \frac{y(u)}{(\log u)^{2/3}} \le M - 2\varepsilon \tag{3.2}$$

for some $\varepsilon > 0$. Then there exists a $u_\varepsilon > 0$ such that

$$y(u) \le (M-\varepsilon)(\log u)^{2/3} \quad \text{if} \quad u > u_\varepsilon.$$

Hence

$$y''(u) \le -\frac{2}{\sqrt{M-\varepsilon}}\frac{1}{u^2}(\log u)^{-1/3} \quad \text{if} \quad u > u_\varepsilon$$

and, because by assumption $y'(\infty) = 0$,

$$y'(u) \ge \frac{2}{\sqrt{M-\varepsilon}} J(u),$$

where

$$J(u) = \int_u^\infty \frac{dt}{t^2(\log t)^{1/3}} \sim \frac{1}{u(\log u)^{1/3}} \quad \text{as} \quad u \to \infty.$$

Thus

$$y(u) \ge y(u_\varepsilon) + \frac{2}{\sqrt{M-\varepsilon}} \int_{u_\varepsilon}^u J(t)\,dt,$$

whence

$$\liminf_{u\to\infty} \frac{y(u)}{(\log u)^{2/3}} \ge \frac{2}{\sqrt{M-\varepsilon}} \lim_{u\to\infty} \frac{\int_{u_\varepsilon}^u J(t)\,dt}{(\log u)^{2/3}} = \frac{3}{\sqrt{M-\varepsilon}}, \tag{3.3}$$

because by l'Hôpital's rule

$$\lim_{u\to\infty} \frac{\int_{u_\varepsilon}^u J(t)\,dt}{(\log u)^{2/3}} = \lim_{u\to\infty} \frac{J(u)}{\frac{2}{3}(\log u)^{-1/3}\frac{1}{u}} = \frac{3}{2}.$$

Remark. Note that

$$\int_{u_\varepsilon}^u J(t)\,dt \to \infty \quad \text{as} \quad u \to \infty.$$

We conclude from (3.3) that

$$\limsup_{u\to\infty}\frac{y(u)}{(\log u)^{2/3}} \geq \frac{3}{\sqrt{M-\varepsilon}} > \frac{3}{\sqrt{M}} = 3^{2/3}, \tag{3.4}$$

which contradicts the assumption (3.2).

Similarly one can prove the following lemma.

LEMMA 3.2. *We have*

$$\liminf_{u\to\infty}\frac{y(u)}{(\log u)^{2/3}} \leq 3^{2/3}.$$

Define

$$\varphi(u) = (3\log u)^{2/3}.$$

Then

$$\varphi'' + \frac{2}{u^2}\frac{1}{\sqrt{\varphi}} = -\frac{2}{u^2\varphi^2} < 0. \tag{3.5}$$

We proceed to compare $y(u)$ and $\varphi(u)$.

THEOREM 3.3. *Let $y(u)$ be a solution of* (2.7) *such that $y'(u) \to 0$ as $u \to \infty$. Then*

$$\lim_{u\to\infty}\frac{y(u)}{\varphi(u)} = 1. \tag{3.6}$$

Proof. We distinguish two cases:

(a) $y(u) - \varphi(u) \neq 0$ for all $u > 1$;

(b) $\exists u_0 > 1$ such that $y(u_0) - \varphi(u_0) = 0$.

Case (a). Suppose that $y < \varphi$. Then

$$\limsup_{u\to\infty}\frac{y(u)}{\varphi(u)} \leq 1.$$

Proceeding as is the proof of Lemma 3.1, with $\varepsilon = 0$, we conclude that

$$\liminf_{u\to\infty}\frac{y(u)}{\varphi(u)} \geq 1,$$

from which we deduce (3.6). If $y > \varphi$ the argument is the same.

Case (b). Let

$$y(u_0) - \varphi(u_0) = 0$$

and let $y > \varphi$ in a right neighbourhood of u_0. We assert that $y > \varphi$ on (u_0, ∞). In fact, assume for contradiction that

$$u_1 \equiv \sup\{u > u_0 \,:\, y > \varphi \text{ on } (u_0, u)\} < \infty.$$

Setting $z = y - \varphi$ and taking (3.5) into account, we find that

$$z'' > 0 \quad \text{and} \quad z > 0 \quad \text{on} \quad (u_0, u_1).$$

This and $z(u_0) = z(u_1) = 0$ contradict the maximum principle (or the convexity of z). Thus $y > \varphi$ on (u_0, ∞) and as in Case (a) we conclude that (3.6) holds.

Finally, let $y < \varphi$ on $(u_0, u_0 + \delta)$ for some $\delta > 0$. If $y \leq \varphi$ in (u_0, ∞), then (3.6) follows as before. On the other hand, if

$$u_1 \equiv \sup\{u > u_0 \,:\, y \leq \varphi \text{ on } (u_0, u)\} < \infty,$$

then $y > \varphi$ on (u_1, ∞) and (3.6) follows again. This completes the proof of Theorem 3.3.

Remark. Theorem 3.3 only involves local arguments near $u = +\infty$. In particular, the condition $y(1) = 0$ of (2.8) is not assumed.

THEOREM 3.4. *The solution of problem* (I) *satisfies* (1.1).

Proof. Using the change of variables (2.6), the relation between $u(x)$ and $y(u)$ can be written as

$$u'(x) = -\sqrt{y(u(x))}, \qquad -\infty < x \leq 0.$$

We divide by $\sqrt{y}$ and integrate over $(x, 0)$. This yields

$$\int_1^{u(x)} \frac{dv}{\sqrt{y(v)}} = -x,$$

and by Theorem 3.3

$$\int_1^{u(x)} \frac{dv}{(3\log v)^{1/3}} \sim -x \quad \text{as} \quad x \to -\infty. \tag{3.7}$$

On the other hand, by l'Hôpital's rule

$$\int_1^u \frac{dv}{(3\log v)^{1/3}} \sim \frac{u}{(3\log u)^{1/3}} \quad \text{as} \quad u \to +\infty.$$

Therefore, (3.7) yields

$$\frac{u(x)}{\{3\log u(x)\}^{1/3}} \sim -x \quad \text{as} \quad x \to -\infty$$

and hence

$$\log u(x) \sim \log|x| \quad \text{as} \quad x \to -\infty.$$

The last two relations imply that

$$u(x) \sim -x(3\log|x|)^{1/3} \quad \text{as} \quad x \to -\infty,$$

which is what we set out to prove.

Remark. It also follows that as $x \to -\infty$

$$u'(x) \sim -(3\log|x|)^{1/3} \quad \text{and} \quad u''(x) \sim -\frac{1}{x(3\log|x|)^{2/3}}.$$

Finally, we briefly deal with the behaviour as $x \to +\infty$.

THEOREM 3.5. *Let* $u(x)$ *be the solution of problem* (I). *Then*

$$K \equiv \lim_{x\to+\infty} u''(x) \ \textit{exists.}$$

Moreover, K is positive and finite so that

$$u(x) \sim \frac{1}{2}Kx^2 \quad \textit{as} \quad x \to +\infty$$

Proof. Since $u''' > 0$ and $u''(-\infty) = 0$, it is clear that the limit K exists and that it is positive. Thus, we only need to prove that it is finite. Because K is positive, it follows that $u(x) > Cx^2$ for some positive constant C and x sufficiently large. This implies by the differential equation (2.1) that $u'''(x) < C^{-2}x^{-4}$ so that u''' is integrable near infinity and hence that u'' tends to a finite limit.

4. Problem (II).

We recall that problem (II) is

$$u''' = -1 + \frac{1}{u^2}, \tag{4.1}$$

$$u(x) \to 1 \quad \text{as} \quad x \to -\infty. \tag{4.2}$$

In this section we always assume that u is a nontrivial solution of problem (II). By definition, we require that u be of class C^3. This implies that u must be positive because of the singularity at $u = 0$ of the differential equation (4.1). Hence, the assertion that u is everywhere positive is included in global existence (Theorem 4.5 below).

The linearization of equation (4.1) about $u = 1$ is $u''' = -2u$ and has the eigenvalues

$$-\sqrt[3]{2}, \qquad \sqrt[3]{2}\left(\frac{1}{2} \pm i\frac{\sqrt{3}}{2}\right).$$

Hence, by standard linearization theory, nontrivial solutions of problem (II) actually exist, have infinitely many zeros near $-\infty$, and satisfy

$$u(x) \to 1, \quad u'(x) \to 0, \quad u''(x) \to 0 \qquad \text{as} \qquad x \to -\infty. \tag{4.3}$$

Notice that for the moment we only know that these solutions are defined near $-\infty$.

We define the following four auxiliary functions:

$$\Phi_1 \equiv (1-u)u'' + \frac{1}{2}u'^2,$$

$$\Phi_2 \equiv u + \frac{1}{u} + u'u'',$$

$$\Phi_3 \equiv -u'u''' + \frac{1}{2}u''^2,$$

$$\Phi_4 \equiv \frac{3}{2}\left(u + \frac{1}{u}\right) + \frac{1}{2}u'u''.$$

LEMMA 4.1. *If u is a nontrivial solution of problem* (II) *then the functions Φ_1, $\Phi_2 - 2$, Φ_3 and $\Phi_4 - 3$ are strictly increasing and positive. Furthermore, Φ_4 is convex.*

Proof. Since Φ_1, $\Phi_2 - 2$, Φ_3 and $\Phi_4 - 3$ tend to zero as $x \to -\infty$ by (4.3), it is enough to prove that they are strictly increasing and that Φ_4 is convex. Taking into account the differential equation, we obtain

$$\Phi_1' = (1-u)u''' = \frac{(1-u)^2(1+u)}{u^2} \geq 0,$$

$$\Phi_2' = u''^2 \quad \text{because} \quad \frac{d}{dx}\left(u + \frac{1}{u}\right) = -u'u''',$$

$$\Phi_3' = -u'u^{(4)} = \frac{2u'^2}{u^3},$$

$$\Phi_4' = \Phi_3 > 0 \quad \text{and hence} \quad \Phi_4'' = \Phi_3' \geq 0.$$

Notice that the monotonicity of these functions is strict. In fact, if one of the derivatives Φ_i' were zero in an interval, then equation (4.1) would imply that $u = 1$ in that interval and hence u would be the trivial solution by standard ODE uniqueness theorems.

We proceed to analyze the structure of the arches of u.

LEMMA 4.2. *Suppose that a_n and a_{n+1} are two consecutive zeros of $u-1$. Then in the interval $[a_n, a_{n+1}]$, the derivative u' has a unique zero b_n and also u'' has a unique zero c_n. Furthermore,*

$$u'(a_n)u''(a_n) > 0 \quad \text{and} \quad a_n < c_n < b_n < a_{n+1}.$$

Proof. Assume that $u > 1$ in (a_n, a_{n+1}). (The case in which $u < 1$ is analogous.) Then $u''' < 0$, u'' is decreasing, and u' is concave in this interval. From the positivity of $\Phi_2 - 2$, we deduce that

$$u'(a_n) > 0, \quad u''(a_n) > 0, \quad u'(a_{n+1}) < 0, \quad \text{and} \quad u''(a_{n+1}) < 0.$$

This proves the existence and uniqueness of b_n and c_n. Finally, $u''(b_n) < 0$ by the positivity of Φ_1 and hence $c_n < b_n$.

In what follows a_n, b_n, and c_n stand for zeros of $u-1$, u', and u'', respectively, and we introduce the notation

$$\begin{aligned} &(a_{2n}, a_{2n+1}) && \text{for intervals where} && u < 1, \\ &(a_{2n+1}, a_{2n+2}) && \text{for intervals where} && u > 1. \end{aligned}$$

LEMMA 4.3. *As n increases, the minima $u(b_{2n})$ are decreasing and the maxima $u(b_{2n+1})$ are increasing.*

Proof. From the monotonicity of Φ_2 (Lemma 4.1), we deduce that

$$u(b_n) + \frac{1}{u(b_n)} \quad \text{is increasing.}$$

The lemma follows by observing that the numeric function

$$s \mapsto s + \frac{1}{s}$$

is decreasing if $0 < s < 1$ and increasing if $s > 1$.

LEMMA 4.4. *For all n,*

$$u(b_{2n+1}) - 1 \leq (a_{2n+2} - a_{2n+1})^3. \tag{4.4}$$

Proof. Observe that in the intervals where $u > 1$, we have that $\sup |u'''| \leq 1$. The lemma follows by integration, since u'', u', and $u-1$ have zeros in $[a_{2n+1}, a_{2n+2}]$.

Our next result is global existence.

THEOREM 4.5. *All solutions of problem* (II) *exist on the whole line* $\mathbf{R}$.

Proof. Let $(-\infty, b)$ be the maximal interval of existence of u. Assume (for contradiction) that $b < \infty$. We consider two cases.

Case 1. Suppose that $u - 1$ and hence u''' have infinitely many zeros near b. By Lemma 4.3

$$\lim_{n\to\infty} u(b_{2n}) \text{ exists.}$$

This limit can only be zero. Otherwise equation (4.1) implies that u''' would be bounded in a neighbourhood of b and the interval would not be maximal. Hence by Lemma 4.1 $\Phi_2(b^-) = \infty$, and by Lemma 4.3

$$\lim_{n\to\infty} u(b_{2n+1}) = \infty.$$

But this and (4.4) contradict $b < \infty$.

Case 2. If u''' has constant sign near b, then $u'(b^-)$ and $u(b^-)$ exist. As before $u(b^-) = 0$ since $(-\infty, b)$ is maximal, and $u'(b^-) \leq 0$. Furthermore, $u'(b^-) \neq -\infty$ because $u'''(b^-) = +\infty$. Thus, $u'(b^-)$ is *finite* and in a left neighbourhood of b, we have

$$u(x) \leq C(b-x).$$

This implies by equation (4.1) and one integration that

$$u'''(x) \geq \frac{C}{(b-x)^2} \qquad \text{and} \qquad u''(x) \geq \frac{C}{b-x},$$

where C denotes a positive constant, possibly a different one at each occurrence. The last inequality implies that $u'(b^-) = \infty$. This contradiction completes the proof of Theorem 4.5.

Remark. The exponent 2 in the denominator of equation (4.1) is a borderline value in the sense that the above argument of Case 2 works for the equation $u''' = -1 + 1/u^p$ if and only if $p \geq 2$ (cf. [3]).

LEMMA 4.6. *The function $u - 1$ has infinitely many zeros near $+\infty$.*

Proof. By linearization theory we know that $u - 1$ has zeros. Hence it is enough to prove that

$$u(a) = 1 \quad \Longrightarrow \exists b > a \quad \text{such that} \quad u(b) = 1.$$

Assume (for contradiction) that $u > 1$ in $(a, +\infty)$. Then $u''' < 0$ and the limits $u''(+\infty)$, $u'(+\infty)$, and $u(+\infty)$ exist. This and the differential equation (4.1) imply that

$$u'''(+\infty) = 0\,, \ u(+\infty) = 1\,, \ u'(+\infty) = 0, \ \text{and} \quad u''(+\infty) = 0.$$

Hence u'' is decreasing and positive, u' increasing and negative, and $u - 1$ decreasing and positive in $(a, +\infty)$. This contradicts $u(a) = 1$.

A similar argument excludes the possibility that $u < 1$ on (a, ∞) and completes the proof of Lemma 4.6.

LEMMA 4.7. *As $n \to +\infty$,*

$$u(b_{2n}) \to 0 \quad \textit{and} \quad u(b_{2n+1}) \to \infty.$$

Proof. By Lemma 4.3 we know that these limits exist. On the other hand, Lemma 4.1 states that Φ_4 is increasing and convex, and hence

$$\Phi_4(x) \to \infty \quad \text{as} \quad x \to +\infty. \tag{4.5}$$

Letting $x \to +\infty$ along the sequence $\{b_n\}$, we obtain that

$$u(b_n) + \frac{1}{u(b_n)} \to \infty \quad \text{as} \quad n \to +\infty,$$

from which the desired conclusions follow.

LEMMA 4.8. *The sequences $|u'(a_n)|$ and $|u''(a_n)|$ are increasing and tend to infinity as $n \to +\infty$.*

Proof. The sequences are increasing because the functions Φ_1 and Φ_3 are increasing by Lemma 4.1. Notice that $u'''(a_n) = 0$. By (4.5)

$$|u'(a_n)\, u''(a_n)| \to \infty \quad \text{as} \quad n \to +\infty. \tag{4.6}$$

Let us show that each of the factors also tends to infinity. Since the functions Φ_i of Lemma 4.1 are increasing, we may compute their limits as $x \to +\infty$ along different sequences. Considering Φ_1 and using Lemma 4.7, we obtain that as $n \to +\infty$

$$\frac{1}{2}\lim\{u'(a_n)\}^2 = \lim u''(b_{2n}),$$

while considering Φ_3 it follows that

$$\lim u''(b_{2n}) = \lim |u''(b_n)| = \lim |u''(a_n)|.$$

The last two relations and (4.6) conclude the proof.

Remark. It follows that as $x \to +\infty$, the four functions Φ_i of Lemma 4.1 tend to infinity so that the sequences $|u'(c_n)|$ and $|u''(b_n)|$ are also increasing and tend to infinity as $n \to +\infty$.

LEMMA 4.9. *We have*

$$a_{2n+2} - a_{2n+1} \to \infty \quad \textit{as} \quad n \to +\infty.$$

Proof. This limit follows at once from (4.4) and Lemma 4.7.

LEMMA 4.10. *We have*

$$a_{2n+1} - a_{2n} \to 0 \quad \textit{as} \quad n \to +\infty.$$

The proof follows from the inequality

$$1 - u(b_{2n}) > \frac{1}{4}|u'(a_{2n})|(a_{2n+1} - a_{2n}) \tag{4.7}$$

and Lemmas 4.7 and 4.8, according to which $u(b_{2n}) \to 0$, $|u'(a_{2n})| \to \infty$ as $n \to \infty$.

Proof of (4.7). We are going to perform a comparison argument in the intervals where $u < 1$. To simplify the notation we set

$$h = a_{2n+1} - a_{2n}, \quad b = b_{2n}, \quad \alpha = -u'(a_{2n}) > 0$$

and translate the origin to a_{2n}. In this setting we have that

$$u(0) = 1, \qquad u(h) = 1, \qquad u < 1 \quad \text{in} \quad (0, h),$$

and (4.7) takes the form

$$1 - u(b) > \frac{1}{4}\alpha h.$$

We define a polynomial of second degree P such that

$$P(0) = 0, \qquad P(h) = 0, \qquad P'(0) = \alpha.$$

Hence

$$P(x) = \alpha x\left(1 - \frac{x}{h}\right).$$

Next we consider the function

$$Q = 1 - u - P.$$

Since

$$Q(0) = Q(h) = Q'(0) = 0 \qquad \text{and} \qquad Q''' = -u''' < 0 \quad \text{in} \quad (0, h),$$

it readily follows that $Q > 0$ in $(0, h)$ and, in particular,

$$1 - u(b) > P(h/2) = \frac{1}{4}\alpha h.$$

This completes the proof of (4.7).

5. Properties A–F of §1. The lemmas and theorems of §4 prove all the properties of solutions of problem (II) stated in §1 as follows:

Property **A** is Theorem 4.5.

Property **B** includes Lemma 4.6; the remainder follows from a standard linearization analysis near $x = -\infty$.

Properties **C** and **D** are Lemmas 4.2, 4.3, and 4.7.

Property **E** is contained in Lemmas 4.9 and 4.10.

Property **F** is dealt with in Lemma 4.8 and its associated remark.

REFERENCES

[1] M. P. Brenner and A. L. Bertozzi, *Spreading of droplets on a solid surface*, Phys. Rev. Lett., 71 (1993), pp. 593–596.

[2] A. L. Bertozzi, M. P. Brenner, T. F. Dupont, and L. P. Kadanoff, *Singularities and similarities in interface flows*, in Trends and Perspectives in Applied Mathematics, L. Sirovich, ed., Appl. Math. Sci. Vol. 100, Springer-Verlag, Berlin, 1994.

[3] E. Beretta, J. Hulshof, and L. A. Peletier, *On an ODE from forced coating flow*, to appear.

[4] S. Boatto, L. P. Kadanoff, and P. Olla, *Traveling-wave solutions to thin-film equations*, Phys. Rev. E, 48 (1993), pp. 4423–4431.

[5] C. M. Bender and S. A. Orszag, *Advanced mathematical methods for scientists and engineers*, McGraw–Hill, New York, 1978, pp. 155–158.

[6] F. Bernis, L. A. Peletier, and S. M. Williams, *Source type solutions of a fourth order nonlinear degenerate parabolic equation*, Nonlinear Anal., 18 (1992), pp. 217–234.

[7] J. A. Moriarty and L. W. Schwartz, *Effective slip in numerical calculations of moving-contact-line problems*, J. Engrg. Math., 26 (1992), pp. 81–86.

[8] J. A. Moriarty, L. W. Schwartz, and E. O. Tuck, *Unsteady spreading of thin liquid films with small surface tension*, Phys. Fluids. A, 3 (1991), pp. 733–742.

[9] L. H. Tanner, *The spreading of silicone oil drops on horizontal surfaces*, J. Phys. D: Appl. Phys., 12 (1979), pp. 1473–1484.

[10] S. Taliaferro, *On the positive solutions of $y'' + \phi(t)y^{-\lambda} = 0$*, Nonlinear Anal., 2 (1978), pp. 437–446.

[11] W. C. Troy, *Solutions of third-order differential equations relevant to draining and coating flows*, SIAM J. Math. Anal., 24 (1993), pp. 155–171.

[12] E. O. Tuck and L. W. Schwartz, *A numerical and asymptotic study of some third-order ordinary differential equations relevant to draining and coating flows*, SIAM Rev., 32 (1990), pp. 453–469.

SIAM J. MATH. ANAL.
Vol. 27, No. 2, pp. 528–542, March 1996

011

GLOBAL BIFURCATION OF AN ELASTIC CONDUCTING ROD IN A MAGNETIC FIELD*

PETER WOLFE †

Abstract. We study the equilibrium states of a nonlinearly elastic conducting rod in a magnetic field, a problem we have considered in several previous papers. We are now able to prove a global bifurcation theorem for this problem. To do this, two difficulties must be overcome. The first is the presence of the rotation group $SO(2)$ as a symmetry group for the problem. The second is that, for some values of certain parameters, the linearized problem is a nonstandard eigenvalue problem. The former difficulty is overcome by applying an idea due to Healey, who observed the existence of an additional symmetry in a related problem first posed by the present author. The latter problem is handled by using some nonstandard tools from functional analysis.

Key words. Cosserat rods, eigenvalues, global bifurcation, nonlinear elasticity, symmetry

AMS subject classifications. 34A47, 34B15, 34L05, 47B50, 58E05, 73C50, 73K05

1. Introduction. In this paper we contine our study, begun in [6] and [7], of the equilibrium states of a nonlinearly elastic conducting rod in a magnetic field. The rod is assumed to be welded to fixed supports. The magnetic field is assumed to be constant and directed parallel to the line between the supports. The rod can undergo flexure, tension, shear, and extension. The elastic properties of the rod are embodied in the the constitutive functions which relate the strains which measure the above quantities to the contact force and contact couple. We assume the rod is homogeneous and transversely isotropic. We will also assume that the rod is hyperelastic so that the constitutive functions are derived from a strain energy function. This assumption was essential in previous work but is less so here. However it is convenient to retain this assumption. The fundamental parameter is $\lambda = IB$, where I is the current in the rod and B is the strength of the magnetic field. For all real λ there exists a trivial state in which the rod is straight and untwisted. We are interested in the existence of nontrivial solutions. This problem can be posed as a boundary value problem for a system of nonlinear ordinary differential equations. In order to study bifurcation from the trivial solution (corresponding to the trivial state), we must study the linear eigenvalue problem obtained by linearization about the trivial solution. It was at this point in our previous work that we ran into an obstruction. In order to apply the standard results of bifurcation theory [2], it is necessary that the eigenspace corresponding to an eigenvalue be one dimensional. However, in our case this condition cannot hold since the problem admits $SO(2)$ as a symmetry group. We are now able to surmount this obstruction. For this we use an idea of Healey [5]. In [5] Healey considered a problem previously posed by the present author [8], that of a rotating conducting wire in a magnetic field. Healey noticed that in addition to the $SO(2)$ symmetry, that problem also possesses a Z_2 symmetry which he called a "subtle symmetry." This observation enabled him to reduce the problem to one in which the eigenvalues of the linearized sytem are simple and thus amenable to standard global bifurcation analysis. We show

*Received by the editors February 15, 1994; accepted for publication (in revised form) July 25, 1994.

†Department of Mathematics, University of Maryland, College Park, MD 20742 (pnw@math.umd.edu).

that the same symmetry is present in our problem. Thus we are able to obtain a global bifurcation result in this case.

In the study of the linearized system in [7], we assumed that the parameters appearing in the linearized eigenvalue problem satisfy an inequality, which enabled us to use standard techniques to prove the existence of (real) eigenvalues. Here we show that, by using some nonstandard tools from functional analysis, we can mostly dispense with this condition (which from a physical point of view is completely artificial).

In §2 we will formulate the problem. In §3 we will outline the theory of global bifurcation in the presence of symmetry. In §4 we will discuss the symmetries of the problem. In §5 we will deal with the linearized problem, while in §6 we present our global bifurcation result.

Vectors (elements of three-dimensional Euclidean space E^3) will be denoted by $\mathbf{r}, \mathbf{d_p}$, etc., while elements of $\mathbf{R}^n$ (n-tuples of real numbers) will be denoted by $\boldsymbol{u}$,$\boldsymbol{v}$, etc. The summation convention will be used throughout. An equation containing an index which is not repeated (and therefore not summed) is assumed to hold for the values 1, 2, 3 of that index.

2. Formulation of the problem. Our model for the rod is the *special Cosserat theory.* In this theory, the configuration of the rod is specified by a position vector $\mathbf{r}$ and an orthonormal pair of vector functions $\mathbf{d_1}$ and $\mathbf{d_2}$ of the real variable $s \in [0,1]$. We interpret s as a scaled arclength parameter of the line of centroids of the rod (a slender three-dimensional body) in a reference configuration, so s identifies material sections of the rod. Thus $\mathbf{r(s)}$ is the position in a deformed configuration of the material point at the centroid of the section s. The vectors $\mathbf{d_1(s)}$ and $\mathbf{d_2(s)}$, which determine a plane in space and a line in that plane, characterize the deformed configuration of the section s. We set

$$\mathbf{d_3} = \mathbf{d_1} \times \mathbf{d_2} \tag{2.1}$$

to obtain an orthonormal triple, called the *directors.* Let derivatives with respect to s be denoted by primes. Since $\{\mathbf{d_p}\}$ is orthonormal there exists a vector function $\mathbf{u}$ such that

$$\mathbf{d_p'} = \mathbf{u} \times \mathbf{d_p}. \tag{2.2}$$

Conversely, given an artbitrary vector function $\mathbf{u}$, any solution $\{\mathbf{d_p}\}$ of (2.2) is orthonormal if $\{\mathbf{d_p(0)}\}$ is. The components of $\mathbf{u}$ with respect to $\{\mathbf{d_p}\}$ are

$$u_p = \frac{1}{2} e_{pqr} \mathbf{d_q'} \cdot \mathbf{d_r}, \tag{2.3}$$

where $\{e_{pqr}\}$ are components of the alternating symbol. We decompose the vector $\mathbf{r}'$ into its components with respect to the basis $\{\mathbf{d_p}\}$ by

$$\mathbf{r}' = v_p \mathbf{d_p} \tag{2.4}$$

and set

$$\boldsymbol{u} = (u_1, u_2, u_3)^T, \quad \boldsymbol{v} = (v_1, v_2, v_3)^T. \tag{2.5}$$

The triples $\boldsymbol{u}$ and $\boldsymbol{v}$ are the *strains* of the theory. They determine the configuration of the rod uniquely to within a rigid body motion by integration of (2.2) and (2.4).

The physical interpretation of these quantities is as follows: v_1 and v_2 measure the amount of shear, $v_3 = \mathbf{r}' \cdot (\mathbf{d_1} \times \mathbf{d_2})$ measures volume change, $(v_k v_k)^{1/2}$ measures axial stretch, u_1 and u_2 measure *flexure*, and u_3 measures *twist*.

Let $\{\mathbf{i_1} = \mathbf{i}, \mathbf{i_2} = \mathbf{j}, \mathbf{i_3} = \mathbf{k}\}$ represent the standard orthonormal basis for E^3. Let

$$d_{ij} = \mathbf{d_j} \cdot \mathbf{i_i}. \tag{2.6}$$

Then we may consider $D = (d_{ij})$ as an element of $SO(3)$, the group of 3×3 orthogonal matrices with determinant 1. We can then write (2.2) as

$$D' = DU, \tag{2.7}$$

where

$$U = \begin{pmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix}. \tag{2.8}$$

We write

$$\mathbf{r} = x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k} = x_i\mathbf{i_i}. \tag{2.9}$$

Then if we set

$$\boldsymbol{r} = (x_1, x_2, x_3)^T, \tag{2.10}$$

we may rewrite (2.4) as

$$\boldsymbol{r}' = D\boldsymbol{v}. \tag{2.11}$$

For later developments it (unfortunately) becomes necessary to introduce Euler angles in order to parameterize D. The ones we introduce here are slightly different from those of [7]. Thus we set

$$D = \begin{pmatrix} c(\psi)c(\phi) - s(\psi)s(\theta)s(\phi) & -s(\psi)c(\theta) & c(\psi)s(\phi) + s(\psi)s(\theta)c(\phi) \\ s(\psi)c(\phi) + c(\psi)s(\theta)s(\phi) & c(\psi)c(\theta) & s(\psi)s(\phi) - c(\psi)s(\theta)c(\phi) \\ -c(\theta)s(\phi) & s(\theta) & c(\phi)c(\theta) \end{pmatrix}, \tag{2.12}$$

where $c(\psi) = \cos\psi$, $s(\psi) = \sin\psi$, etc. These angles are such that $\theta = \phi = \psi = 0$ corresponds to $D = I$. Also, the polar singularities occur at $\theta = (n + \frac{1}{2})\pi$. From (2.7) and (2.12),

$$\begin{aligned} u_1 &= \theta' \cos\phi - \psi' \cos\theta \sin\phi, \\ u_2 &= \phi' + \psi' \sin\theta, \\ u_3 &= \theta' \sin\phi + \psi' \cos\theta \cos\phi, \end{aligned} \tag{2.13}$$

while $\boldsymbol{v}$ can be written as a function of $\boldsymbol{r}'$ and (θ, ϕ, ψ) by rewriting (2.11) as

$$\boldsymbol{v} = D^T \boldsymbol{r}' \tag{2.14}$$

and using (2.12).

The set of admissible strains is restricted by the inequality

$$v_3 > H(u_1, u_2). \tag{2.15}$$

The inequality (2.15) is the rod-theoretic analogue of the requirement that the Jacobian of the transformation in three-dimensional kinematics be positive. Here H is a given function which depends on the cross-sectional shape. We require that H satisfies

$$\begin{gathered} H(0,0) = 0, \\ H(u_1, u_2) > 0 \text{ for } u_1^2 + u_2^2 > 0, \\ H \text{ is homogeneous of degree 1.} \end{gathered} \tag{2.16}$$

If the cross sections of the rod are disks of radius h, then

$$H(u_1, u_2) = h\sqrt{u_1^2 + u_2^2}.$$

For $s \in (0,1)$, let

$$\mathbf{n(s)} = n_k(s)\mathbf{d_k(s)} \tag{2.17}$$

be the contact force and

$$\mathbf{m(s)} = m_k(s)\mathbf{d_k(s)} \tag{2.18}$$

be the contact couple exerted by the material of $[0, s)$ on the material of $[s, 1]$ in the configuration $\{\mathbf{r}, \mathbf{d_k}\}$. We let

$$\int_s^1 \mathbf{f(t)}\, dt$$

be the resultant of all other forces exerted on the material of $[s, 1]$ in this configuration. We assume that the resultant of all other couples exerted on the material of $[s, 1]$ in this configuration is $\mathbf{0}$. Then the classical equilibrium equations for forces and moments are

$$\mathbf{n}' + \mathbf{f} = \mathbf{0}, \tag{2.19}$$

$$\mathbf{m}' + \mathbf{r}' \times \mathbf{n} = \mathbf{0}. \tag{2.20}$$

In this paper we assume that the rod is a conductor carrying a current I. There is also a constant magnetic field $\mathbf{B} = B\mathbf{k}$ present. (Recall that $\mathbf{k} = \mathbf{i_3}$.) The force on the rod is then given by

$$\mathbf{f(s)} = I\mathbf{r}'\mathbf{(s)} \times B\mathbf{k} = \lambda\mathbf{r}'\mathbf{(s)} \times \mathbf{k}, \tag{2.21}$$

where

$$\lambda = IB. \tag{2.22}$$

The elastic properties of the rod are embodied in the constitutive equations relating the stresses $\boldsymbol{m} = (m_1, m_2, m_3)^T$ and $\boldsymbol{n} = (n_1, n_2, n_3)^T$ to the strains $\boldsymbol{u}$ and

$\boldsymbol{v}$. Thus we assume that twice continuously differentiable functions $\widehat{\boldsymbol{m}}$ and $\widehat{\boldsymbol{n}}$ taking values in $\mathbf{R}^3$ are defined on the set

$$S = \{(\boldsymbol{u}, \boldsymbol{v}) \in \mathbf{R}^3 \times \mathbf{R}^3 : v_3 > H(u_1, u_2)\} \tag{2.23}$$

so that

$$\boldsymbol{m}(s) = \widehat{\boldsymbol{m}}(\boldsymbol{u}(s), \boldsymbol{v}(s)), \quad \boldsymbol{n}(s) = \widehat{\boldsymbol{n}}(\boldsymbol{u}(s), \boldsymbol{v}(s)). \tag{2.24}$$

We further assume that the rod is *hyperelastic*, meaning that there exists a *stored-energy function* $\Phi : S \to \mathbf{R}$ such that

$$\widehat{\boldsymbol{m}}(\boldsymbol{u}, \boldsymbol{v}) = \frac{\partial \Phi(\boldsymbol{u}, \boldsymbol{v})}{\partial \boldsymbol{u}}, \quad \widehat{\boldsymbol{n}}(\boldsymbol{u}, \boldsymbol{v}) = \frac{\partial \Phi(\boldsymbol{u}, \boldsymbol{v})}{\partial \boldsymbol{v}}. \tag{2.25}$$

We assume that Φ is three times continuously differentiable and convex. We assume that $\widehat{\boldsymbol{m}}$ and $\widehat{\boldsymbol{n}}$ tend to infinity as $|\boldsymbol{u}| + |\boldsymbol{v}| \to \infty$ or as $v_3 - H(u_1, u_2) \to 0$.

We also assume that the material of the rod is *transversely isotropic*, i.e.,

$$\widehat{\boldsymbol{m}}(Q\boldsymbol{u}, Q\boldsymbol{v}) = Q\widehat{\boldsymbol{m}}(\boldsymbol{u}, \boldsymbol{v}), \quad \widehat{\boldsymbol{n}}(Q\boldsymbol{u}, Q\boldsymbol{v}) = Q\widehat{\boldsymbol{n}}(\boldsymbol{u}, \boldsymbol{v}) \tag{2.26}$$

for each orthogonal matrix Q of the form

$$Q = \begin{pmatrix} Q_{11} & Q_{12} & 0 \\ Q_{21} & Q_{22} & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.27}$$

Finally, we assume

$$\widehat{\boldsymbol{m}}(\mathbf{0}, \boldsymbol{v}) = \mathbf{0}, \quad \hat{n}_\alpha = 0 \text{ if } v_\alpha = 0, \ \alpha = 1, 2. \tag{2.28}$$

It follows from these assumptions [1] that $\hat{m}_3$ and $\hat{n}_3$ depend on

$$u_1^2 + u_2^2, \ v_1^2 + v_2^2, \ u_1 v_1 + u_2 v_2, \ u_3, \ v_3 \tag{2.29}$$

and that $\{\hat{m}_\beta, \hat{n}_\beta, \beta = 1, 2\}$ have the form

$$\hat{m}_\beta = \hat{\sigma} u_\beta, \quad \hat{n}_\beta = \hat{\tau} v_\beta, \tag{2.30}$$

where $\hat{\sigma}$ and $\hat{\tau}$ depend on the arguments listed in (2.29). The convexity of Φ together with (2.30) implies that

$$\hat{\sigma} > 0, \quad \hat{\tau} > 0. \tag{2.31}$$

We assume that the rod is welded to fixed supports at $\mathbf{0}$ and $b\mathbf{k}$, where $b > 1$. Thus the boundary conditions are

$$\mathbf{r}(\mathbf{0}) = \mathbf{0}, \quad \mathbf{r}(\mathbf{1}) = b\mathbf{k}, \tag{2.32a}$$

$$\mathbf{d}_\mathbf{k}(\mathbf{0}) = \mathbf{i}_\mathbf{k}, \quad \mathbf{d}_\mathbf{k}(\mathbf{1}) = \mathbf{i}_\mathbf{k}. \tag{2.32b}$$

It terms of the Euler angles, (2.32b) becomes

$$\theta(0) = \phi(0) = \psi(0) = 0, \ \theta(1) = \phi(1) = \psi(1) = 0. \tag{2.32c}$$

In order to convert the equilibrium equations (2.19) and (2.20) to a form which is amenable to further analysis, we take the dot product of these equations with $\mathbf{d}_i$ and use (2.2). The result is

$$n_1' - u_3n_2 + u_2n_3 + \lambda(x_2'd_{11} - x_1'd_{21}) = 0, \tag{2.33a}$$

$$n_2' + u_3n_1 - u_1n_3 + \lambda(x_2'd_{12} - x_1'd_{22}) = 0, \tag{2.33b}$$

$$n_3' - u_2n_1 + u_1n_2 + \lambda(x_2'd_{13} - x_1'd_{23}) = 0, \tag{2.33c}$$

$$m_1' - u_3m_2 + u_2m_3 + v_2n_3 - v_3n_2 = 0, \tag{2.33d}$$

$$m_2' + u_3m_1 - u_1m_3 + v_3n_1 - v_1n_3 = 0, \tag{2.33e}$$

$$m_3' - u_2m_1 + u_1m_2 + v_1n_2 - v_2n_1 = 0. \tag{2.33f}$$

In (2.33) the arguments of $\boldsymbol{m}$ and $\boldsymbol{n}$ are $\boldsymbol{u}$ and $\boldsymbol{v}$. The convexity of Φ implies that the symmetric matrix

$$\begin{pmatrix} \partial m_i/\partial u_j & \partial m_i/\partial v_j \\ \partial n_i/\partial u_j & \partial n_i/\partial v_j \end{pmatrix} \tag{2.34}$$

is positive definite. Thus equation (2.33) can be solved for $\boldsymbol{u}'$ and $\boldsymbol{v}'$. Furthermore, as long as $|\theta| < \pi/2$, we can differentiate (2.13) and (2.14) and solve these equations for the second derivatives of $x_1, x_2, x_3, \theta, \phi, \psi$. Thus if we let

$$\boldsymbol{X} = (x_1, x_2, z, \theta, \phi, \psi)^T, \tag{2.35}$$

where

$$z(s) = x_3(s) - bs, \tag{2.36}$$

we arrive at an equation of the form

$$-\boldsymbol{X}'' = \boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}', \lambda), \tag{2.37}$$

which is equivalent to (2.33). The boundary conditions (2.32) can be restated as

$$\boldsymbol{X}(0) = \boldsymbol{X}(1) = \mathbf{0}. \tag{2.38}$$

Thus our boundary value problem consists of (2.33), (2.13), and (2.14) along with the constitutive equations (2.24) and the boundary conditions (2.32) or, alternatively, (2.37) with the boundary condition (2.38).

For every real value of λ, the problem admits a *trivial solution* in which the rod is straight and untwisted. In this solution $\boldsymbol{u} = \mathbf{0}, v_1 = v_2 = 0, v_3 = b, \mathbf{r} = bs\mathbf{k}, \theta = \phi = \psi = 0, \boldsymbol{m} = \mathbf{0}, n_1 = n_2 = 0$, and

$$n_3 = \hat{n}_3(\mathbf{0}, (0,0,b)^T) \equiv n_0, \tag{2.39}$$

where

$$n_0 > 0. \tag{2.40}$$

Inequality (2.40) is consistent with the assumption that the rod is in tension when no force is applied, that is, the assumption $b > 1$. The object of this paper is to study the existence of nontrivial solutions.

3. The global bifurcation theorem. In this section we state the global bifurcation theorem which we will use to obtain our results. We follow the treatment in [4]. For further references, please refer to that paper.

Let B be a real Banach space, Ω an open connected subset of $B \times R$ such that $(0, \lambda) \in \Omega$ for all $\lambda \in R$ and $f : \Omega \to B$ be m times Fréchet differentiable. We consider the problem

$$f(x, \lambda) = 0. \tag{3.1}$$

We wish to determine Σ, the solution set of (3.1).

We assume that (3.1) models a system characterized by a symmetry group $\mathcal{G}$. In particular, we assume that (3.1) is *equivariant* under a specific representation T of $\mathcal{G}$ on B, i.e.,

$$f(T_g x, \lambda) = T_g f(x, \lambda), \quad \forall g \in \mathcal{G}, \tag{3.2}$$

where it is assumed that if $(x, \lambda) \in \Omega$ then $(T_g x, \lambda) \in \Omega$ for all $g \in \mathcal{G}$. Let $\mathcal{H}$ be a (not necessarily proper) subgroup of $\mathcal{G}$. We define the *$\mathcal{H}$-fixed point set $B_{\mathcal{H}}$* as

$$B_{\mathcal{H}} = \{u \in B : T_g u = u \ \ \forall g \in \mathcal{H}\}.$$

If $\mathcal{H}$ is a compact group, $B_{\mathcal{H}}$ is a Banach space. We let $\Omega_{\mathcal{H}} = \Omega \cap (B_{\mathcal{H}} \times R)$. It then follows from (3.1) and (3.2) that if $(u, \lambda) \in \Omega_{\mathcal{H}}$, $g \in \mathcal{H}$, then

$$T_g f(u, \lambda) = f(T_g u, \lambda) = f(u, \lambda). \tag{3.3}$$

Thus $f(u, \lambda) \in B_{\mathcal{H}}$ for all $(u, \lambda) \in \Omega_{\mathcal{H}}$ and $f : \Omega_{\mathcal{H}} \to B_{\mathcal{H}}$. So we see that a point $(x_0, \lambda) \in \Omega_{\mathcal{H}}$ is a solution of (3.1) if and only if it is a solution of the *$\mathcal{H}$-reduced problem*

$$f_{\mathcal{H}}(u, \lambda) = 0, \tag{3.4}$$

where $f_{\mathcal{H}} = f|_{\Omega_{\mathcal{H}}}$. The solution set of (3.4), denoted by $\Sigma_{\mathcal{H}}$, is called the *$\mathcal{H}$-solution set.*

We assume that $f_{\mathcal{H}}$ has the form

$$f_{\mathcal{H}} \equiv u - c_{\mathcal{H}}(u, \lambda), \tag{3.5}$$

where $c_{\mathcal{H}} : \Omega_{\mathcal{H}} \to B_{\mathcal{H}}$ is completely continuous. Suppose

$$f(0, \lambda) = 0, \quad \forall \lambda \in R. \tag{3.6}$$

The set $\Sigma_t \equiv (\{0\} \times R)$ is called the trivial solution branch of (3.1). A solution $(0, \lambda_0) \in \Sigma_t$ is said to be a *bifurcation point* of (3.1) if every neighborhood of $(0, \lambda_0)$ contains solution pairs $(u_*, \lambda_*) \in \Sigma$ with $u_* \neq 0$. Define $L(\lambda) \equiv D_1 f(0, \lambda)$, the Fréchet derivative of f with respect to u at $(0, \lambda)$. A necessary condition for $(0, \lambda_0)$ to be a bifurcation point is that $L(\lambda_0) : B \to B$ be noninvertible.

For any subgroup $\mathcal{H} \subset \mathcal{G}$, since $0 \in B_{\mathcal{H}}$, we have

$$f_{\mathcal{H}}(0, \lambda) = 0, \ \forall \lambda \in R. \tag{3.7}$$

Suppose $L(\lambda_0)$ is singular, $y \in \mathcal{N}(L(\lambda_0))$ is such that the *isotropy subgroup* of $\mathcal{G}$ at y

$$\mathcal{H} = \{g \in \mathcal{G} \mid T_g y = y\} \tag{3.8}$$

is proper. It then follows that $L_{\mathcal{H}}(\lambda_0) = D_1 f_{\mathcal{H}}(0, \lambda_0) = L(\lambda_0)|_{B_{\mathcal{H}}}$ is singular. We then have the *equivariant bifurcation theorem.*

THEOREM 3.1 [4]. *Suppose that $f \in C^2$ and there exists a vector $y \in \mathcal{N}(L(\lambda_0))$ which defines a proper isotropy subgroup $\mathcal{H}$. Assume*

(i) $\dim \mathcal{N}(L_{\mathcal{H}}(\lambda_0))$ *is odd, and*

(ii) $L'_{\mathcal{H}}(\lambda_0)v \notin \mathcal{R}(L_{\mathcal{H}}(\lambda_0))\ \forall v \in (L_{\mathcal{H}}(\lambda_0))\backslash\{0\}$,

where prime denotes differentiation with respect to λ. Then $(0, \lambda_0)$ is a bifurcation point of (3.1) such that in every sufficiently small neighborhood of $(0, \lambda_0)$ there are nontrivial solutions $(u_*, \lambda_*) \in \Sigma_{\mathcal{H}}$. In particular, if $\dim \mathcal{N}(L_{\mathcal{H}}(\lambda_0)) = 1$, there exists a unique, local bifurcating branch of solutions of the form $\sigma \to (\hat{u}(\sigma), \hat{\lambda}(\sigma))$. Moreover there exists a connected subset $\zeta_{\mathcal{H}} \subseteq \Sigma_{\mathcal{H}} \cap \overline{\Sigma\backslash\Sigma_t}$ containing $(0, \lambda)$ that is characterized by at least one of the following properties:

(a) $\zeta_{\mathcal{H}}$ *is unbounded in $B \times R$.*

(b) $\overline{\zeta_{\mathcal{H}}} \cap \partial\Omega \neq \emptyset$

(c) *There exists a pair $(0, \lambda_*) \in \zeta_{\mathcal{H}} \cap \Sigma_t$ with $\lambda_* \neq \lambda_0$.*

The set $\zeta_{\mathcal{H}}$ is called a global $\mathcal{H}$-symmetric bifurcating branch *of* (3.1) *through* $(0, \lambda_0)$.

The rest of the paper consists of verifying that we can apply Theorem (3.1) to the problem described in §2.

4. Symmetries of the problem. In this section we examine the symmetries of the problem. There are two types of symmetry, the "obvious" rotational symmetry and the "subtle" reflection symmetry. It is the latter which we will use to obtain our results.

We first consider the action of the group $SO(2)$ on the problem. The action of an element of this group $T_\alpha, \alpha \in R(\text{mod}2\pi)$ is defined by the matrix

$$Q_\alpha = \begin{pmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{4.1}$$

T_α acts on the vector $\mathbf{r}$ by counterclockwise rotation through the angle α about the x_3 axis. If $\boldsymbol{r}$ is given by (2.10), the components of $T_\alpha \mathbf{r}$ are given by $Q_\alpha \boldsymbol{r}$. Similarly, T_α acts on the triples $\boldsymbol{u}$ and $\boldsymbol{v}$ by

$$(4.2) \qquad T_\alpha(\boldsymbol{u}, \boldsymbol{v}) = (Q_\alpha \boldsymbol{u}, Q_\alpha \boldsymbol{v}).$$

The action of T_α on the matrices D and U is

$$T_\alpha D = Q_\alpha D Q_\alpha^T, \quad T_\alpha U = Q_\alpha U Q_\alpha^T. \tag{4.3}$$

(The second equation of (4.3) is equivalent to $T_\alpha \boldsymbol{u} = Q_\alpha \boldsymbol{u}$.) By (4.2), (2.24), and (2.26) it follows that

$$T_\alpha(\boldsymbol{m}, \boldsymbol{n}) = (Q_\alpha \boldsymbol{m}, Q_\alpha \boldsymbol{n}). \tag{4.4}$$

It is easily checked that if the configuration $\{\mathbf{r}, D\}$ is a solution of our problem (in the form (2.33), (2.32a,b)), so is $\{T_\alpha \mathbf{r}, T_\alpha D\}$. It does not appear that the Euler angles transform in a nice (i.e., geometric) way under the transformation T_α. Since we do not intend to use the rotational invarience in our global bifurcation theorem, we shall not identify a mapping f which is equivarient under $SO(2)$. Of course, as has been noted in [7], this $SO(2)$ symmetry is a source of difficulty when one attempts to apply standard bifurcation theory since it forces the kernel of the linearized operator (at a singular point) to have an even dimension. However, as in [5], this problem also admits a "subtle symmetry" (Healey's term), which we now describe. This symmetry corresponds to a 180° rotation of the rod about the line $x_2 = 0$, $x_3 = b/2$. If we denote this mapping by R (so that $R^2 = I$), we have

$$R(x_1, x_2, x_3)(s) = (x_1, -x_2, b - x_3)(1 - s). \tag{4.5}$$

Let

$$E = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{4.6}$$

Then R acts on the triples $\boldsymbol{u}$ and $\boldsymbol{v}$ by

$$R(\boldsymbol{u}, \boldsymbol{v})(s) = (E\boldsymbol{u}, E\boldsymbol{v})(1 - s). \tag{4.7}$$

The action of R on the matrices D and U is given by

$$RD(s) = EDE(1 - s), \quad RU(s) = -EUE(1 - s). \tag{4.8}$$

As regards the Euler angles, if we let $\Psi = (\theta, \phi, \psi)^T$, the action of R on Ψ is given by

$$R\Psi(s) = -E\Psi(1 - s). \tag{4.9}$$

Again, by (4.2), (2.24), and (2.26) with $Q = E$,

$$R(\boldsymbol{m}, \boldsymbol{n})(s) = (E\boldsymbol{m}, E\boldsymbol{n})(1 - s). \tag{4.10}$$

We are now ready to cast our problem into the form suitable for applying Theorem 3.1.

We let $C_0^1 = C_0^1[0, 1]$ be the space of continuously differentable functions vanishing at $s = 0$ and $s = 1$. For $x \in C_0^1[0, 1]$, we define

$$\|x\|_{C^1} = \max_{0 \le s \le 1} |x(s)| + \max_{0 \le s \le 1} |x'(s)|.$$

We let $B = (C_0^1[0, 1])^6$. For $\boldsymbol{X} = (\xi_1, \ldots, \xi_6)^T \in B$, we define

$$\|\boldsymbol{X}\|_B = \sum_{i=1}^{6} \|\xi_i\|_{C^1}. \tag{4.11}$$

We identify our state variables with $\boldsymbol{X} \in B$ by (2.35),(2.36). Associated with $\boldsymbol{X} \in B$ are the variables D defined by (2.12), $\boldsymbol{u}$ by (2.13), and $\boldsymbol{v}$ by (2.14). Note that when $\boldsymbol{X} = \boldsymbol{0}$, $\boldsymbol{v} = (0, 0, b)^T$. The set Ω is defined as

$$\Omega = \{(\boldsymbol{X}, \lambda) : v_3 > h(u_1, u_2), |\theta| < \pi/2, \lambda \in R\}. \tag{4.12}$$

For the mapping $\boldsymbol{f}(\boldsymbol{X}, \lambda) : \Omega \to B$, we let $K(s, t)$ be the Green's function for the problem $-x''(s) = g(s)$, $0 \le s \le 1$, $x(0) = x(1) = 0$. Then (2.37), (2.38) is equivalent to

$$\boldsymbol{f}(\boldsymbol{X}, \lambda)(s) \equiv \boldsymbol{X}(s) - \int_0^1 K(s, t)\boldsymbol{F}(\boldsymbol{X}(t), \boldsymbol{X}'(t), \lambda)\, dt = \boldsymbol{0}. \tag{4.13}$$

Clearly, $\boldsymbol{f}$ is of the form (3.5). Thus we are now ready to apply the theory of §3.

For $\boldsymbol{X} \in B$, we define $R\boldsymbol{X}$ by

$$(R\boldsymbol{X})(s) = (x_1, -x_2, -z, \theta, -\phi, -\psi)^T(1-s). \tag{4.14}$$

Remark. The transformation law for z given in (4.14) is consistent with (4.5) since $RZ(s) = R(x_3(s) - bs) = Rx_3(s) - R(bs) = b - x_3(1-s) - (b - b(1-s)) = b(1-s) - x_3(1-s) = -z(1-s)$.

We also have

$$\frac{d}{ds}(R\boldsymbol{X}))(s) = (-x_1', x_2', z', -\theta', \phi', \psi')^T(1-s) \tag{4.15}$$

THEOREM 4.1. *The mapping* (4.13) *is equivarient under the representation of* Z_2 *defined by* R, *i.e.,*

$$\boldsymbol{f}(R\boldsymbol{X}, \lambda) = R\boldsymbol{f}(\boldsymbol{X}, \lambda). \tag{4.16}$$

Proof. Observe that if the left sides of (2.33) are formed into a 6-vector (in the order in which they appear), this 6-vector is equivarient under the mapping (4.14). It follows that this equivariance is preserved in the transformation to (2.37). If we write (2.37) as

$$\boldsymbol{X}'' + \boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}', \lambda) = \mathbf{0}, \tag{4.17}$$

the transformed left-hand side of (4.17) must be of the form

$$\pm((R\boldsymbol{X})'' + \boldsymbol{f}(R\boldsymbol{X}, (R\boldsymbol{X})', \lambda)) = \mathbf{0}. \tag{4.18}$$

But by (4.15) we see that the + sign must obtain in (4.18), so the operator on the left side of (4.17) is equivariant under R. Therefore, so is $\boldsymbol{f}$. □

Thus our reduced problem consists of restricting the mapping (4.13) to functions $\boldsymbol{X} \in B$ satisfying $R\boldsymbol{X} = \boldsymbol{X}$. Note that the full symmetry group for the problem is $O(2) \simeq \{T_\alpha, RT_\alpha : \alpha \in R(\text{mod} 2\pi)\}$. We next turn to the linearization of the problem.

5. The linearized problem. In this section we will consider the linearization of the problem about the trivial solution. For this we set $\boldsymbol{r} = (0,0,bs)^T + \epsilon(x_1, x_2, x_3)^T, \boldsymbol{v} = (0,0,b)^T + \epsilon(v_1, v_2, v_3), \boldsymbol{u} = \epsilon(u_1, u_2, u_3)^T$, and $\Psi = \epsilon(\theta, \phi, \psi)^T$. Linearization of (2.12) gives $D = I + \epsilon D_1 + O(\epsilon^2)$, where

$$D_1 = \begin{pmatrix} 0 & -\psi & \phi \\ \psi & 0 & -\theta \\ -\phi & \theta & 0 \end{pmatrix}, \tag{5.1}$$

while linearization of (2.13) gives

$$u_1 = \theta', \; u_2 = \phi', \; u_3 = \psi'. \tag{5.2}$$

Finally, linearization of (2.11) yields

$$x_1' = v_1 + b\phi, \; x_2' = v_2 - b\theta, \; x_3' = v_3. \tag{5.3}$$

At the trivial solution, the matrix (2.34) is given by

$$\operatorname{diag}(\sigma, \sigma, m_{33}, \tau, \tau, n_{33}), \tag{5.4}$$

where σ and τ are the values of $\hat{\sigma}$ and $\hat{\tau}$ (defined by (2.30)) evaluated at the trivial solution, while $m_{33} = \partial m_3/\partial u_3$, $n_{33} = \partial n_3/\partial v_3$ evaluated there. Of course all the constants in (5.4) are positive. We now use (5.1)–(5.4) to linearize (2.33). The result is

$$\tau x_1'' + (n_0 - \tau b)\phi' + \lambda x_2' = 0, \tag{5.5a}$$
$$\tau x_2'' - (n_0 - \tau b)\theta' - \lambda x_1' = 0, \tag{5.5b}$$
$$n_{33} x_3'' = 0, \tag{5.5c}$$
$$\sigma\theta'' + b(n_0 - \tau b)\theta + (n_0 - \tau b)x_2' = 0, \tag{5.5d}$$
$$\sigma\phi'' + b(n_0 - \tau b)\phi - (n_0 - \tau b)x_1' = 0, \tag{5.5e}$$
$$m_{33}\psi'' = 0. \tag{5.5f}$$

The boundary conditions are

$$x_1 = x_2 = x_3 = \theta = \phi = \psi = 0 \text{ at } s = 0 \text{ and } s = 1. \tag{5.6}$$

Our linearized problem is (5.5), (5.6). Of course it follows immediately from (5.5c), (5.5f), and (5.6) that $x_3 \equiv 0$, $\phi \equiv 0$. The remaining equations can be simplified. We let

$$a = n_0 - \tau b, \quad X = x_1 + i x_2, \quad W = \theta + i\phi. \tag{5.7}$$

Then equations (5.5a), (5.5b), (5.5d), and (5.5f) can be written in *complex* form;

$$\tau X'' - iaW' - i\lambda X' = 0, \tag{5.8a}$$
$$\sigma W'' + abW - iaX' = 0. \tag{5.8b}$$

The boundary conditions become

$$X(0) = X(1) = W(0) = W(1) = 0. \tag{5.9}$$

Our problem is to find the eigenvalues of (5.8), (5.9). The form (5.8) shows that the dimension of the eigenspaces for (5.5), (5.6) must be even. This is because the dimension of an eigenspace for (5.5), (5.6) is twice the dimension of the corresponding eigenspace for (5.8), (5.9). In [7] we proved the existence of an infinite set of real eigenvalues of this system under the assumption that

$$a < \frac{1}{2}(\sqrt{b^2\tau^2 + 4\pi^2\tau\sigma} - b\tau). \tag{5.10}$$

This assumption implies that, if we write our system in the form $Au = \lambda Bu$ with u in a suitable Hilbert space, the operator A is positive definite. In this paper we show that we can, for the most part, dispense with the condition (5.10). Here we only assume

$$ab \neq n^2\pi^2\sigma, \quad n = 1, 2, 3, \ldots. \tag{5.11}$$

(It is by no means clear that the result is false if (5.11) is violated. Of course, if (5.11) is false but (5.10) holds, we can still appeal to the argument of [7].)

Under the assumption that (5.10) holds, we can solve (5.8b) for W in terms of X:

$$W = iaK_1(X'), \tag{5.12}$$

where K_1 is the Green's operator for (5.8b) with the boundary conditions $W(0) = W(1) = 0$. We next insert this into (5.8a) to obtain

$$\tau X'' + a^2 K_1(X')' - i\lambda X' = 0. \tag{5.13}$$

We let K_0 be the Green's operator which inverts the operator $\tau X''$ with the boundary conditions $X(0) = X(1) = 0$. Then (5.8), (5.9) is equivalent to

$$X + a^2 K_0[K_1(X')'] - i\lambda K_0(X') = 0. \tag{5.14}$$

We consider (5.14) on the space $H_1^0[0,1]$ equipped with the norm

$$\|x\|^2 = \int_0^1 |x'(s)|^2\, ds.$$

LEMMA 5.1. *The mapping $u \to K_0[K_1(u')']$ is self-adjoint and compact on H_1^0.*

Proof. Let $\langle \cdot, \cdot \rangle$ denote the H_1^0 inner product and $(\cdot,\cdot)$ denote the L_2 inner product, so that $\langle u, v\rangle = (u', v')$. Then if $u, v \in C^2 \cap H_1^0$, a dense subset of H_1^0,

$$\begin{aligned}\langle u, K_0[K_1(v')']\rangle &= (u', K_0[K_1(v')']') = -(u'', K_0[K_1(v')']) = -\frac{1}{\tau}(u, K_1(v')')\\ &= \frac{1}{\tau}(u', K_1(v')') = \frac{1}{\tau}(K_1(u'), v') = -\frac{1}{\tau}(K_1(u')', v) = -(K_1(u')', K_0(v''))\\ &= -(K_0[K_1(u')'], v'') = (K_0[K_1(u')']', v') = \langle K_0[K_1(u')], v\rangle.\end{aligned}$$

To prove compactness we need only count derivatives. We find that the mapping maps H_1^0 into H_3, thus is compact on H_1^0. □

LEMMA 5.2. *The mapping $u \to iK_0(u')$ is self-adjoint, compact and injective on H_1^0.*

Proof. Let u and v be as in the proof of Lemma 5.15. Then

$$\begin{aligned}\langle u, iK_0(v')\rangle &= (u', iK_0(v')') = -(u'', iK_0(v')) = -(K_0(u''), iv)') = -\frac{1}{\tau}(u, iv')\\ &= -\frac{1}{\tau}(iu', v) = -(iu', K_0(v'')) = -(iK_0(u'), v'') = -(iK_0(u')', v') = \langle iK_0(u'), v\rangle.\end{aligned}$$

The compactness and injectivity are clear. □

From Lemmas 5.1 and 5.2 we see that (5.13) can be written in the form

$$M(\lambda)u = 0, \tag{5.15}$$

where $M(\lambda)$ is the linear pencil

$$M(\lambda) = I - T - \lambda H \tag{5.16}$$

with T and H compact and self-adjoint and H injective with its set of eigenvalues square summable. Of course, if $I - T$ or H were positive definite, the spectral theory of (5.15), (5.16) would be straightforward. (This is essentially the case considered in [7].) The theory of such pencils is treated in [3, Chap. V]. This theory is a chapter in the theory of operators on an inner product space with an indefinite inner product, in particular a *Pontryagin space.*

It is shown in [3] that, without loss of generality, we may assume that $I - T$ is invertible. We then introduce a new (possibly indefinite) inner product on H_1^0 by setting

$$\{u, v\} = \langle (I - T)^{-1}u, v\rangle. \tag{5.17}$$

With respect to this inner product, the operator $A = H(I - T)^{-1}$ is self-adjoint and (5.15) is equivalent to

$$(I - \lambda A)v = 0, \quad v = (I - T)u. \tag{5.18}$$

If $I - T$ and therefore $(I - T)^{-1}$ are positive definite, then $\{u, u\}$ is equivalent to $\langle u, u\rangle$ and the usual theory applies. In the case where this is not true, we have the following result.

THEOREM 5.3 [3]. *Under the above hypotheses, we can write*

$$H_1^0 = N \oplus P,$$

where the following hold:

- *N and P are invariant subspaces with respect to the operator A.*
- *N is finite dimensional.*
- *The inner product $\{\cdot, \cdot\}$ is positive on P.*
- *The operator $A|_P$ has a complete set of eigenvectors which correspond to real eigenvalues and are orthogonal with respect to the inner product $\{\cdot, \cdot\}$.*

From this result, it follows immediately that (5.14) and hence (5.8), (5.9) have infinitely many real eigenvalues.

In the next section, when we apply all of this to prove our main result, we will need to verify condition (ii) of Theorem 3.1. We will now do this for the operator L, the linearization of (4.13). It will then hold a fortiori for the operator $L_{\mathcal{H}}$ which we shall define in the next section.

So suppose we have

$$L(\lambda)v_0 = 0, \ L'(\lambda)v_0 = L(\lambda)v_1, \ v_0, v_1 \in B, \ v_0 \neq 0, \ \lambda \neq 0. \tag{5.19}$$

The components of v_0 and v_1 are in H_1^0. Thus, if we retrace the steps which led to (5.15), (5.16), we see that (5.19) is equivalent to

$$(I - T - \lambda H)y_0 = 0, \ (I - T - \lambda H)y_0 = Hy_0, \ y_0 \neq 0. \tag{5.20}$$

We take the inner product of the second equation of (5.20) with y_0 and use the fact that $I - T - \lambda H$ is self-adjoint. The result is $\langle Hy_0, y_0\rangle = 0$. But

$$\langle Hy_0, y_0\rangle = \frac{1}{\lambda}\langle (I - T)y_0, y_0\rangle = \frac{1}{\lambda}\{(I - T)y_0, (I - T)y_0\} \neq 0$$

by Theorem 5.3 since $(I - T)y_0 \in P$. Thus we have proven the following.

PROPOSITION 5.4. *If $v \in N(L(\lambda))\backslash\{0\}$, $\lambda \neq 0$, then $L'(\lambda)v \notin R(L(\lambda))$.*

6. Global bifurcation. We are now ready to prove our main result by applying Theorem 3.1. As our subgroup of $O(2)$, $\mathcal{H}$, we of course take Z_2, whose representation on B is given by $\{I, R\}$. Thus $B_{\mathcal{H}}$ consists of all elements of B which are invariant under the mapping R defined by (4.14). We need only check that $\dim(L_{\mathcal{H}}(\lambda_0)) = 1$, where λ_0 is an eigenvalue of the linearized equation (5.8), (5.9). In terms of the complex variables X and W, the $\mathcal{H}$-invariance condition is

$$(6.1) \qquad \overline{X}(1-s) = X(s), \ \overline{W}(1-s) = W(s).$$

Now it may happen that for certain values of the parameters in (5.8), the eigenspace corresponding to an eigenvalue λ_0 has dimension greater that one (over the complex numbers). However, generically this will not happen. (In principle, the eigenvalues can be found by algebraic means since (5.8) is a system of ordinary differential equations with constant coefficients.) So let us suppose that λ_0 is an eigenvalue of (5.8), (5.9) whose eigenspace is one dimensional over C (two dimensional over R). Let $\{X(s), W(s)\}$ be an eigenfunction of (5.8), (5.9). Then there is a complex constant C such that

$$(6.2) \qquad \overline{X}(1-s) = CX(s), \ \overline{W}(1-s) = CW(s).$$

We must have $|C| = 1$. To see this, take absolute values in (6.2) and integrate from 0 to 1. Let

$$(6.3) \qquad \{X_1(s), W_1(s)\} = \alpha\{X(s), W(s)\}.$$

In order for $\{X_1, W_1\}$ to satisfy (6.1), we observe that

$$\overline{X_1}(1-s) = \overline{\alpha}\overline{X}(1-s) = \overline{\alpha}CX(s) = \frac{\overline{\alpha}}{\alpha}CX_1(s),$$

so that we must have

$$\frac{\overline{\alpha}}{\alpha}C = 1.$$

Hence $\{X_1, W_1\}$ satisfies (6.1) if

$$(6.4) \qquad \alpha = e^{i\gamma/2} \text{ where } C = e^{i\gamma}.$$

We therefore have proven the following theorem.

THEOREM 6.1 (global bifurcation theorem). *Suppose the parameters appearing in* (5.8) *satisfy* (5.10) *or* (5.11). *Then the system* (5.8), (5.9) *has an infinite set of real eigenvalues and, except for perhaps an exceptional set of these parameters, for each of these eigenvalues λ, the point $(\mathbf{0}, \lambda)$ is a bifurcation point for the problem* (2.32), (2.33). *Moreover, there is a global bifurcating branch $\zeta_{\mathcal{H}}$ of solutions through $(\mathbf{0}, \lambda)$ which are invariant under the mapping* (4.14). *$\zeta_{\mathcal{H}}$ is at least locally a curve and satisfies the conclusion of Theorem* 3.1 *where Ω is defined by* (4.12).

Remarks.

1. It seems likely that (5.11) is needed only for the proof we give, and even if it is violated, the conclusion of Theorem 6.1 remains valid. Likewise, if for some value of the parameters (5.8), (5.9) does have an eigenvalue of complex multiplicity greater than one, most of the eigenvalues should be simple.

2. Since the Euler angles represent local coordinates on $SO(3)$, it should be possible to remove the restriction $|\theta| < \pi/2$ from the definition of Ω.
3. By the rotational invariance, if $(\boldsymbol{X}, \lambda) \in \zeta_{\mathcal{H}}$, then $(T_\alpha \boldsymbol{X}, \lambda)$ is a solution of the problem. Thus the branch of solutions bifurcating from $(\mathbf{0}, \lambda)$ is, in fact, $\zeta_{\mathcal{H}} \times R(\mathrm{mod} 2\pi)$.

REFERENCES

[1] S. S. Antman and C. S. Kenney, *Large buckled states of nonlinearly elastic rods under torsion, thrust and gravity*, Arch. Rational Mech. Anal., 76 (1981), pp. 289–338.

[2] M. Crandall and P. Rabinowitz, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321–340.

[3] I. C. Gohberg and M. G. Krein, *Introduction to the theory of linear nonselfadjoint operators*, American Mathematical Society, Providence, RI, 1969.

[4] T. J. Healey, *Global bifurcation in the presence of symmetry with an application to solid mechanics*, SIAM J. Math. Anal., 19 (1988), pp. 824–840.

[5] ———, *Large rotating states of a conducting elastic wire in a magnetic field: Subtle symmetry and multiparameter bifurcation*, J. Elasticity, 24 (1990), pp. 211–227.

[6] T. Seidman and P. Wolfe, *Equilibrium states of an elastic conductor in a magnetic field*, Arch. Rational Mech. Anal., 102 (1988), pp. 307–329.

[7] P. Wolfe, *Bifurcation theory of an elastic conducting rod in a magnetic field*, Quart. J. Mech. Appl. Math., 41 (1988), pp. 265–279.

[8] ———, *Rotating states of an elastic conductor*, in Physical Mathematics and Nonlinear Partial Differential Equations, J. Lightbourne and S. Rankin, eds., Dekker, New York, 1985, pp. 213–222.

SIAM J. MATH. ANAL.
Vol. 27, No. 2, pp. 543–557, March 1996

012

ON SEMILINEAR PROBLEMS WITH NONLINEARITIES DEPENDING ONLY ON DERIVATIVES*

A. CAÑADA[†] AND P. DRÁBEK[‡]

Abstract. In this paper we deal with the semilinear boundary value problems (BVPs)

$$u''(t) + \lambda_1 u(t) + g(t, u'(t)) = f(t), \ t \in I, \ \ (Bu)(t) = 0, \ \ t \in \partial I,$$

where $I = [0, \pi]$, B denotes either the Dirichlet or the Neumann or the periodic boundary conditions, respectively, and λ_1 is the first eigenvalue of the corresponding linear problem

$$u'' + \lambda u(t) = 0, \ \ t \in I, \ \ (Bu)(t) = 0, \ \ t \in \partial I.$$

This kind of problem is very important in applications where the quantity $g(t, u')$ may be regarded as a nonlinear damping term. The nonlinear function g is supposed to be bounded and, in some cases, satisfies additional differentiability assumptions and asymptotic conditions. We emphasize the dependence of g on the derivative of the solution $u'(t)$ in order to show the qualitative difference of this case and the "classical" Landesman–Lazer-type problem in which the nonlinearity g depends only on the solution $u(t)$.

Key words. semilinear boundary value problems, ordinary differential equations of second order, solvability, nonlinear damping, bounded nonlinearities, alternative method

AMS subject classifications. 34B15, 34C25

1. Introduction.

In this paper we study semilinear problems of the type

$$\begin{cases} u''(t) + \lambda_1 u(t) + g(t, u'(t)) = f(t), & t \in I, \\ (Bu)(t) = 0, & t \in \partial I, \end{cases} \tag{1.1}$$

where I is a closed interval $[0, \pi]$, B denotes either the *Dirichlet boundary conditions*

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

or the *Neumann boundary conditions*

$$u'(0) = u'(\pi) = 0, \tag{1.3}$$

or the *periodic boundary conditions*

$$u(0) = u(\pi), \ u'(0) = u'(\pi), \tag{1.4}$$

respectively, and λ_1 is the first eigenvalue of the eigenvalue problem

$$\begin{cases} u'' + \lambda u(t) = 0, & t \in I, \\ (Bu)(t) = 0, & t \in \partial I. \end{cases} \tag{1.5}$$

The main purpose of this work is to *emphasize the qualitative difference* between our case and the "classical" Landesman–Lazer-type problem where the nonlinearity g depends only on the solution u and not on its derivative u'. This qualitative difference

* Received by the editors January 3, 1994; accepted for publication July 26, 1994.

[†] Departamento de Analisis Matematico, Universidad Granada, 18071, Granada, Spain (acanada@goliat.ugr.es). This author has been supported by Direccioń General de Investigacíon Cientifica y Tecnológica, Ministry of Education and Science (Spain), grant PB92-0941.

[‡] Department of Mathematics, University of West Bohemia Plzeň, Americká 42, 306 14 Plzeň, Czech Republic (pdrabek@kma.zcu.cz). This author was supported by Grant Agency of the Czech Republic grant 201/94/0008.

will be shown by describing the range of the operator $u''(t)+\lambda_1 u(t)+g(t,u'(t))$ (under $(Bu)(t)=0$) and then comparing the obtained results with the corresponding ones for the case $g(t,u(t))$. Our paper is not an exhaustive treatment of the problems of this kind (in fact, after reading §4 of this paper, one would have the impression that such a paper would be too long); we deal only with the case of ordinary differential equations (ODEs) of the second order and with the boundary conditions of the type (1.2)–(1.4). Note that the *techniques* used here are *typical for ODEs.* In the case of Neumann or periodic boundary value problems (BVPs), respectively, we use the *shooting method* (more precisely, the continuous dependence of the solution of an associated initial-value problem on the parameter and on the initial conditions, and some differentiability property of this dependence—see, e.g., [7]). The Dirichlet problem is studied by an *alternative method* combined with the method of *lower* and *upper* solutions (which should also be applied in the case of partial differential equations (PDEs)) but here we use *essentially the shape* of the *first eigenfunction* $\sin t$ and *its derivative* $\cos t$.

Many *natural questions* arise during the proofs; that is why we formulate some of them explicitly *in the last section* of this paper.

It should be also mentioned here that this is not the first attempt to give some satisfactory answer to the question of solvability of problems like (1.1). Motivated by the celebrated papers [15] and [13], various authors dealt with semilinear problems where the nonlinearity depends not only on the solution but also on the derivatives (or on the gradient) of it. Let us mention, e.g., papers [3], [4], [9], [19], [6], and [5]. In all these papers different variants of Landesman–Lazer conditions were considered in order to give (at least) sufficient conditions for the solvability of investigated BVPs (for analogous results concerning the nonlinearity g depending only on the solution u we refer the reader to [8] and [10] and to the references therein). However, these results were not completely satisfactory, because they do not allow one to deal with the most simple case $g=g(u')$ (i.e., g does not depend on t). Thus it was not clear up to now how to characterize the solvability of the BVP

$$\begin{gathered} u''(t)+u(t)+g(u'(t))=f(t), \quad t\in(0,\pi), \\ u(0)=u(\pi)=0 \end{gathered}$$

(see [3]) or related problems with boundary conditions (1.3) and (1.4), respectively.

Note that semilinear problems of the type (1.1) are very important in applications. Let us mention the problems arising in viscosity, nonlinear oscillations, electric circuits, etc. The term $g(u')$ may be regarded as a *nonlinear damping term* in *resonance problems* and it appears, e.g., in Rayleigh's equation (which is closely connected with a theory of oscillation of a violin string), in oscillations of a simple pendulum under the action of viscous damping, in dry (or Coulomb) friction (which occurs when the surfaces of two solids are in contact and in relative motion without lubrication), and in some cases of van der Pol oscillators. (Refer to [11], [20], [14], and [18] for bibliography).

2. Preliminaries. Let us assume that $g: I\times\mathbb{R}^2\to\mathbb{R}$ is a continuous function, bounded by a constant $M>0$:

$$|g(t,\eta,\xi)|\le M \tag{2.1}$$

for $t\in I$ and $(\eta,\xi)\in\mathbb{R}^2$. Consider the BVP

$$\left\{\begin{aligned} u''(t)+\lambda_1 u(t)+g(t,u(t),u'(t))&=f(t), \quad t\in I, \\ (Bu)(t)&=0, \quad t\in\partial I. \end{aligned}\right. \tag{2.2}$$

Let $u_1 = u_1(t)$ be the *positive eigenfunction* associated with $\lambda = \lambda_1$ of the eigenvalue problem (1.5), which satisfies $\int_0^\pi u_1^2(t)dt = 1$.

PROPOSITION 2.1. *Let $f \in C(I)$ be written in the form $f(t) = su_1(t) + \tilde{f}(t)$, $s \in$ $\mathbb{R}$, $\tilde{f} \in C(I)$, $\int_0^\pi \tilde{f}(t)u_1(t)dt = 0$. Then, for any $\tilde{f}$ there exists a nonempty, connected, and bounded set $\mathcal{J}_{\tilde{f}} \subset \mathbb{R}$ such that the BVP (2.2) has at least one solution $u \in C^2(I)$ if and only if $s \in \mathcal{J}_{\tilde{f}}$.*

Sketch of the proof of Proposition 2.1. Let Z denote the Banach space $Z = C(I)$ with the norm $\|z\|_0 = \max_{t\in I} |z(t)|$ for any $z \in Z$. By U we denote the Banach space $U = \{u \in C^1(I) : (Bu)(t) = 0,\ t \in \partial I\}$ with the norm $\|u\|_1 = \max\{\|u\|_0, \|u'\|_0\}$. If we define $L : \mathrm{dom}L \subset U \to Z$ by $\mathrm{dom}L = \{u \in U : \ u \in C^2(I)\}$, $Lu = u'' + \lambda_1 u$ and $N : \ U \to Z$ by $(Nu)(t) = f(t) - g(t, u(t), u'(t))$, for any $u \in U$, and $t \in I$, then our problem (2.2) is equivalent to solving the operator equation $Lu = Nu$. It is well known (see, e.g., [10]) that there exist continuous projections $P : \ U \to U$ and $Q : \ Z \to Z$ such that $Lu = Nu$ is equivalent to the alternative system

$$u - Pu = K(I - Q)Nu,$$

$$QNu = 0,$$

where K is the (continuous) inverse of the mapping $L : \ \mathrm{dom}L \cap \mathrm{Ker}P \to \mathrm{Im}L \equiv \mathrm{Ker}Q$. Now, writing $u \in U$ in the form $u(t) = cu_1(t) + v(t)$, $c \in \mathbb{R}$, $\int_0^\pi v(t)u_1(t)\ dt = 0$, the BVP (2.2) is equivalent to the system

$$v = K(I - Q)N(cu_1(\cdot) + v), \tag{2.3}$$

$$QN(cu_1(\cdot) + v) = 0. \tag{2.4}$$

Applying the Schauder fixed-point theorem we get that for any fixed $c \in \mathbb{R}$ there exists at least one $v_c \in U \cap C^2(I)$ such that (2.3) holds (see, e.g., [10]).

Equation (2.4) is now

$$QN(cu_1(\cdot) + v_c) = 0,$$

which, taking into account the expression for Q (see again [10]), becomes

$$\int_0^\pi g(t, cu_1(t) + v_c(t), cu_1'(t) + v_c'(t))u_1(t)\ dt = s. \tag{2.4a}$$

Hence, for a given $\tilde{f} \in C(I)$, $\int_0^\pi \tilde{f}(t)u_1(t)\ dt = 0$, the BVP (2.2) with $f(t) = su_1(t) + \tilde{f}(t)$ has at least one solution if and only if s belongs to the range of the (multivalued, in general) function $\Gamma_{\tilde{f}} : \mathbb{R} \to \Gamma_{\tilde{f}}(\mathbb{R})$,

$$\Gamma_{\tilde{f}}(c) = \int_0^\pi g(t, cu_1(t) + v_c(t), cu_1'(t) + v_c'(t))u_1(t)\ dt, \tag{2.5}$$

where $v_c \in \{v \in U \cap C^2(I) : \ v \text{ is a solution of (2.3) for fixed } c\}$. But $\mathcal{J}_{\tilde{f}} \equiv \Gamma_{\tilde{f}}(\mathbb{R})$ is a connected set. In fact, let s^* and $s^\sharp$ belong to $\mathcal{J}_{\tilde{f}}$ and $s^* \leq s^\sharp$. Then the BVP (2.2) with $f^* = s^*u_1 + \tilde{f}$ and $f^\sharp = s^\sharp u_1 + \tilde{f}$ has solutions u^* and $u^\sharp$, respectively. If we consider now the BVP (2.2) with $f = su_1 + \tilde{f}$, where $s \in [s^*, s^\sharp]$, then u^* is an upper solution and $u^\sharp$ is a lower solution to this problem. Due to a result in [1], there exists at least one solution, i.e., s belongs to $\mathcal{J}_{\tilde{f}}$. Moreover, since g is bounded, the range of $\Gamma_{\tilde{f}}$ is bounded.

Remark 2.1. The proof of Proposition 2.1 is based on the alternative method (see, e.g., [10]) and on an interesting result related to the method of lower and upper solutions due to the authors in [1]. In fact, it follows from [1] that the same assertion as in Proposition 2.1 holds true also for *more general* BVPs of the type

$$\begin{aligned} \Delta u + \lambda_1 u + g(x,u,\nabla u) = f(x) \quad &\text{in} \quad \Omega,\\ Bu = 0 \quad &\text{on} \quad \partial\Omega, \end{aligned}$$

with g bounded.

Remark 2.2. Note that we have used the fact that *existence* for the BVP (2.2) follows even if u^* and $u^\sharp$ *are not ordered* (see [1]). The periodic problem is not considered in [1] but the reasoning is the same.

Remark 2.3. Further information concerning the properties of the "solution set" $\mathcal{J}_{\tilde{f}}$ from Proposition 2.1 will be obtained from the *concrete form* of the nonlinearity g in next section. For this purpose, it is useful to define $s_1(\tilde{f}) = \inf \mathcal{J}_{\tilde{f}}$, $s_2(\tilde{f}) = \sup \mathcal{J}_{\tilde{f}}$.

Remark 2.4. We must also remark that the bounded interval $\mathcal{J}_{\tilde{f}}$ may be open, or closed, or open from below and closed from above (or vice versa) (see next section).

3. Main results. Let us point out that there are *many results* concerning the structure of the range of a multivalued function $\Gamma_{\tilde{f}}$ in the case when *nonlinearity g does not depend on the derivative* of the solution u'. One of the most illustrative characterizations of the range of $\Gamma_{\tilde{f}}$ is given by Landesman–Lazer-type conditions formulated at first in [13]. This result has been generalized in many subsequent papers (see the bibliographies in [10] and [8]). Many papers document that the structure of the right-hand sides f for which the BVP (2.2) has at least one solution is closely related to the shape of nonlinearity g and to its qualitative properties (the growth or limits at infinity, smoothness, oscillatory properties, sign condition, etc.).

On the other hand there are *few results* for the case when *nonlinearity g depends also on the derivative.* Only several attempts have been made to adopt Landesman–Lazer-type conditions to guarantee the solvability of (2.2). However, in these cases either the fact that $g = g(t,u,u')$ depends also on u is essential (see, e.g., [9]), or an additional monotonicity hypothesis on g is considered (see [19]), or a suitable dependence of g on $t \in I$ is required (see, e.g., [3, 4]). To illustrate the latter case let us consider the BVP (1.1), where $\lambda_1 = 1$ and B denotes the Dirichlet boundary condition. Then the Landesman–Lazer-type conditions

$$\begin{aligned} &\int_0^{\pi/2} g(t,+\infty)\sin t\,dt + \int_{\pi/2}^{\pi} g(t,-\infty)\sin t\,dt < \int_0^{\pi} f(t)\sin t\,dt \\ &\qquad < \int_{\pi/2}^{\pi} g(t,+\infty)\sin t\,dt + \int_0^{\pi/2} g(t,-\infty)\sin t\,dt \end{aligned} \tag{3.1}$$

are sufficient for the solvability of (1.1) (see [3, 4]) (here we denote $g(t,\pm\infty) = \lim_{\xi\to\pm\infty} g(t,\xi)$). Observe that if, moreover,

$$\begin{aligned} g(t,+\infty) < g(t,u'), &\quad \forall t \in (0,\pi/2), \quad \forall u' \in \mathbb{R},\\ g(t,-\infty) < g(t,u'), &\quad \forall t \in (\pi/2,\pi), \quad \forall u' \in \mathbb{R},\\ g(t,+\infty) > g(t,u'), &\quad \forall t \in (\pi/2,\pi), \quad \forall u' \in \mathbb{R},\\ g(t,-\infty) > g(t,u'), &\quad \forall t \in (0,\pi/2), \quad \forall u' \in \mathbb{R}, \end{aligned}$$

then the previous condition is not only sufficient but also necessary to the solvability of (1.1). So that, if for $f \in C(I)$ we write $f(t) = s\sqrt{2/\pi}\sin t + \tilde{f}(t)$,

$s = \sqrt{2/\pi}\int_0^\pi f(t)\sin t\,dt$, $\int_0^\pi \tilde{f}(t)\sin t\,dt = 0$, then (1.1) has solution if and only if $s \in (s_1(\tilde{f}), s_2(\tilde{f}))$, where

$$s_1(\tilde{f}) = \sqrt{\frac{2}{\pi}}\left[\int_0^{\pi/2} g(t,+\infty)\sin t\,dt + \int_{\pi/2}^{\pi} g(t,-\infty)\sin t\,dt\right],$$

$$s_2(\tilde{f}) = \sqrt{\frac{2}{\pi}}\left[\int_{\pi/2}^{\pi} g(t,+\infty)\sin t\,dt + \int_0^{\pi/2} g(t,-\infty)\sin t\,dt\right]$$

(realize that now $s_1(\tilde{f})$ and $s_2(\tilde{f})$ are both independent of $\tilde{f}$!), which shows that $\mathcal{J}_{\tilde{f}}$, in Proposition 2.1, may be open.

However, it should be verified immediately that *conditions* (3.1) *are empty* in the case $g = g(u')$ (i.e., when g does not depend on t). Hence this approach does not apply in this more simple case, which establishes a deep difference with respect to the case where g depends only on u.

It appears that the question of *solvability* of the BVP

$$\begin{aligned} u''(t) + \lambda_1 u(t) + g(u'(t)) &= f(t), \quad t \in I,\\ (Bu)(t) &= 0, \quad t \in \partial I \end{aligned}$$

is *qualitatively different* from the case

$$\begin{aligned} u''(t) + \lambda_1 u(t) + g(u(t)) &= f(t), \quad t \in I,\\ (Bu)(t) &= 0, \quad t \in \partial I. \end{aligned}$$

Since this difference depends essentially on the type of boundary conditions, we consider separately B of the forms (1.2), (1.3), and (1.4).

3.1. The Dirichlet problem. Let us study the BVP

$$\left\{\begin{aligned} u''(t) + u(t) + g(u'(t)) &= f(t), \quad t \in (0,\pi),\\ u(0) &= u(\pi) = 0. \end{aligned}\right. \tag{3.2}$$

We will consider nonlinear function g which will document the difference between the situation considered here and in [13].

In this subsection we will write the right-hand side of the BVP (3.2) in the following form:

$$f(t) = s\sqrt{\frac{2}{\pi}}\sin t + \tilde{f}(t),\ \ s \in \mathbb{R},\quad \int_0^\pi \tilde{f}(t)\sin t\,dt = 0.$$

THEOREM 3.1. *Let g be a bounded and continuous real function of a real variable satisfying $g(+\infty) = g(-\infty)$ and $g(\xi) < g(+\infty)$ for any $\xi \in \mathbb{R}$. Then for any $\tilde{f} \in C(I)$, $\int_0^\pi \tilde{f}(t)\sin t\,dt = 0$ there exists a real number $g_{\tilde{f}} < 2\sqrt{2/\pi}g(+\infty)$ such that the Dirichlet BVP* (3.2) *has at least one solution $u \in C^2(I)$ if and only if*

$$s \in \left[g_{\tilde{f}}, 2\sqrt{\frac{2}{\pi}}g(+\infty)\right).$$

Proof. The starting point is Proposition 2.1. First, note that now $u_1(t) = \sqrt{2/\pi}\sin t$. Due to the considerations in the sketch of the proof of Proposition 2.1, it is sufficient to show that for a given $\tilde{f} \in C(I)$, $\int_0^\pi \tilde{f}(t)\sin t\, dt = 0$ we have

$$\Gamma_{\tilde{f}}(\mathbb{R}) = \left[g_{\tilde{f}}, 2\sqrt{\frac{2}{\pi}}g(+\infty)\right).$$

The (possibly multivalued) function $\Gamma_{\tilde{f}}$ has the following form:

$$\Gamma_{\tilde{f}}(c) = \sqrt{\frac{2}{\pi}}\int_0^\pi g(c\cos t + v_c'(t))\sin t\, dt,$$

where $c \in \mathbb{R}$ and $v_c \in C^2(I)$ verify the equation (2.3). In particular, it follows from the boundedness of g that there exists a constant $D > 0$ such that

$$\|v_c\|_{C^1} \leq D \tag{3.3}$$

for any $c \in \mathbb{R}$ (see, e.g., [10] for related estimates). Thus we have

$$c\cos t + v_c'(t) \to \pm\infty \text{ for } t \in \left(0, \frac{\pi}{2}\right) \text{ and } c \to \pm\infty,$$

$$c\cos t + v_c'(t) \to \mp\infty \text{ for } t \in \left(\frac{\pi}{2}, \pi\right) \text{ and } c \to \pm\infty.$$

Applying the Lebesgue-dominated convergence theorem we obtain

$$\int_0^\pi g(c\cos t + v_c'(t))\sin t\, dt \to \int_0^{\pi/2} g(\pm\infty)\sin t\, dt + \int_{\pi/2}^\pi g(\mp\infty)\sin t\, dt \tag{3.4}$$

for $c \to \pm\infty$. Due to $g(+\infty) = g(-\infty)$ we get

$$\Gamma_{\tilde{f}}(c) \to 2\sqrt{\frac{2}{\pi}}g(+\infty) \text{ for } c \to \pm\infty.$$

The assumption $g(\xi) < g(+\infty)$, $\xi \in \mathbb{R}$, and (3.3) yield

$$\Gamma_{\tilde{f}}(c) < 2\sqrt{\frac{2}{\pi}}g(+\infty)$$

for any $c \in \mathbb{R}$. Let us denote

$$g_{\tilde{f}} = \inf_{c\in\mathbb{R}} \Gamma_{\tilde{f}}(c).$$

It is sufficient to prove that this infimum is achieved. We use the standard compactness argument. Let $\{s_n\} \subset \Gamma_{\tilde{f}}(\mathbb{R})$ be such that $s_n \to g_{\tilde{f}}$ and $\{c_n\}$ be the corresponding minimizing sequence, i.e., $u_n = c_n\sqrt{2/\pi}\sin t + v_{c_n}(t)$ are the solutions of the BVP (3.2) with the right-hand sides $f_n = s_n\sqrt{2/\pi}\sin t + \tilde{f}(t)$. Then the sequence $\{c_n\}$ is bounded due to $g(\xi) < g(+\infty)$, $\xi \in \mathbb{R}$, (3.3), and (3.4). Applying a standard compactness argument usual in alternative methods (see, e.g., [10]), we show that $c_n \to c$ (at least for a subsequence) and that $u(t) = c\sqrt{2/\pi}\sin t + v_c(t)$ is a solution of the BVP (3.2) with $f(t) = g_{\tilde{f}}\sqrt{2/\pi}\sin t + \tilde{f}(t)$. Hence the infimum is achieved in c. □

Remark 3.1. The necessary and sufficient condition from Theorem 3.1 can be written in the following equivalent form:

$$\sqrt{\frac{\pi}{2}}g_{\tilde{f}} \leq \int_0^\pi f(t)\sin t\, dt < 2g(+\infty). \tag{3.5}$$

Note that the solvability is determined not only by the asymptotic behavior of g but also that the behavior of g on bounded intervals in $\mathbb{R}$ is essential (a similar result for the case where g depends only on u may be seen in [2], which shows that, in some particular cases, the solvability of problem (3.2) may be like the case $g = g(u)$). Observe also that $g_{\tilde{f}}$ depends, in general, on $\tilde{f}$. The exact determination of $g_{\tilde{f}}$, for g and $\tilde{f}$ in general seems to be a difficult question.

Example 3.1. Let $g(\xi) = \arctan\xi$ for $\xi \geq 0$, $g(\xi) = g(-\xi)$ for $\xi < 0$. Then the condition (3.5) has the form

$$\sqrt{\frac{\pi}{2}}g_{\tilde{f}} \leq \int_0^\pi f(t)\sin t\, dt < \pi,$$

where $0 \leq g_{\tilde{f}} < \sqrt{2\pi}$ for any $\tilde{f} \in C(I)$, $\int_0^\pi \tilde{f}(t)\sin t\, dt = 0$. Moreover, it should be easily verified that $g_{\tilde{f}} = 0$ for $\tilde{f} \equiv 0$ in $(0, \pi)$.

Remark 3.2. The assumptions on g from Theorem 3.1 imply that the "solution set" $\mathcal{J}_{\tilde{f}}$ is an interval

$$\mathcal{J}_{\tilde{f}} = [s_1(\tilde{f}), s_2(\tilde{f})).$$

Considering nonlinearity $-g$ in Theorem 3.1, we obtain that $\mathcal{J}_{\tilde{f}}$ is of the following form:

$$\mathcal{J}_{\tilde{f}} = (s_1(\tilde{f}), s_2(\tilde{f})].$$

Last, it is clear that if $g \equiv 0$, then $s_1(\tilde{f}) = s_2(\tilde{f}) = 0$, and in this case $\mathcal{J}_{\tilde{f}}$ is a degenerated interval.

As we have pointed out in Theorem 3.1, the solvability of problem (3.2) may be, in some particular cases, like the case $g = g(u)$. The following result (where $g(-\infty) \neq g(+\infty)$ is allowed) shows that in other situations the conditions for solvability of (3.2) may be completely different with respect to the case $g = g(u)$ (see Proposition 6.4 in [2]).

THEOREM 3.2. *Let g be a bounded and continuous function satisfying*

$$g(0) = 0 < g(+\infty) + g(-\infty).$$

Then, there exist two numbers $g_1 \leq 0$, $g_2 \geq \sqrt{(2/\pi)}(g(+\infty) + g(-\infty))$ such that the BVP

$$\begin{aligned} u''(t) + u(t) + g(u'(t)) &= s\sqrt{\tfrac{2}{\pi}}\sin t, \quad t \in (0, \pi), \\ u(0) \;=\; u(\pi) &= 0 \end{aligned}$$

satisfies

(i) *If* $s \in [g_1, g_2)$, *it has a solution.*

(ii) *If* $s \notin [g_1, g_2]$, *it has no solution.*

Proof. As in Theorem 3.1, we have

$$\Gamma_0(c) \to \sqrt{\frac{2}{\pi}}(g(+\infty) + g(-\infty)) \text{ for } c \to \pm\infty.$$

Since $g(0) = 0$, $0 \in \Gamma_0(0)$, which proves that $g_1 \equiv \inf_{c \in \mathbb{R}} \Gamma_0(c) = s_1(0)$ verifies $g_1 \leq 0 < \sqrt{(2/\pi)}(g(+\infty) + g(-\infty))$. Again, as in Theorem 3.1, one may see that g_1 is achieved, i.e., $g_1 \in \Gamma_0(\mathbb{R})$. So, the theorem is proved if we denote $g_2 \equiv \sup_{c \in \mathbb{R}} \Gamma_0(c) = s_2(0)$. □

The case where $g_1 = 0$, $g_2 = \sqrt{(2/\pi)}(g(+\infty) + g(-\infty))$ is possible (see Theorem 3.1 and Example 3.1).

3.2. The Neumann problem. Let us study the BVP

$$\left\{ \begin{array}{l} u''(t) + g(u'(t)) = f(t), \ t \in (0,\pi), \\ u'(0) \ = u'(\pi) = 0. \end{array} \right. \tag{3.6}$$

In this case $\lambda_1 = 0$ and $u_1 \equiv 1/\sqrt{\pi}$. It should be mentioned here that the alternative method does not provide enough information concerning the structure of the "solution set" $\mathcal{J}_{\tilde{f}}$. This is due to the fact that u_1 is a constant function and consequently $u_1' = 0$, which means that we cannot obtain information from the alternative equation (2.4). Instead, we apply a change of variables and then a shooting method.

It is possible to observe immediately that the solution of the BVP (3.6) is invariant under the translation $v = u + c$, where $c \in \mathbb{R}$ is an arbitrary constant, i.e., u is the solution of (3.6) if and only if v is the solution of (3.6). Thus we can restrict ourselves to the functions with mean value zero. Let us denote $\tilde{C}^i(I) = \{u \in C^i(I); \int_0^\pi u(t)\,dt = 0\}$, $i = 0, 1, 2$ with the norm $\|\cdot\|_i$, where $\|\cdot\|_0$ means the uniform norm, $\|u\|_1 = \|u\|_0 + \|u'\|_0$, and $\|u\|_2 = \|u\|_0 + \|u'\|_0 + \|u''\|_0$. For $u \in \tilde{C}^2(I)$ let us introduce a new function, $w \in C^1(I)$, by $w(t) = u'(t)$. Then the BVP (3.6) transforms to

$$w'(t) + g(w(t)) = f(t), \ t \in (0,\pi), \ w(0) = w(\pi) = 0. \tag{3.7}$$

Let us split the right-hand side $f \in C(I)$ as follows:

$$f(t) = s + \tilde{f}(t), \ s \in \mathbb{R}, \ \tilde{f} \in \tilde{C}(I). \tag{3.8}$$

Define the map $\mathcal{F} : C^1(I) \times \mathbb{R} \times \tilde{C}(I) \to C^1(I) \times \mathbb{R}$ in the following way:

$$\mathcal{F}(w, s, \tilde{f}) = \begin{pmatrix} F_1(w, s, \tilde{f}) \\ F_2(w, s, \tilde{f}) \end{pmatrix} = \begin{pmatrix} w(t) - st - \int_0^t \tilde{f}(\tau)\,d\tau + \int_0^t g(w(\tau))\,d\tau \\ w(\pi) \end{pmatrix}.$$

Let us assume that $g : \mathbb{R} \to \mathbb{R}$ satisfies

$$|g(\xi)| \leq M \tag{3.9}$$

for any $\xi \in \mathbb{R}$ with some constant $M > 0$.

THEOREM 3.3. *Let g be a continuously differentiable function satisfying* (3.9) *and let $f(t)$ in* (3.6) *be of the form* (3.8). *Then for any $\tilde{f} \in \tilde{C}(I)$ there exists precisely one $s = s(\tilde{f})$ such that* (3.6) *has a solution. In this case, the Neumann BVP* (3.6) *has a family of solutions*

$$u_c(t) = u(t) + c;$$

$c \in \mathbb{R}$ is arbitrary, where

$$u(t) = \int_0^t w_{s(\tilde f)}(\tau)\,d\tau,$$

and $w_{s(\tilde f)}$ is the unique solution of (3.7). *Moreover, the map $\tilde C(I) \to \mathbb{R}$, $\tilde f \to s(\tilde f)$ is continuously differentiable and $s(\tilde f) \in [-M, M]$ for any $\tilde f \in \tilde C(I)$.*

Proof. The idea of the proof consists in the application of the implicit function theorem to $\mathcal{F}$. Note that $\mathcal{F}(w, s, f) = \boldsymbol{o}$ is equivalent to (3.7). Let us divide the proof into two steps.

Step 1. For any $\tilde f \in \tilde C(I)$ there exist precisely one $s = s(\tilde f) \in [-M, M]$ and $w = w_{s(\tilde f)} \in C^1(I)$ such that

$$\mathcal{F}(u_{s(\tilde f)}, s(\tilde f), \tilde f) = \boldsymbol{o}. \tag{3.10}$$

Consider the initial-value problem (IVP)

$$w'(t) = s + \tilde f(t) - g(w(t)), \qquad t \in [0, \pi], \qquad w(0) = 0, \tag{3.11}$$

with parameter $s \in \mathbb{R}$. The existence and uniqueness theorems and the continuous dependence of the solution to the IVP (3.11) on the parameter s (see, e.g., [7]) yield the following facts: the solution $w = w_s$ of the IVP (3.11) is uniquely determined and depends continuously on $s \in \mathbb{R}$. In particular, $\varphi(s) = w_s(\pi)$ is well defined and it is a continuous function. It follows from (3.11) that

$$w_s(t) = st + \int_0^t \tilde f(\tau)\,d\tau - \int_0^t g(w_s(\tau))\,d\tau, \qquad t \in [0, \pi],$$

and hence we have

$$\varphi(s) > 0 \text{ for } s > M \text{ and } \varphi(s) < 0 \text{ for } s < -M, \tag{3.12}$$

due to (3.9). To prove Step 1 it is sufficient to show that $\varphi'(s) > 0$ for any $s \in \mathbb{R}$. Really, then, this fact together with (3.12) and the continuity of φ imply the existence of a unique $s \in [-M, M]$ satisfying $\varphi(s) = w_s(\pi) = 0$. Let us fix $s_0 \in \mathbb{R}$. Since the right-hand side of (3.11) is a continuously differentiable function of variables s and w, we obtain (see [7]) that $w_s = w_s(t)$ is a continuously differentiable function of s and, moreover, $\frac{dw_s(t)}{ds}\big|_{s=s_0}$ is the solution of the linear IVP

$$\begin{aligned} z'(t) &= 1 - g'(w_{s_0}(t))z(t), \\ z(0) &= 0. \end{aligned} \tag{3.13}$$

Then

$$\left.\frac{dw_s(t)}{ds}\right|_{s=s_0} = \int_0^t e^{\int_0^\tau g'(w_{s_0}(\sigma))\,d\sigma}\,d\tau \cdot e^{-\int_0^t g'(w_{s_0}(\tau))\,d\tau}.$$

In particular, it follows from here that $\frac{dw_s(\pi)}{ds}\big|_{s=s_0} > 0$, i.e., $\varphi'(s)\big|_{s=s_0} > 0$. This completes the proof of Step 1.

Step 2. The assumptions of the implicit function theorem are satisfied at any point $(w, s, \tilde f)$ satisfying (3.10).

Let $(w_0, s_0, \tilde{f}_0) \in C^1(I) \times \mathbb{R} \times \tilde{C}(I)$ be a fixed point satisfying $\mathcal{F}(w_0, s_0, \tilde{f}_0) = \boldsymbol{o}$. Then the function $\mathcal{F}$ and the partial Fréchet differentials $\partial\mathcal{F}/\partial w, \partial\mathcal{F}/\partial s, \partial\mathcal{F}/\partial\tilde{f}$ are continuous in the neighborhood of $(w_0, s_0, \tilde{f}_0)$, and we have

$$\mathcal{T}(w,s) = \frac{\partial\mathcal{F}}{\partial(w,s)}\bigg|_{(w_0,s_0,\tilde{f}_0)} = \left.\begin{pmatrix} \dfrac{\partial F_1}{\partial w} & \dfrac{\partial F_1}{\partial s} \\[2ex] \dfrac{\partial F_2}{\partial w} & \dfrac{\partial F_2}{\partial s} \end{pmatrix}\right|_{(w_0,s_0,\tilde{f}_0)} \begin{pmatrix} w \\ s \end{pmatrix}$$
$$= \begin{pmatrix} w(t) + \int_0^t g'(w_0(\tau))w(\tau)\,d\tau - st \\ w(\pi) \end{pmatrix}.$$

It follows from here that $\mathcal{T} : C^1(I) \times \mathbb{R} \to C^1(I) \times \mathbb{R}$ is *a continuous linear map.* Let $(h, r) \in C^1(I) \times \mathbb{R}$. Then the equation $\mathcal{T}(w, s) = (h, r)^T$ is equivalent to the problem of finding the solution of the IVP

$$w'(t) + g'(w_0(t))w(t) = s + h'(t), w(0) = h(0) \tag{3.14}$$

satisfying

$$w(\pi) = r. \tag{3.15}$$

It is easy to see that the solution of (3.14) is of the form

$$w(t) = h(0)e^{-\int_0^t g'(w_0(\tau))\,d\tau} + se^{-\int_0^t g'(w_0(\tau))\,d\tau}\int_0^t e^{\int_0^\sigma g'(w_0(\tau))\,d\tau}\,d\sigma$$
$$+ e^{-\int_0^t g'(w_0(\tau))\,d\tau}\int_0^t h'(\sigma)e^{\int_0^\sigma g'(w_0(\tau))\,d\tau}\,d\sigma.$$

It follows from this expression that for any $(h, r) \in C^1(I) \times \mathbb{R}$ we can find a unique $s \in \mathbb{R}$ and the solution $w = w_s(t)$ of the IPV (3.14) satisfying (3.15). This proves the *surjectivity* of $\mathcal{T}$. It follows also from this explicit form of w that for $h(t) = 0, t \in [0, \pi]$, and $r = 0$, the relations (3.14) and (3.15) are verified only by $w(t) = 0, t \in [0, \pi]$, and $s = 0$. Thus $\mathcal{T}$ is *injective.* Applying the Banach open mapping theorem (see, e.g., [21]) we get that $\mathcal{T}^{-1}$ is continuous. Thus $\mathcal{T} : C^1(I) \times \mathbb{R} \to C^1(I) \times \mathbb{R}$ is an *isomorphism.* This completes the proof of Step 2.

The assertion of Theorem 3.3 follows immediately from Steps 1 and 2 and from the considerations at the beginning of this subsection. □

Remark 3.3. From the results of the previous two subsections we can assert that the structure of the range of the operator $u''(t) + \lambda_1 u(t) + g(u'(t))$ with the Dirichlet boundary conditions is completely different from the case of the Neumann ones.

Remark 3.4. Also, the qualitative structure of the range of the operator $u''(t) + g(u'(t))$ under the Neumann boundary conditions is quite distinct from the case $g = g(u)$ (see [17]), where, in many cases, infinitely many values of the parameter s exist for a given $\tilde{f} \in \tilde{C}(I)$, for which (3.6) has a solution.

3.3. The periodic problem.

In this subsection we will study the BVP

$$\begin{cases} u''(t) + g(u'(t)) = f(t), & t \in (0, \pi), \\ u(0) = u(\pi), & u'(0) = u'(\pi). \end{cases} \tag{3.16}$$

Again, the alternative method is not a useful tool to study this type of problem (realize that, as in the case of Neumann problems, the positive eigenfunction associated with the first eigenvalue, $\lambda_1 = 0$, is a constant function). We shall also apply a change of variables and a shooting method. In any event, the periodic boundary conditions make the problem a little more complicated than the Neumann ones, although the final conclusion about the range of the corresponding operator will be the same. By the same reasoning as in the previous subsection, we can restrict ourselves to the functions $u \in \tilde{C}^2(I)$, and by introducing the function $w(t) = u'(t)$, we transform the BVP (3.16) to the problem

$$(3.17) \qquad w'(t) + g(w(t)) = f(t), \quad t\in(0,\pi), \quad w(0) = w(\pi), \quad \int_0^\pi w(t)\,dt = 0.$$

Using the same notation as in the previous subsection let us define the map

$$\mathcal{G} : C^1(I) \times \mathbb{R}^2 \times \tilde{C}(I) \to C^1(I) \times \mathbb{R}^2$$

in the following way:

$$\mathcal{G}(w, s, \alpha, \tilde{f}) = \begin{pmatrix} G_1(w, s, \alpha, \tilde{f}) \\ G_2(w, s, \alpha, \tilde{f}) \\ G_3(w, s, \alpha, \tilde{f}) \end{pmatrix} = \begin{pmatrix} w(t) - st - \int_0^t \tilde{f}(\tau)\,d\tau + \int_0^t g(w(\tau))\,d\tau - \alpha \\ \int_0^\pi w(\tau)\,d\tau \\ w(\pi) - \alpha \end{pmatrix}.$$

THEOREM 3.4. *Let g be a continuously differentiable function satisfying* (3.9) *and let $f(t)$ in* (3.16) *be of the form* (3.8). *Then for any $\tilde{f} \in \tilde{C}(I)$ there exists precisely one $s = s(\tilde{f})$ such that* (3.16) *has a solution. In this case, the periodic BVP* (3.16) *has a family of solutions $u_c(t) = u(t) + c, c \in \mathbb{R}$ is arbitrary, where*

$$u(t) = \int_0^t w_{s(\tilde{f})}(\tau)\,d\tau,$$

and $w_{s(\tilde{f})}$ is the unique solution of (3.17). *Moreover, the map $\tilde{C}(I) \to \mathbb{R}$, $\tilde{f} \to s(\tilde{f})$ is continuously differentiable and $s(\tilde{f}) \in [-M, M]$ for any $\tilde{f} \in \tilde{C}(I)$.*

Proof. The idea is again to apply the implicit function theorem to $\mathcal{G}$ because the equation $\mathcal{G}(w, s, \alpha, \tilde{f}) = o$ is equivalent to (3.17).

Let us consider the IVP

$$(3.18) \qquad w'(t) = s + \tilde{f}(t) - g(w(t)), \qquad t \in [0, \pi], \qquad w(0) = \alpha,$$

with s and α as parameters. Integrating (3.18) we get the integral equation

$$(3.19) \qquad w(t) = st + \int_0^t \tilde{f}(\tau)\,d\tau - \int_0^t g(w(\tau))\,d\tau + \alpha, \qquad t \in [0, \pi]$$

for all solutions of the IVP (3.18). On the other hand, every solution of (3.19) (belonging to $C^1(I)$) is the solution of the IVP (3.18). Let $w = w_{s,\alpha}(t)$ be the solution of the IVP (3.18) and define $\tilde{G}_2, \tilde{G}_3 : \mathbb{R}^2 \to \mathbb{R}$ by

$$\tilde{G}_2(s, \alpha) = \int_0^\pi w_{s,\alpha}(\tau)\,d\tau \qquad \tilde{G}_3(s, \alpha) = w_{s,\alpha}(\pi) - \alpha.$$

The proof of the theorem follows in three steps.

Step 1. There exists at least one $(s, \alpha) \in \mathbb{R}^2$ such that

$$\tilde{G}_2(s, \alpha) = \tilde{G}_3(s, \alpha) = 0. \tag{3.20}$$

Observe that $\tilde{G}_2$ and $\tilde{G}_3$ are C^1 functions due to the continuous and differentiable dependence of the solution of the IVP (3.18) on the parameter and on the initial condition (see [7]). It follows from (3.9) and (3.19) that for any $\alpha \in \mathbb{R}$ and arbitrarily small $\varepsilon > 0$ we have $\tilde{G}_3(-M - \varepsilon, \alpha) < 0$ and $\tilde{G}_3(M + \varepsilon, \alpha) > 0$; therefore, for any $\alpha \in \mathbb{R}$ the equation

$$\tilde{G}_3(s, \alpha) = 0$$

has at least one solution $s \in [-M, M]$. As in the Neumann problem, one may prove that

$$\frac{\partial \tilde{G}_3(s, \alpha)}{\partial s} > 0 \qquad \text{for any } (s, \alpha) \in \mathbb{R}^2, \tag{3.21}$$

so that for any $\alpha \in \mathbb{R}$ there exists precisely one $s = s(\alpha) \in [-M, M]$ such that

$$\tilde{G}_3(s(\alpha), \alpha) = 0 \qquad \text{for any } \alpha \in \mathbb{R}.$$

Clearly, the mapping $\mathbb{R} \overset{\varphi_1}{\to} [-M, M]$, $\varphi_1(\alpha) = s(\alpha)$ is continuous.

On the other hand, for any fixed $s \in \mathbb{R}$ we get a constant $\alpha_{s,\tilde{f}} > 0$ such that $\tilde{G}_2(s, -\alpha_{s,\tilde{f}}) < 0$ and $\tilde{G}_2(s, \alpha_{s,\tilde{f}}) > 0$; therefore, for any $s \in \mathbb{R}$ the equation

$$\tilde{G}_2(s, \alpha) = 0$$

has at least one solution. Moreover the function $v_0(t) \equiv \frac{\partial w_{s,\alpha}(t)}{\partial \alpha} |_{\alpha = \alpha_0}$, $t \in [0, \pi]$, is the solution of the linear IVP

$$v'(t) + g'(w_{s,\alpha_0}(t))v(t) = 0, \quad t \in [0, \pi], \quad v(0) = 1$$

(see [7]), and it is easy to see that $v_0(t) > 0$ for $t \in [0, \pi]$. In particular this implies that

$$\frac{\partial \tilde{G}_2(s, \alpha)}{\partial \alpha} > 0 \text{ for any } (s, \alpha) \in \mathbb{R}^2, \tag{3.22}$$

so that for any $s \in \mathbb{R}$ there exists a unique $\alpha = \alpha(s)$ such that

$$\tilde{G}_2(s, \alpha(s)) = 0 \text{ for any } s \in \mathbb{R}.$$

Also, the mapping $\mathbb{R} \overset{\varphi_2}{\to} \mathbb{R}$, $\varphi_2(s) = \alpha(s)$ is continuous. Now, if we consider the continuous mapping $\varphi_1 \circ \varphi_2 : [-M, M] \to [-M, M]$, then there is at least one point $s_0 \in [-M, M]$ such that $\varphi_1(\varphi_2(s_0)) = s_0$. Lastly, $(s_0, \varphi_2(s_0))$ satisfies (3.20).

Step 2. The point (s, α) with the property (3.20) is unique.

We argue by contradiction. Let (s_1, α_1) and (s_2, α_2) be two points satisfying (3.20). We can assume, without loss of generality, that $\alpha_1 > \alpha_2$. Then due to the fact that both $w_{s_i,\alpha_i}, i = 1, 2$ must solve (3.17), we have to find two points $t_1, t_2 \in (0, \pi)$ such that

$$w_{s_1,\alpha_1}(t_i) = w_{s_2,\alpha_2}(t_i), \ i = 1, 2$$

and

$$w'_{s_1,\alpha_1}(t_1) \le w'_{s_2,\alpha_2}(t_1), \qquad w'_{s_1,\alpha_1}(t_2) \ge w'_{s_2,\alpha_2}(t_2).$$

Then the first inequality implies $s_1 \le s_2$ and the second one implies $s_1 \ge s_2$. Hence $s_1 = s_2$, but this contradicts $\alpha_1 > \alpha_2$ and (3.22).

Step 3. The assumptions of the implicit function theorem are satisfied at any point $(w_0, s_0, \alpha_0, \tilde{f}_0) \in C^1(I) \times \mathbb{R}^2 \times \tilde{C}(I)$ satisfying

$$\mathcal{G}(w_0, s_0, \alpha_0, \tilde{f}_0) = \boldsymbol{o}.$$

The function $\mathcal{G}$ and the partial Fréchet differentials $\frac{\partial \mathcal{G}}{\partial w}$, $\frac{\partial \mathcal{G}}{\partial s}$, $\frac{\partial \mathcal{G}}{\partial \alpha}$, and $\frac{\partial \mathcal{G}}{\partial f}$ are continuous in the neighborhood of $(w_0, s_0, \alpha_0, \tilde{f}_0)$. We have

$$\mathcal{S}(w,s,\alpha) = \left.\frac{\partial \mathcal{G}}{\partial(w,s,\alpha)}\right|_{(w_0,s_0,\alpha_0,\tilde{f}_0)} = \begin{pmatrix} w(t) + \int_0^t g'(w_0(\tau))w(\tau)\,d\tau - st - \alpha \\ \int_0^\pi w(\tau)\,d\tau \\ w(\pi) - \alpha \end{pmatrix}.$$

Then $\mathcal{S} : C^1(I) \times \mathbb{R}^2 \to C^1(I) \times \mathbb{R}^2$ is a *continuous linear map.* Let $(h, \beta, r) \in C^1(I) \times \mathbb{R}^2$. Then the equation $\mathcal{S}(w,s,\alpha) = (h,\beta,r)^T$ is equivalent to the problem of finding the solution of the IVP

$$w'(t) + g'(w_0(t))w(t) = s + h'(t), \ w(0) = \alpha + h(0) \tag{3.23}$$

satisfying

$$\int_0^\pi w(\tau)\,d\tau = \beta, \ w(\pi) = \alpha + r. \tag{3.24}$$

Since the solution of the linear IVP (3.23) is of the explicit form

$$w(t) = [\alpha + h(0)]e^{-\int_0^t g'(w_0(\tau))\,d\tau} + e^{-\int_0^t g'(w_0(\tau))\,d\tau} \cdot \int_0^t h'(\sigma)e^{\int_0^\sigma g'(w_0(\tau))\,d\tau}\,d\sigma$$

$$+ s{\cdot}e^{-\int_0^t g'(w_0(\tau))\,d\tau} \cdot \int_0^t e^{\int_0^\sigma g'(w_0(\tau))\,d\tau}\,d\sigma,$$

we derive from here that

(i) for any $(h, \beta, r) \in C^1(I) \times \mathbb{R}^2$ we can find $(s, \alpha) \in \mathbb{R}^2$ and the solution $w = w_{s,\alpha}(t)$ of the IVP (3.23) satisfying (3.24);

(ii) for $h(t) = 0$, $t \in [0, \pi]$, $r = \beta = 0$, the relations (3.23) and (3.24) are verified only by $w(t) = 0$, $t \in [0, \pi]$, and $s = \alpha = 0$.

It follows from (i) that $\mathcal{S}$ is *surjective* and from (ii) that $\mathcal{S}$ is *injective.* Then $\mathcal{S}$ is an *isomorphism* due to the Banach open mapping theorem. This completes the proof of Step 3 and also the proof of Theorem 3.4. □

Remark 3.5. Taking into account the results obtained in previous subsections we can affirm that the structure of the range of the operator $u''(t)+\lambda_1 u(t)+g(u'(t))$ under periodic boundary conditions is similar to the Neumann case and completely different from the case of Dirichlet boundary conditions. Also, the qualitative structure of such a range is quite distinct from the case $g = g(u)$ (see [16], [12]) where, in many cases, infinitely many values of the parameter s exist for a given $\tilde{f} \in \tilde{C}(I)$ for which (3.16) has a solution.

4. Concluding remarks, possible extensions, and generalizations.

Remark 4.1. Obviously, the results of Theorems 3.1–3.4 should be formulated also for nonlinearities of the type $g = g(t, u')$ (with g depending also on $t \in [0, \pi]$).

Remark 4.2. It follows from the results in §3 that the "solution set" $\mathcal{J}_{\tilde{f}}$ from Proposition 2.1 has a structure that depends on the nonlinearity of g and on the type of boundary conditions. In the case of a Dirichlet problem adding some appropriate asymptotic conditions on g, it is, in general, an interval with a *nonempty interior*, while in the case of a Neumann (or periodic) problem, it is a point without any assumptions on the asymptotic behavior of g.

Remark 4.3. Let us point out that further assumptions on g may allow us to get more information about the behavior of the maps $\tilde{f} \to s_i(\tilde{f})$, $i = 1, 2$ (e.g., the limits of $s_i(\tilde{f})$ for $\| \tilde{f} \| \to 0$ or $\| \tilde{f} \| \to \infty$, etc.).

Remark 4.4. In the case of the Dirichlet BVP we restrict ourselves only to the case of bounded nonlinearity with the *finite limits* in $\pm\infty$. We had several reasons to do so:

(i) to avoid tedious calculations and to keep the idea clear,

(ii) to give not only sufficient but also necessary conditions for the solvability,

(iii) to emphasize the difference between the BVP (1.1) and the "classical" Landesman–Lazer-type problem with nonlinearity depending only on the solution u.

However, by using the *Fatou lemma* instead of the *Lebesgue theorem* in the proof of Theorem 3.1 one can deal with more general nonlinearities (e.g., unbounded ones) and formulate sufficient conditions for the solvability of the BVP (3.3) in terms of $\liminf_{\xi \to \pm\infty} g(\xi)$ and $\limsup_{\xi \to \pm\infty} g(\xi)$. It should also be mentioned that the *alternative method does not provide enough information* concerning the *solvability* of *Neumann* and *periodic* problems. That is why we use a different approach based on the implicit function theorem in these cases.

Remark 4.5. It should be interesting to study related problems for *higher-order equations* or for the *systems of equations.* One can expect that similar results also hold true in the case of boundary value problems for *partial differential equations.* However, a *different approach* should be applied because our method works only in the case of ordinary differential equations. Another possible extension of our result is to consider the solvability of the BVP (1.1) with λ_n instead of λ_1, with general $n \geq 1$.

Acknowledgment. P. Drábek is grateful to Universidad Granada for the pleasant hospitality during preparation of this paper.

REFERENCES

[1] H. Amann, A. Ambrosetti, and G. Mancini, *Elliptic equations with noninvertible Fredholm linear part and bounded nonlinearities*, Math. Z., 158 (1978), pp. 179–194.

[2] A. Ambrosetti and G. Mancini, *Existence and multiplicity results for nonlinear elliptic problems with linear part at resonance*, J. Differential Equations, 28 (1978), pp. 220–245.

[3] A. Cañada, *Existencia de soluciones para problemas de contorno elipticos no lineales y no necesariamente autoadjuntos*, C.E.D.Y.A., Granada, 1984, pp. 79–84.

[4] ———, *Nonselfadjoint semilinear elliptic boundary value problems*, Ann. Mat. Pura Appl. (4), CXLVIII (1987), pp. 237–250.

[5] C. Cesari and P. Pucci, *Existence theorems for nonselfadjoint semilinear elliptic boundary value problems*, Nonlinear Anal., 9 (1985), pp. 1227–1241.

[6] L. Cesari and T. T. Bowman, *Existence of solutions to nonselfadjoint boundary value problems for ordinary differential equations*, Nonlinear Anal., 9 (1985), pp. 1211–1225.

[7] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, Toronto, London, 1955.

[8] P. Drábek, *Solvability and bifurcations of nonlinear equations*, Pitman Res. Notes Math. Ser., 264, Longman, London, 1992.

[9] P. Drábek and F. Nicolosi, *Semilinear boundary value problems at resonance with general nonlinearities*, Differential Integral Equations, 5 (1992), pp. 339–355.

[10] S. Fučík; *Solvability of Nonlinear Equations and Boundary Value Problems*, D. Reidel, Boston, MA, Dordrecht, The Netherlands, North–Holland, Amsterdam, 1980.

[11] W. Jordan and P. Smith, *Nonlinear Ordinary Differential Equations*, Oxford University Press, London, 1977.

[12] R. Kannan and R. Ortega, *An asymptotic result in forced oscillations of pendulum-type equations*, Appl. Anal., 22 (1986), pp. 45–53.

[13] E. Landesman and A. C. Lazer, *Nonlinear perturbations of linear boundary value problems at resonance*, J. Math. Mech., 19 (1970), pp. 609–623.

[14] B. J. Lazan, *Damping of Materials and Members in Structural Mechanics*, Pergamon Press, Elmsford, NY, 1968.

[15] A. C. Lazer and D. E. Leach, *Bounded perturbations of forced harmonic oscillators at resonance*, Ann. Mat. Pura Appl., 82 (1969), pp. 49–68.

[16] J. Mawhin and M. Willem, *Multiple solutions of the periodic boundary value problem for some forced pendulum-type equations*, J. Differential Equations, 52 (1984), pp. 264–287.

[17] J. Mawhin, *Problèmes aux limites du type de Neumann pour certaines équations différentielles on aux dérivées partielles non linéaires*, in Equations différentielles et fonctionnelles non linéaires, P. Janssens, J. Mawhin, and M. Rouche, eds., Hermann, Paris, 1973, pp. 124–134.

[18] N. Minorsky, *Nonlinear Oscillations*, Van Nostrand, Princeton, NJ, 1962.

[19] K. Nagle, K. Pothoven, and K. Singkoffer, *Nonlinear elliptic equations at resonance where the nonlinearity depends essentially on the derivatives*, J. Differential Equations, 38 (1980), pp. 210–225.

[20] A. H. Nayfech and D. T. Mook, *Nonlinear Oscillations*, John Wiley, New York, 1979.

[21] A. Taylor, *Introduction to Functional Analysis*, John Wiley, New York, 1958.

SIAM J. MATH. ANAL.
Vol. 27, No. 2, pp. 558–577, March 1996

013

TRACKING INVARIANT MANIFOLDS UP TO EXPONENTIALLY SMALL ERRORS*

C. K. R. T. JONES[†], TASSO J. KAPER[‡], AND NANCY KOPELL[§]

Abstract. This work establishes a new tool for proving the existence of multiple-pulse homoclinic orbits in perturbed Hamiltonian systems and general multidimensional singular-perturbation problems. The center-stable and center-unstable manifolds of slow manifolds in these problems intersect transversely at angles that are of the same order as the asymptotically small parameter in the problem, which can be either an amplitude or a frequency. To deal with the difficulties associated with small angles of intersection, we develop the exchange lemma with exponentially small error (ELESE), which is the main technical result of this work. This lemma enables highly accurate tracking of invariant manifolds while orbits on them spend long intervals of time near slow manifolds.

Key words. multiple-pulse heteroclinic orbits, singular perturbations, Fenichel coordinates, tracking invariant manifolds, exchange lemma, Hamiltonian systems

AMS subject classifications. 58F30, 34C37, 70H05, 34E15

1. Introduction.

Homoclinic orbits play central roles in many systems. For reaction-diffusion systems and nerve impulse equations, for example, travelling waves are realized as homoclinic orbits in associated systems of ordinary differential equations. In addition, for finite-dimensional systems, transverse homoclinic and heteroclinic orbits are the source of Smale horseshoe chaos in the ambient phase space.

In singularly perturbed systems with two asymptotically distinct time scales, one can construct formal homoclinic orbits in the limit in which the small parameter vanishes. These orbits are formal in that they consist of a finite number of fast $\mathcal{O}(1)$-time-duration jumps, or "pulses," between manifolds of critical points interspersed with appropriate slow $\mathcal{O}(\frac{1}{\varepsilon})$-time-duration orbit segments on those manifolds, where $0 < \varepsilon \ll 1$. These jumps and orbit segments are obtained from the "fast" and "slow" limiting version of the equations, respectively, which are individually easier to study than the entire system. One of the major goals of geometric singular-perturbation theory is to establish conditions under which such formally constructed singular homoclinic and heteroclinic orbits perturb to real orbits for small, nonzero values of the perturbation parameter.

The problem of constructing heteroclinic orbits in the context of singularly perturbed systems has been attacked with many techniques. Topological methods for proving the existence of orbits homoclinic to fixed points near a singular limit were pioneered in [5]. The results of [5] concerned the Fitzhugh–Nagumo and Hodgkin–Huxley systems, in which the homoclinic orbit models the travelling nerve impulse. Similar, but more analytic, methods were simultaneously developed in [17]. The authors of [14] extended the topological methods to systems with more diffusing variables. A combination of analytic and other techniques have been developed by many

* Received by the editors December 9, 1993; accepted for publication (in revised form) June 14, 1994.

[†] Division of Applied Mathematics, Brown University, Providence, RI 02912. The research of this author was partially supported by National Science Foundation grant DMS-9100085 and by Office of Naval Research grant N00014-92-J-1401.

[‡] Mathematics Department, Boston University, Boston, MA 02215. The research of this author was partially supported by National Science Foundation grant DMS-9307074 and the Lefschetz Center for Dynamical Systems during 1991 and 1992, when this work was done.

[§] Mathematics Department, Boston University, Boston, MA 02215. The research of this author was partially supported by National Science Foundation grant DMS-9200131.

authors for determining homoclinic orbits and related structures in singular systems; among these authors' works are [6]–[8], [9], [15], [19], [28], and [33]. We refer the reader to [34] for a more complete list of references.

Recently, two of us established a technical tool known as the exchange lemma for tracking invariant manifolds in singularly perturbed systems; see [21]. This work was motivated by a large class of travelling-wave problems. The exchange lemma is a general tool that makes it possible to demonstrate the existence of multiple-pulse orbits homoclinic to fixed points even when these points lie on higher-dimensional center manifolds. The pulses are the fast excursions the orbit makes going from a neighborhood of one slow manifold to a neighborhood of another slow manifold. These results are obtained under the assumption that certain transversality conditions hold between the center-stable and center-unstable manifolds of the slow manifolds in the space of the dependent variables and the wave speed. In particular, the relevant intersections of the center-stable and center-unstable manifolds of the slow manifolds occur at an angle of $\mathcal{O}(1)$.

There are many interesting classes of singularly perturbed systems, however, in which the center-stable and center-unstable manifolds of invariant sets on the slow manifolds coincide in the $\varepsilon = 0$ limit of the equations, so that the leading $\mathcal{O}(1)$ term of an expansion of the distance between the perturbed stable and unstable manifolds is identically zero. These systems include perturbed Hamiltonian problems and coupled travelling wave problems. A partial listing includes [1]–[4], [10]–[12], [16], [23]–[27], and [29]–[32]. See also Chapter 4 of [35]. In these systems, the perturbations lift the degeneracy at $\varepsilon = 0$, and the stable and unstable manifolds intersect at an $\mathcal{O}(\varepsilon)$ angle.

Our goal in this work is to produce a technique that is applicable to a wide range of problems, such as those listed above. We call this technique the exchange lemma with exponentially small error (henceforth referred to by the acronym ELESE). The sharpness of the error estimates given by ELESE enables us to overcome the previously unsurmounted difficulties associated with transverse intersections at angles of order $\mathcal{O}(\varepsilon)$ between the relevant center-stable and center-unstable manifolds. Moreover, our technique also applies to problems in which the angle of intersection is order $\mathcal{O}(\varepsilon^{\alpha})$ for any $\alpha \geq 0$.

As introduced in [20] and [21], the exchange lemma tracks invariant manifolds while trajectories on them spend long intervals of time near slow manifolds. Consider a trajectory which lies on a locally invariant manifold M and which spends an $\mathcal{O}(\frac{1}{\varepsilon})$ interval of time in the neighborhood of a slow manifold S. In order for an orbit to spend an $\mathcal{O}(\frac{1}{\varepsilon})$ amount of time in this neighborhood, it must be exponentially close to the stable manifold of S on entry into the neighborhood and for at least half of the time interval it is in the neighborhood. Now, if one assumes that M is transverse to the stable manifold of S at the point where the orbit enters the neighborhood; then the exchange lemma states that at the point where the orbit exits the neighborhood, M is C^1–$\mathcal{O}(\varepsilon)$ close to the unstable manifold of S restricted to an orbit segment on S. In particular, at the exit point, the tangent hyperplane to M is $\mathcal{O}(\varepsilon)$ close to the space spanned by the unstable directions and by the one center direction tangent to the restricted flow on the slow manifold.

As a consequence of the exchange lemma, we know that when the unstable manifold of one slow manifold, restricted to a slow orbit segment, transversely intersects at an angle of $\mathcal{O}(1)$ the stable manifold of another (not necessarily distinct) slow manifold, then the tracked manifold M will also be transverse to the stable manifold of this second slow manifold. In this sense, transversality information is exchanged. By induction, this process can be repeated finitely many times to study homoclinic orbits with finitely many pulses. In particular, when orbits on the tracked manifold M leave the neighborhood of the second slow manifold, M and its tangent plane are exponentially close to the unstable manifold of that second slow manifold restricted to the appropriate slow orbit segment. Thus, since the restricted unstable manifold transversely intersects the stable manifold of a third slow manifold, the tracked manifold will as well.

The exchange lemma enables one to capture the dynamics in all of the fast and slow directions. By contrast, following the tangents to individual trajectories one loses information about the dynamics in the center directions because all trajectories exit the neighborhood of the slow manifold approximately tangent to the fast unstable space.

By establishing ELESE, we improve on the result of the exchange lemma. In particular, we demonstrate that, at the point at which the orbit on M exits the neighborhood, M is actually C^1– $\mathcal{O}(e^{-\frac{c}{\varepsilon}})$ close to the unstable manifold of S restricted to an orbit segment on S. Therefore, when the unstable manifold of one slow manifold transversely intersects at an angle of $\mathcal{O}(\varepsilon^{\alpha})$ the stable manifold of another (not necessarily distinct) slow manifold for any $\alpha \geq 0$, then the tracked manifold M will also be transverse to the stable manifold of this second slow manifold. In fact, ELESE is precisely the result needed to establish the existence of multipulse orbits in the problems listed above. Finally, the sets to which these multipulse orbits are asymptotic can consist of unstable equilibria, periodic orbits, tori, or more complicated invariant sets on the slow manifolds. They can also be an entire slow manifold.

The improvements obtained here are made without requiring further assumptions on the equations (see (1) below) under consideration. Rather, we take fuller advantage of the structure in the normal form for the flow near a slow normally hyperbolic invariant manifold. See equations (4) and the discussion in §2.

Finally, in order to illustrate the technique, we use ELESE to obtain the existence of multiple-pulse heteroclinic orbits in a model problem which consists of a quasiperiodic, slowly modulated, nonlinear pendulum. In particular, we demonstrate the existence of orbits that (i) are forward and backward asymptotic to periodic orbits on the two-dimensional tori that are the slow manifolds in this problem and (ii) have finitely many pulses, or fast excursions from one slow torus to another. Furthermore, in between each pair of adjacent excursions, these heteroclinic orbits spend long $\mathcal{O}(\frac{1}{\varepsilon})$ intervals of time near the slow torus.

The central feature of this model is that the stable and unstable manifolds of the slow manifolds intersect transversely at an angle of $\mathcal{O}(\varepsilon)$. Hence, error estimates that are sharper than those of existing work are necessary in order to obtain useful bounds from tracking the invariant manifolds during their long passages near slow manifolds. The exponentially small error estimates of ELESE are more than sharp enough.

Geometric simplicity guided the construction of this model. In particular, the number of dimensions and the qualitative features have been kept to a minimum. The model is intended only to elucidate the essential ideas. We refer the reader to the references cited above for several important problems exhibiting $\mathcal{O}(\varepsilon)$ and $\mathcal{O}(\varepsilon^2)$ transversality in which ELESE can be used.

This paper is organized as follows. In §2, we state the equations under study and their normal forms near the slow manifolds. In §3, we establish a C^0 version of the main result. Then, using differential forms, we prove the full ELESE in §4. Finally, the example is discussed in §5.

2. Fenichel coordinates. We study systems with two time (or length) scales for which the governing equations have the form

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}, \mathbf{y}, \varepsilon),$$

$$\mathbf{y}' = \varepsilon \mathbf{g}(\mathbf{x}, \mathbf{y}, \varepsilon), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$ with $n > 0$ and $\mathbf{f}$ and $\mathbf{g}$ are smooth.

The unperturbed equations, (1) with $\varepsilon = 0$, are autonomous and are parametrized by the $\mathbf{y}$ variables. The relation $\mathbf{f}(\mathbf{x}, \mathbf{y}, 0) = \mathbf{0}$ determines a manifold of equilibria. For this work, we are interested in the case in which some subset of this manifold is normally hyperbolic; i.e., we assume that there exists a subset S_0 of the manifold of equilibria such that the eigenvalues of $D_{\mathbf{x}}\mathbf{f}(\mathbf{p}, 0)$ have nonzero real parts for all points $\mathbf{p} \in S_0$. The reason for this choice is that a normally hyperbolic manifold S_0 persists as an invariant manifold S_ε in the full system (1) along with its local stable and unstable manifolds for $0 < \varepsilon \ll 1$. See [13] and [18]. Furthermore, S_ε lies $\mathcal{O}(\varepsilon)$ close to S_0 and may be expressed as a graph over S_0 when the eigenvalues of $D_{\mathbf{x}}\mathbf{f}(\mathbf{p}, 0)$ are uniformly bounded away from zero for all $\mathbf{p} \in S_0$. In addition, the perturbed local stable and unstable manifolds of S_ε lie $\mathcal{O}(\varepsilon)$ close to their unperturbed counterparts. Finally, Fenichel theory states that the vector field restricted to S_ε has magnitude $\mathcal{O}(\varepsilon)$, which leads to the label slow manifold for S_ε.

We briefly discuss the structure of the unperturbed vector field near S_0. For a fixed $\mathbf{y}$, the local stable and unstable manifolds of S_0 give rise to a new coordinate system in which the dynamics near S_0 are best studied. Let $\mathbf{a} \in \mathbb{R}^{\mathbf{k}}$ and $\mathbf{b} \in \mathbb{R}^{\ell}$, where $k + \ell = m$, be a new coordinate system such that the local stable and unstable manifolds of S_0 are the axes $\mathbf{a} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$, respectively. Then the equations are

$$\mathbf{a}' = \Lambda(\mathbf{a}, \mathbf{b}, \mathbf{y})\mathbf{a},$$

$$\mathbf{b}' = \Gamma(\mathbf{a}, \mathbf{b}, \mathbf{y})\mathbf{b},$$

$$\mathbf{y}' = \mathbf{0}, \tag{2}$$

where Λ and Γ are matrix-valued functions. The normal hyperbolicity of S_0 implies that for any $\Delta > 0$ sufficiently small, there are constants λ_0 and γ_0 such that for any eigenvalue $\tilde{\lambda}_i$ of $\Lambda(\mathbf{0}, \mathbf{0}, \mathbf{y})$ or any eigenvalue $\tilde{\gamma}_i$ of $\Gamma(\mathbf{0}, \mathbf{0}, \mathbf{y})$, we have $\text{Re}\tilde{\lambda}_i > \lambda_0 > 0$ and $\text{Re}\tilde{\gamma}_i < \gamma_0 < 0$ for all points in the box $\mathcal{B} \equiv \{(\mathbf{a}, \mathbf{b}, \mathbf{y}) | |\mathbf{a}|, |\mathbf{b}| \leq \Delta\}$, where the $\mathbf{y}$ variables lie in a compact subset of $\mathbb{R}^{\mathbf{n}}$.

The persistence theory cited above gives an explicit constructon of the perturbed counterpart of this local coordinate system. In particular, there exist $(\mathbf{a}, \mathbf{b}, \mathbf{y})$ coordinates, which have been dubbed Fenichel coordinates (see [21]) such that the perturbed local stable and unstable manifolds correspond to the manifolds $\mathbf{a} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$, respectively. Furthermore, these manifolds are $\mathcal{O}(\varepsilon)$ close, as functions of the variables $(\mathbf{x}, \mathbf{y}, \varepsilon)$, to the corresponding unperturbed manifolds. Using equations (11.6)(a) and (11.7)(a) from Theorem 11.1 of [13], the equations near S_ε may be written as

$$\mathbf{a}' = \Lambda(\mathbf{a}, \mathbf{b}, \mathbf{y}, \varepsilon)\mathbf{a},$$

$$\mathbf{b}' = \Gamma(\mathbf{a}, \mathbf{b}, \mathbf{y}, \varepsilon)\mathbf{b},$$

$$\mathbf{y}' = \varepsilon \mathbf{m}(\mathbf{y}, \varepsilon) + \varepsilon \mathbf{h}(\mathbf{a}, \mathbf{b}, \mathbf{y}, \varepsilon)\mathbf{ab}, \tag{3}$$

where $\mathbf{h}(\mathbf{a}, \mathbf{b}, \mathbf{y}, \varepsilon)$ is a rank-three tensor. In component notation, the third equation is $y_i{}' = \varepsilon m_i(\mathbf{y}, \varepsilon) + \varepsilon \sum_{m=1}^{\ell} \sum_{j=1}^{k} h_{ijm} a_j b_m$.

The terms $\mathbf{a}$ and $\mathbf{b}$ factor out in the slow ($\mathbf{y}$) equations in (3) because the coordinate system is determined by the perturbed center manifold and its fast stable and unstable foliations. Furthermore, it is precisely this factoring that makes the tracking with exponentially small error possible, since the product of these two terms remains exponentially small during an orbit's $\mathcal{O}(\frac{1}{\varepsilon})$ long time of passage near a slow manifold.

Finally, we make a technical refinement of (3). If one first rectifies $\mathcal{S}_\varepsilon$ in (1) (taking into account that this may have to be done using more than one chart) so that the slow flow of (1) is in the y_1 direction, as is done in [21], then the normal form has the following useful structure:

$$\begin{aligned} \mathbf{a}' &= \Lambda(\mathbf{a}, \mathbf{b}, \mathbf{y}, \varepsilon)\mathbf{a}, \\ \mathbf{b}' &= \Gamma(\mathbf{a}, \mathbf{b}, \mathbf{y}, \varepsilon)\mathbf{b}, \\ \mathbf{y}' &= \varepsilon \left(\mathbf{U} + \mathbf{h}(\mathbf{a}, \mathbf{b}, \mathbf{y}, \varepsilon)\mathbf{a}\mathbf{b}\right), \end{aligned} \tag{4}$$

where $\mathbf{U} = (1, 0, \ldots, 0)$. This is the normal form we shall use throughout the paper.

Remark 1. In October 1992, Fenichel pointed out to one of us (C. Jones) that the C^0 and C^1 ELESE of §§3 and 4 also hold when the factor of ε in front of the second term of the $\mathbf{y}'$ equations in (3) is not present.

Remark 2. The technical refinement used in (4) is made without loss of generality. The flow on $\mathcal{S}_\varepsilon$ can, and in general does, exhibit nontrivial dynamics, including fixed points, periodic orbits, homoclinic orbits, etc.

3. The C^0 ELESE. In this section, we establish the C^0 version of the main result, namely that, in the coordinates of (4), the $\mathbf{b}$ coordinates are exponentially close in ε to $\mathbf{b} = \mathbf{0}$ and that the y_i ($i > 0$) coördinates are exponentially close in ε to the values of y_i they have on entry into $\mathcal{B}$.

PROPOSITION 3.1. *Let q be a point in $M \cap \{|\mathbf{b}| = \Delta\}$ whose trajectory exits $M \cap \{|\mathbf{a}| = \Delta\}$ at $\bar{q}$ after a time τ that is $\mathcal{O}(\frac{1}{\varepsilon})$. Let V be a neighborhood of q in M. Then for V sufficiently small, the image of V under the time τ map is $\mathcal{O}(e^{-\frac{c}{\varepsilon}})$ C^0 close for some $c > 0$ to $\{|\mathbf{b}| = 0, y_i - y_i(0) = 0, i > 1\}$, where $y_i(0)$ denotes the value of y_i at q.*

Remark. $\mathcal{O}(\varepsilon)$ C^0-closeness is established in Corollary 3.1 of [21]. See also [6] and [8]. Here we are able to refine the estimate since both an $\mathbf{a}$ and a $\mathbf{b}$ factor out in the normal form (4).

Before proving this proposition, we state a technical result which gives bounds on the rate of growth and decay of the variables $\mathbf{a}$ and $\mathbf{b}$, respectively, while the orbit is in $\mathcal{B}$. Most importantly, the proposition shows that the integral of the product of $|\mathbf{a}|$ and $|\mathbf{b}|$ over any subinterval of the time a trajectory is in $\mathcal{B}$ is exponentially small. This proof uses the fact that solutions spending long times in $\mathcal{B}$ must have exponentially small $\mathbf{a}$ components for at least the first half of the subinterval and exponentially small $\mathbf{b}$ components for at least the second half of the subinterval, so that the product of $\mathbf{a}$ and $\mathbf{b}$ is exponentially small during the orbits entire stay in $\mathcal{B}$.

LEMMA 3.1. *For ε sufficiently small, there exist positive constants c_a, c_b, c, and K such that, for $s \le t$,*

$$|\mathbf{b}(t)| \le c_b |\mathbf{b}(s)| e^{\gamma_0 (t-s)}, \tag{i}$$

(ii) $$|\mathbf{a}(t)| \geq c_a |\mathbf{a}(s)| e^{\lambda_0 (t-s)},$$

(iii) $$\int_s^t |\mathbf{a}(\zeta)||\mathbf{b}(\zeta)| d\zeta \leq K e^{c(s-t)},$$

independently of ε, *where* $K = 2\Delta^2 \max(\frac{1}{c_a \lambda_0}, \frac{c_b}{|\gamma_0|})$ *and* $c = \frac{1}{2}\min(\lambda_0, |\gamma_0|)$, *so long as the trajectory stays in* $\mathcal{B}$.

Remark. Parts (i) and (ii) of this lemma are identical to parts (i) and (ii) of Proposition 3.1 in [21], and so we refer the reader there for their proofs. We prove part (iii) here.

Proof of part (iii). Split the integral on the left hand side of (iii) in two pieces: (I) from s to $\frac{s+t}{2}$ and (II) from $\frac{s+t}{2}$ to t. For (I), we first rewrite part (ii) of the lemma as

(5) $$|\mathbf{a}(\zeta)| \leq \frac{|\mathbf{a}(t)|}{c_a} e^{\lambda_0(\zeta - t)},$$

where $s \leq \zeta \leq t$. Then, using (5), $|\mathbf{b}(\zeta)| \leq \Delta$, and $|\mathbf{a}(t)| \leq \Delta$, we get

(6) $$\int_s^{\frac{s+t}{2}} |\mathbf{a}(\zeta)||\mathbf{b}(\zeta)| d\zeta \leq \frac{\Delta^2}{c_a} \int_s^{\frac{s+t}{2}} e^{\lambda_0(\zeta - t)} d\zeta = \frac{\Delta^2}{c_a \lambda_0} e^{\frac{\lambda_0}{2}(s-t)} \left[1 - e^{\frac{\lambda_0}{2}(s-t)}\right].$$

Also, since we have $|\mathbf{b}(\zeta)| \leq c_b |\mathbf{b}(s)| e^{\gamma_0(\zeta - s)}$ from using (i) with $s \leq \zeta \leq t$, the second part of the integral (II) yields

(7) $$\int_{\frac{s+t}{2}}^t |\mathbf{a}(\zeta)||\mathbf{b}(\zeta)| d\zeta \leq \frac{c_b \Delta^2}{|\gamma_0|} e^{\frac{\gamma_0}{2}(t-s)} \left[1 - e^{\frac{\gamma_0}{2}(t-s)}\right].$$

Putting (6) and (7) together proves the lemma.

Remark. As a consequence of (5), the fact that $|\mathbf{a}|$ is bounded by Δ at time t implies that $|\mathbf{a}|$ is exponentially small at least in the first half of the interval. Similarly, the fact that $|\mathbf{b}(\zeta)|$ is bounded by Δ at time s implies that $|\mathbf{b}|$ is exponentially small during at least the second half of the time interval.

This technical lemma leads to the following proof.

Proof of Proposition 3.1. The variable $\mathbf{b}(t)$ is exponentially small at $\bar{q}$ by part (i) of Lemma 3.1 because the trajectory spends an $\mathcal{O}(\frac{1}{\varepsilon})$ amount of time in $\mathcal{B}$. Hence, we only need to consider y_i for $i > 1$. From (4), we know

(8) $$y_i{}' = \varepsilon h_i(\mathbf{a}, \mathbf{b}, \mathbf{y}, \varepsilon)\mathbf{ab}.$$

Thus,

$$\int_0^t y_i{}' d\tau = \varepsilon \int_0^t h_i(\mathbf{a}, \mathbf{b}, \mathbf{y}, \varepsilon)\mathbf{ab} d\zeta,$$

(9) $$|y_i(t) - y_i(0)| \leq \varepsilon d_i \int_0^t |\mathbf{a}||\mathbf{b}| d\zeta,$$

where d_i is a bound on $|h_i|$. Finally, because $t = \mathcal{O}(\frac{1}{\varepsilon})$, part (iii) of Lemma 3.1 implies that the right-hand side of (9) is $\mathcal{O}(e^{-\frac{c}{\varepsilon}})$ as $\varepsilon \to 0$ for some $c > 0$, which establishes the proposition.

4. The C^1 ELESE. In order to get the C^1 version of the main result, which we present in this section, we follow [21] and study the vector field induced on the $(k+1)$-dimensional tangent planes of invariant manifolds using the $(k+1)$-forms which are dual to them. By tracking $(k+1)$-dimensional manifolds and their tangent planes, we capture the dynamics in all directions, including all of the center directions.

We are interested in the $(k+1)$-forms

$$P_{\sigma_1,\ldots,\sigma_{k+1}} \equiv \delta\sigma_1 \wedge \cdots \wedge \delta\sigma_{k+1},$$

where the σ_j range over the variables a_i, b_i, and y_i and where $\delta\sigma_i$ is the 1-form dual to the coordinate σ_i. Each $(k+1)$-form associates with the $(k+1)$-dimensional plane Q a number that is proportional to the volume of the projection of a unit cube of Q onto the coordinate plane of the $k+1$ coordinates specified by P. We assume that the σ_i are ordered in P according to the rule

$$a_1 < a_2 < \cdots < a_k < b_1 < \cdots < b_\ell < y_1 < \cdots < y_n,$$

and we work with the projectivized version of the $(k+1)$-forms

$$\hat{P}_{\sigma_1,\ldots,\sigma_{k+1}} \equiv \frac{P_{\sigma_1,\ldots,\sigma_{k+1}}}{P_{a_1,a_2,\ldots,a_k,y_1}}.$$

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

ELESE. *Let M be a $(k+1)$-dimensional invariant manifold. Assume that $\rho_\varepsilon = M \cap \{|\mathbf{b}| = \Delta\}$ intersects $\{\mathbf{a} = \mathbf{0}\}$ transversely. Let $q \in \rho_\varepsilon$ be a point whose trajectory exits from $\{|\mathbf{a}| = \Delta\}$ at $\bar{q}$ after a time $t = \mathcal{O}(\frac{1}{\varepsilon})$. Then, for ε sufficiently small, the manifold M is $\mathcal{O}(e^{-\frac{c}{\varepsilon}})$-$C^1$ close for some $c > 0$ to $\{\mathbf{b} = \mathbf{0}, y_i - y_i(0) = 0, i > 1\}$ at $\bar{q}$.*

Remark. The conclusion of the ELESE may be restated as: Upon exiting the box $\mathcal{B}$ at the point $\bar{q}$, the projectivized $(k+1)$-form $\hat{P}_{\sigma_1,\sigma_2,\ldots,\sigma_{k+1}}$ is $\mathcal{O}(e^{-\frac{c}{\varepsilon}})$ for some constant $c > 0$ and for all $(\sigma_1, \ldots, \sigma_{k+1}) \neq (a_1, \ldots, a_k, y_1)$.

Before proving ELESE, we develop the working estimates that we will need for the differential forms. After these are established, the proof of ELESE requires three short steps.

Let $\mathbf{z} \equiv (\mathbf{a}, \mathbf{b}, \mathbf{y})$. Let Λ_{ij} and Γ_{ij} denote the i,jth entries of Λ and Γ, resp., and let Λ_i and Γ_i denote their ith rows. In addition, let $L_{ij} \equiv \nabla_z \Lambda_{ij} \cdot \delta\mathbf{z}$ and $G_{ij} \equiv \nabla_z \Gamma_{ij} \cdot \delta\mathbf{z}$, where for each i,j pair, $\nabla_z \Lambda_{ij} \in \mathbb{R}^{k+\ell+n}$ and $\nabla_z \Gamma_{ij} \in \mathbb{R}^{k+\ell+n}$. Finally, let $\mathbf{L}_i = (L_{i1}, \ldots, L_{ik})$ and $\mathbf{G}_i = (G_{i1}, \ldots, G_{i\ell})$.

The development of the working estimates begins with the equations of variation of (4):

$$(\delta a_i)' = \Lambda_i \cdot \delta\mathbf{a} + \mathbf{a} \cdot \mathbf{L}_i,$$

$$(\delta b_i)' = \Gamma_i \cdot \delta\mathbf{b} + \mathbf{b} \cdot \mathbf{G}_i,$$

$$(\delta y_i)' = \varepsilon \sum_{j=1}^{k} \sum_{m=1}^{\ell} \{h_{ijm} \delta a_j b_m + h_{ijm} a_j \delta b_m + (\nabla h_{ijm} \cdot \delta\mathbf{z}) a_j b_m\}. \tag{10}$$

These equations of variation readily yield the evolution equations for each of the possible $(k+1)$-forms $\hat{P}_{\sigma_1,\ldots,\sigma_{k+1}}$. We split the forms into two blocks: block I consists of the forms $Z_i \equiv \hat{P}_{a_1 a_2,\ldots,a_k y_i}$ for all i, and block II contains all of the others, which we denote $\mathbf{X}$. From the definition of Z_i and the product rule for differentiation, we have

$$Z_i{}' = \sum_{j=1}^{k} \delta a_1 \wedge \cdots \wedge (\delta a_j)' \wedge \cdots \wedge \delta y_i + \delta a_1 \wedge \cdots \wedge \delta a_k \wedge (\delta y_i)'.$$

Then, using the fact that the wedge product of a 1-form by itself vanishes, specifically that

$$\delta a_1 \wedge \cdots \wedge \delta a_k \wedge \left(\sum_{m=1}^{\ell} \sum_{j=1}^{k} h_{ijm} \delta a_j b_m \right) \equiv 0 \qquad \text{for all } i,$$

we obtain

$$Z_i{}' = (\mathrm{Tr}\Lambda)\, Z_i + \eta_{1i}(\mathbf{Z}, \mathbf{X}, t), \tag{11}$$

where

$$\eta_{1i} = F_i(Z_i, \mathbf{X}, t) + G_i(\mathbf{Z}, \mathbf{X}, t) + Q_i(\mathbf{X}, t) \tag{12}$$

and

$$F_i \equiv \sum_{j=1}^{k} \delta a_1 \wedge \cdots \wedge (\mathbf{a} \cdot \mathbf{L}_j) \wedge \cdots \wedge \delta y_i,$$

$$G_i \equiv \varepsilon \delta a_1 \wedge \cdots \wedge \delta a_k \wedge \left\{ \sum_{m=1}^{\ell} \sum_{j=1}^{k} (\nabla h_{ijm} \cdot \delta \mathbf{z}) a_j b_m \right\},$$

$$Q_i \equiv \varepsilon \delta a_1 \wedge \cdots \wedge \delta a_k \wedge \left\{ \sum_{m=1}^{\ell} \sum_{j=1}^{k} h_{ijm} a_j \delta b_m \right\}. \tag{13}$$

We have kept the dependence of η_{1i} on $\mathbf{a}, \mathbf{b}$, and $\mathbf{y}$ implicit. Clearly, the inhomogeneous term η_{1i} vanishes for each i at $\mathbf{a} = \mathbf{0}$. In addition, the terms in F_i from block I involve only Z_i and are $\mathcal{O}(|\mathbf{a}|)$, and those in F_i from block II are also $\mathcal{O}(|\mathbf{a}|)$. All of the terms in G_i are $\mathcal{O}(\varepsilon|\mathbf{a}||\mathbf{b}|)$. Finally, the terms in Q_i can only be in block II, and they are $\mathcal{O}(\varepsilon|\mathbf{a}|)$. Thus, we have proved the following lemma.

LEMMA 4.1.

$$|F_i(Z_i, \mathbf{X}, t)| \le C_F |\mathbf{a}| (|Z_i| + |\mathbf{X}|),$$

$$|G_i(\mathbf{Z}, \mathbf{X}, t)| \le \varepsilon C_G |\mathbf{a}||\mathbf{b}| (|\mathbf{Z}| + |\mathbf{X}|),$$

$$|Q_i(\mathbf{X}, t)| \le \varepsilon C_Q |\mathbf{a}||\mathbf{X}| \tag{14}$$

for some $C_F, C_G, C_Q < \infty$ and for each i.

Remark. The factor of $|\mathbf{b}|$ in the bound on G_i plays a vital role in the working estimates of Lemma 4.3.

In a similar fashion, one derives the evolution equation for the $(k+1)$-forms $\mathbf{X}$:

$$\mathbf{X}' = B\mathbf{X} + \eta_2(\mathbf{Z}, \mathbf{X}, t), \tag{15}$$

where

$$\eta_2(\mathbf{Z},\mathbf{X},t) \equiv \mathbf{E}(\mathbf{X},t) + \mathbf{H}(\mathbf{Z},\mathbf{X},t) \tag{16}$$

and where $B = B(\mathbf{a},\mathbf{b},\mathbf{y},\varepsilon)$ is a matrix that satisfies

$$\|\exp\left\{\int_0^t (B - \mathrm{Tr}\Lambda)d\zeta\right\}\| \leq \bar{M}e^{-\mu(t-s)} \tag{17}$$

for some $\bar{M} \geq 1$ and $\mu > 0$. The terms included in $\mathbf{E}$ come from those $(k+1)$-forms that have the $\mathbf{a}\cdot\mathbf{L_i}$ term in them from a factor with $(\delta a_i)'$ and from those that have the $\varepsilon\sum_{m=1}^{\ell}\sum_{j=1}^{k} h_{ijm}a_j\delta b_m$ terms in them from a factor with $(\delta y_i)'$. The term $H(\mathbf{Z},\mathbf{X},t)$ consists of the remaining terms, each of which has a a factor of $|\mathbf{b}|$ in them. Therefore, recalling that $\varepsilon|\mathbf{a}| \leq \epsilon\Delta$, we have established the following lemma.

LEMMA 4.2.

$$|E(\mathbf{X},t)| \leq C_E|\mathbf{a}||\mathbf{X}|,$$

$$|H(\mathbf{Z},\mathbf{X},t)| \leq C_H|\mathbf{b}|\left(|\mathbf{Z}| + |\mathbf{X}|\right). \tag{18}$$

for some C_E, $C_H < \infty$.

Next, we use the bounds from Lemmas 4.1 and 4.2 to derive the working estimates. Let $\hat{\mathbf{X}} \equiv \frac{\mathbf{X}}{Z_1}$, and $\hat{Z}_i \equiv \frac{Z_i}{Z_1}$. We show the following.

LEMMA 4.3. *There exist constants $C, K > 0$ such that the following hold:*

$$|Z_1|' \geq \left\{(\mathrm{Tr}\Lambda) - C|\mathbf{a}|(1 + |\hat{\mathbf{X}}| + \varepsilon|\mathbf{b}||\hat{\mathbf{Z}}|)\right\}|Z_1|; \tag{19}$$

$$|\hat{Z}_i|' \leq \{2C|\mathbf{a}| + \alpha(t)\}|\hat{Z}_i| + \alpha(t), \tag{20}$$

where

$$\alpha(t) \equiv C|\mathbf{a}|\left[|\hat{\mathbf{X}}| + \varepsilon|\mathbf{b}||\hat{\mathbf{Z}}|\right]; \tag{21}$$

and

$$|\hat{\mathbf{X}}| \leq \bar{M}\left\{e^{\int_0^t \beta_1(s)ds}|\hat{\mathbf{X}}_0| + \int_0^t e^{\int_s^t \beta_1(r)dr}\beta_2(s)ds\right\}, \tag{22}$$

where

$$\beta_1 \equiv -\mu + C|\mathbf{a}||\hat{\mathbf{X}}| + C\Delta,$$

$$\beta_2 \equiv K|\mathbf{b}|\left[\left(1 + \varepsilon|\mathbf{a}||\hat{\mathbf{X}}|\right)|\hat{\mathbf{Z}}| + 1 + 2\varepsilon|\mathbf{a}||\hat{\mathbf{X}}|\right], \tag{23}$$

and μ and $\bar{M}$ are as stated in (17).

Proof of Lemma 4.3. The proof of this lemma is virtually identical to that of Lemma 3.3 in [21]. We state only the differences. In (19) and (21), there are factors of $|\mathbf{b}|$ in front of the $|\hat{\mathbf{Z}}|$ terms due to the bound on $|G_i|$ given by (14) that are not present in (3.11) and (3.12) of [21]. In addition, there is an overall factor of $|\mathbf{b}|$ in front of the first term in β_2 in (23), due to the sharper bound on H given by (18), and in the proof their β_3 is replaced by $\beta_3 \equiv C^*|\mathbf{a}|(1+|\hat{\mathbf{X}}|) + \varepsilon C^*|\mathbf{a}||\mathbf{b}|(1+|\hat{\mathbf{X}}|) + K|\mathbf{b}|$, for some constant C^*.

The last preliminary result we shall need is an estimate of the initial conditions.

LEMMA 4.4 *There exists a constant $K_1 > 0$ such that $|Z_1| > K_1\varepsilon^{\alpha+1}$, $|\hat{\mathbf{X}}| < \frac{1}{\bar{K}\varepsilon^{\alpha+1}}$, and $|\hat{Z}_i|$ are exponentially small at T_qM.*

Proof of Lemma 4.4. Since $|\mathbf{X}| < 1$ by the choice of normalization, the fact that the tracked manifold M and the stable manifold $\{\mathbf{a} = \mathbf{0}\}$ intersect transversely at an angle of $\mathcal{O}(\varepsilon^\alpha)$ implies that there exists a constant K_1 such that $|Z_1| > K_1\varepsilon^{\alpha+1}$ at T_qM. Therefore, using $|\mathbf{X}| < 1$, and the definition of $\hat{X}$, we also have $|\hat{X}| < \frac{1}{K_1\varepsilon^{\alpha+1}}$ at T_qM.

To establish the third result, we apply the argument given in the proof of Lemma 3.2 in [21]. We observe also that here the situation is even better in that the y_i component of the tangent vector is $\mathcal{O}(|\mathbf{a}||\mathbf{b}|)$ by (4) for each $i > 1$.

This concludes the proof of Lemma 4.4.

We finish the proof of the C^1 ELESE in three steps. Let $T = \mathcal{O}(\frac{1}{\varepsilon})$ denote the time required for the trajectory to pass through $\mathcal{B}$ from q to $\bar{q}$. We assume $t = 0$ corresponds to the time at which the trajectory in question is at the point q.

Remark. Although the same in spirit as those in [21], all three steps involve different estimates. The differences can be traced back to the term $\mathbf{b}$ that is factored out in the normal form (4) and to the fact that our technique is designed to apply to any problem in which ρ_ε and $\{\mathbf{a} = \mathbf{0}\}$ intersect transversely at an angle of $\mathcal{O}(\varepsilon^\alpha)$ for any $\alpha \geq 0$.

Step I. Let $0 < T_1 < T$ be any time of size $\mathcal{O}(\frac{1}{\varepsilon})$ such that $|\mathbf{a}|$ is exponentially small for $t \leq T_1$. Then at $t = T_1$, $|\hat{\mathbf{X}}| = \mathcal{O}(\varepsilon + \Delta)$ and $|\hat{\mathbf{Z}}|$ is exponentially small.

Proof of Step I. In Lemma 4.4, we showed that since the angle of intersection of ρ_ε and $\{\mathbf{a} = \mathbf{0}\}$ is $\mathcal{O}(\varepsilon^\alpha)$, with $\alpha \geq 0$, at $t = 0$, the quantities $|\hat{\mathbf{X}}|$ and $|\hat{\mathbf{Z}}|$ satisfy the bounds $|\hat{\mathbf{X}}| \leq \frac{K_1}{\varepsilon^{\alpha+1}}$ and $|\hat{\mathbf{Z}}| \leq 1$ at $t = 0$. Thus, since $|\mathbf{a}|$ is exponentially small at $t = 0$, it follows that $\beta_1 = -\mu + \mathcal{O}(\Delta) +$ exp small and $\beta_2 \leq 2K\Delta +$ exp small at $t = 0$. Then, as long as $|\hat{\mathbf{Z}}| \leq 1$,

$$\beta_1 \leq -\frac{\mu}{2}, \tag{24}$$

and

$$\beta_2 \leq 3K\Delta, \tag{25}$$

(22)–(25) imply

$$|\hat{\mathbf{X}}| \leq \bar{M}\left\{e^{-\frac{\mu}{2}t} \cdot \frac{1}{\bar{K}\varepsilon^{\alpha+1}} + \frac{6K\Delta}{\mu}\right\}. \tag{26}$$

Hence the a priori estimate $|\hat{\mathbf{X}}| \leq \frac{K_1}{\varepsilon^{\alpha+1}}$ continues to hold.

Next, we observe that since $|\hat{\mathbf{Z}}|$ is exponentially small at q by Lemma 4.4, not only does the a priori bound $|\hat{\mathbf{Z}}| \leq 1$ continue to hold for all t up to and including T_1, but $|\hat{\mathbf{Z}}|$ is exponentially small at $t = T_1$. This follows immediately from applying Gronwall's inequality to (20) and recalling that $|\mathbf{a}|$ is exponentially small for $t \leq T_1$ so that $\alpha(t) + 2C|\mathbf{a}| \leq K_2 e^{-\frac{c_2}{\varepsilon}} \equiv \kappa$:

$$|\hat{Z}_i(t)| \leq |\hat{Z}_i(0)|e^{\kappa t} + \left(e^{\kappa t} - 1\right).$$

Finally, (26) implies that $|\hat{\mathbf{X}}| = \mathcal{O}(\varepsilon + \Delta)$ at $t = T_1$.

Step II. Let $T_1 < T_2 < T$ be any time of size $\mathcal{O}(\frac{1}{\varepsilon})$ such that $T - T_2 = \mathcal{O}(\frac{1}{\varepsilon})$ and $|\mathbf{a}|$ and $|\mathbf{b}|$ are exponentially small for $T_1 \le t \le T_2$. Then at $t = T_2$, $|\hat{\mathbf{X}}|$ and $|\hat{\mathbf{Z}}|$ are exponentially small.

Proof of Step II. The result of Step I implies that the a priori bounds $|\hat{\mathbf{X}}| \le \frac{K_1}{\varepsilon^{\alpha+1}}$ and $|\hat{\mathbf{Z}}| \le 1$ hold at $t = T_1$. We now establish these a priori bounds up to T_2.

From the definition of β_2 in (23) and the fact that $|\mathbf{b}|$ is exponentially small, we know

$$\beta_2 \le K_4 e^{-\frac{c_4}{\varepsilon}}. \tag{27}$$

Hence, as long as $|\hat{\mathbf{Z}}| \le 1$, the estimate (27) on β_2 holds up to T_2, implying that $|\hat{\mathbf{X}}| \le \frac{K_1}{\varepsilon^{\alpha+1}}$ at T_2. Also, while $|\hat{\mathbf{X}}| \le \frac{K_1}{\varepsilon^{\alpha+1}}$ continues to hold in this step, we know that $|\hat{\mathbf{Z}}| \le 1$. This establishes the a priori bounds for the entire step.

Next, using (27) and substituting $t = T_2$ in the inequality (22), we obtain that $|\hat{\mathbf{X}}|$ is exponentially small at $t = T_2$, which is as desired, since T_2 is $\mathcal{O}(\frac{1}{\varepsilon})$. Finally, $|\hat{\mathbf{Z}}|$ is also exponentially small at $t = T_2$, since application of Gronwall's inequality to (20) (in a fashion similar to the application of Gronwall's inequality in Step I) yields $|\hat{Z}_i(t)| \le |\hat{Z}_i(0)|e^{\kappa t} + (e^{\kappa t} - 1)$, where

$$\alpha(t) + 2C|\mathbf{a}| \le K_2 e^{-\frac{c_2}{\varepsilon}} \equiv \kappa.$$

Step III. At $t = T$, $|\hat{\mathbf{X}}|$ and $|\hat{\mathbf{Z}}|$ are exponentially small.

Proof of Step III. During this final step, $|\mathbf{b}|$ is exponentially small while $|\mathbf{a}| \le \mathcal{O}(\Delta)$. We first establish the a priori bounds $|\hat{\mathbf{Z}}| \le 1$ and $|\hat{\mathbf{X}}| \le K_3 e^{-\frac{c_3}{\varepsilon}}$. From Step II, we know that these bounds are more than satisfied at $t = T_2$.

If $|\hat{\mathbf{X}}| \le K_3 e^{-\frac{c_3}{\varepsilon}}$ for some K_3 and c_3, then the bound $\beta_1 \le -\frac{\mu}{2}$ and Lemma 4.3 imply

$$|\hat{\mathbf{X}}| \le \bar{M}\left[e^{-\frac{\mu}{2}(t-T_2)}|\hat{\mathbf{X}}(T_2)| + \text{exp small}\right]. \tag{28}$$

Hence, for $K_3 e^{-\frac{c_3}{\varepsilon}} > \bar{M} \cdot (\text{exp small})$, the a priori bound continues to hold as long as $|\hat{\mathbf{Z}}| \le 1$.

Next, we verify that $|\hat{\mathbf{Z}}| \le 1$ holds throughout this step. Recalling (21),

$$\alpha(t) \equiv C|\mathbf{a}|\left[|\hat{\mathbf{X}}| + \varepsilon|\mathbf{b}||\hat{\mathbf{Z}}|\right],$$

we know that for $|\hat{\mathbf{Z}}| \le 1$ and $|\hat{\mathbf{X}}|$ exponentially small,

$$\alpha(t) \le C|\mathbf{a}|K_5 e^{-\frac{c_5}{\varepsilon}}$$

because $|\mathbf{b}|$ is exponentially small. Therefore, (20) implies

$$|\hat{Z}_i(t)| \le \frac{\tilde{C}}{\varepsilon}|\hat{Z}_i(T_2)| + \varepsilon\int_{T_2}^{t} (\text{exp small})\, ds, \tag{29}$$

which guarantees $|\hat{\mathbf{Z}}| \le 1$. Note that in deriving this bound, we took advantage of the fact that the integrals of $|\mathbf{a}|$ stay finite as $\varepsilon \to 0$.

Observing that the first term on the right-hand side of (29) is exponentially small and that the second term is also exponentially small at $t = T$, we have thus concluded the proof of the C^1 ELESE.

Remark 1. T. Kaper has shown that ELESE can be extended to manifolds M of dimension $k + 2$. After that work was completed, the generalization to the $(k + \sigma)$-dimensional case, with $1 \leq \sigma \leq n$, was carried out by S. K. Tin using our methods. Furthermore, he has removed the rectification requirement by working in Plücker coordinates.

Remark 2. In the applications listed in the introduction, the slow manifolds have two or more dimensions and the flows on them range from simple flows with fixed points to complex flows exhibiting resonance bands. Also, these systems depend on parameters, such as the speed of the travelling wave or the amplitude of forcing or damping. See [23] in particular for an example of two coupled nonlinear oscillators in which there is a two-dimensional slow manifold with a resonance band on it and in which the $k + 2$ version of ELESE is used to establish the existence of eight different types of multipulse orbits heteroclinic to fixed points and periodic orbits, including multipulse Silnikov orbits.

5. Multipulse orbits heteroclinic to periodic orbits: An application of ELESE. We present a model problem, consisting of a quasiperiodically modulated pendulum, to illustrate how ELESE with $k = 1$ can be applied. In particular, we prove the existence of multipulse orbits heteroclinic to periodic orbits. Recall that pulses are defined to be those segments of an orbit which are the fast excursions from one slow manifold to another that the orbit makes in between the long intervals of time it spends near the slow manifolds.

The equations of motion are

$$\ddot{q} + A(\tau_1, \tau_2; \omega_1, \beta_1, \beta_2) \sin q = 0, \tag{30a}$$

$$\dot{\tau}_1 = \epsilon, \tag{30b}$$

$$\dot{\tau}_2 = \epsilon \sin(2\pi\tau_2), \tag{30c}$$

where $q \in \mathbb{R}$, $A(\tau_1, \tau_2; \omega_1, \beta_1, \beta_2) = 1 + \beta_1 \sin(2\pi\omega_1\tau_1) + \beta_2 \sin(2\pi\tau_2)$ with $\omega_1 > 1$ and irrational, and $\beta_1, \beta_2 > 0$ such that $0 < \beta_1 + \beta_2 < 1$, $(\tau_1, \tau_2) \in T^2$, and $0 < \varepsilon \ll 1$.

When $\varepsilon = 0$, the parameters τ_1 and τ_2 are frozen, and the system reduces to that of the classical nonlinear pendulum with Hamiltonian

$$H(q, p, \tau_1, \tau_2) = \frac{p^2}{2} - A(\tau_1, \tau_2; \omega_1, \beta_1, \beta_2) \cos q - A(\tau_1, \tau_2; \omega_1, \beta_1, \beta_2). \tag{31}$$

The term $-A(\tau_1, \tau_2; \omega_1, \beta_1, \beta_2)$, independent of q and p, is included in the Hamiltonian H so that H vanishes identically at the fixed points $H(\pm\pi, 0, \tau_1, \tau_2) \equiv 0$ for all $(\tau_1, \tau_2) \in T^2$. Of course, on the cylinder $S^1 \times \mathbb{R}$, these two saddle fixed points coincide, but they are distinct on the universal cover $\mathbb{R}^2$, which is where we shall study the global geometry.

In the $\varepsilon = 0$ phase space, there exists a pair of normally hyperbolic invariant two-tori, which are the Cartesian products of the saddle fixed points $(q = \pm\pi, p = 0)$ with the two-torus T^2:

$$S_0^1 \equiv (-\pi, 0) \times T^2, \tag{32a}$$

$$S_0^2 \equiv (\pi, 0) \times T^2. \tag{32b}$$

Moreover, every point on S_0^1 is connected by two heteroclinic orbits

$$\Gamma^{\text{up}}_{\tau_1,\tau_2} \equiv (q_h^{\text{up}}(t), p_h^{\text{up}}(t), \tau_1, \tau_2), \tag{33a}$$

$$\Gamma^{\text{down}}_{\tau_1,\tau_2} \equiv (q_h^{\text{down}}(t), p_h^{\text{down}}(t), \tau_1, \tau_2) \tag{33b}$$

to the point on S_0^2 which has the same τ_1 and τ_2 coordinates. The orbit $\Gamma^{\text{up}}_{\tau_1,\tau_2}$ (resp., $\Gamma^{\text{down}}_{\tau_1,\tau_2}$) is asymptotic to the point $(-\pi, 0, \tau_1, \tau_2)$ as $t \to -\infty$ (resp., $t \to +\infty$) and to the point $(\pi, 0, \tau_1, \tau_2)$ as $t \to \infty$ (resp., $t \to -\infty$). The normally hyperbolic invariant manifolds S_0^1 and S_0^2 have smooth three-dimensional stable and unstable manifolds, $W^S(S_0^1)$ and $W^U(S_0^1)$; $W^S(S_0^2)$ and $W^U(S_0^2)$, respectively. One branch of each of $W^U(S_0^1)$ and $W^S(S_0^2)$ and one branch of each of $W^U(S_0^2)$ and $W^S(S_0^1)$ coincide in the homoclinic manifolds

$$\Xi^{\text{up}} \equiv W^U(S_0^1) \bigcap W^S(S_0^2) \equiv \bigcup_{\tau_1,\tau_2 \in T^2} \Gamma^{\text{up}}_{\tau_1,\tau_2}, \tag{34a}$$

$$\Xi^{\text{low}} \equiv W^U(S_0^2) \bigcap W^S(S_0^1) \equiv \bigcup_{\tau_1,\tau_2 \in T^2} \Gamma^{\text{down}}_{\tau_1,\tau_2}. \tag{34b}$$

Now, for $0 < \varepsilon \ll 1$, the persistence theory for normally hyperbolic invariant manifolds presented in [13] guarantees that there exist manifolds S_ε^1 and S_ε^2 that are differentiably $\mathcal{O}(\varepsilon)$ close to S_0^1 and S_0^2. In fact, for (30), we have $S_\varepsilon^1 \equiv S_0^1$ and $S_\varepsilon^2 \equiv S_0^2$ since the perturbation in (30a) was chosen to be in the form of slow amplitude modulation.

The slow flow on S_ε^1 is illustrated in Fig. 1. On S_ε^1, there are two periodic orbits

$$\gamma^{1a} \equiv \left\{ (\tau_1, \tau_2) | \tau_2 = \frac{1}{2} \bmod 1 \right\}, \tag{35a}$$

$$\gamma^{1r} \equiv \{ (\tau_1, \tau_2) | \tau_2 = 0 \bmod 1 \}, \tag{35b}$$

which are attracting and repelling, respectively, for the system (30b) and (30c). The slow flow on S_ε^2 is the same, and we denote its attracting and repelling periodic orbits by γ^{2a} and γ^{2r}, respectively.

The persistent hyperbolic manifolds S_ε^1 and S_ε^2 have smooth local stable and unstable manifolds—$W^S_{\text{loc}}(S_\varepsilon^1)$, and $W^U_{\text{loc}}(S_\varepsilon^1)$; $W^S_{\text{loc}}(S_\varepsilon^2)$, and $W^U_{\text{loc}}(S_\varepsilon^2)$—which are $\mathcal{O}(\varepsilon)$ close to their unperturbed counterparts. However, the perturbed global stable and unstable manifolds no longer coincide. Instead, they intersect transversely along two-dimensional surfaces spanned by heteroclinic orbits connecting one of the slow manifolds to the other. These heteroclinic orbits have only one pulse and may be detected by the adiabatic Melnikov theory; see [31], [30], and [22]. We briefly review the relevant one-pulse results, since we need to know some details about the perturbed global geometry for our construction of multipulse heteroclinic orbits. We refer the reader to [35] for new results and a complete exposition of Melnikov theory in perturbed Hamiltonian systems.

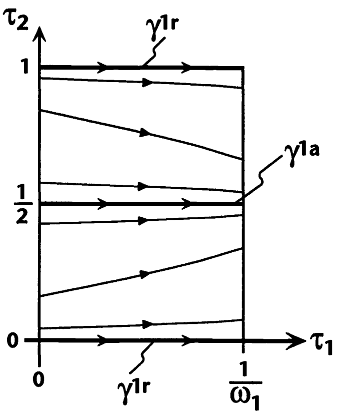


FIG. 1. *The phase portrait of the slow flow on the perturbed slow manifold S^1_ε in the example of* §5.

We work only with the upper pair of manifolds $W^U(S^1_\varepsilon)$ and $W^S(S^2_\varepsilon)$. The work and result for the lower pair are almost identical. The Melnikov function is

$$M(\tau_1,\tau_2;\omega_1,\beta_1,\beta_2) = \int_{-\infty}^{\infty} \left(\nabla_{(\tau_1,\tau_2)}H \cdot \mathbf{g}^\tau\right)(q_h^{\text{up}}(t),p_h^{\text{up}}(t),\tau_1,\tau_2)dt, \tag{36}$$

where $\mathbf{g}^\tau$ is the two-component vector consisting of the coefficients on the $\mathcal{O}(\varepsilon)$ terms in the vector field (30b) and (30c). Melnikov theory implies that if there exists a point $(\bar{\tau}_1,\bar{\tau}_2) \in T^2$ and parameter values $(\bar{\omega}_1,\bar{\beta}_1,\bar{\beta}_2)$ such that

$$M(\bar{\tau}_1,\bar{\tau}_2;\bar{\omega}_1,\bar{\beta}_1,\bar{\beta}_2) = 0, \tag{37a}$$

$$\nabla_{(\tau_1,\tau_2)}M(\bar{\tau}_1,\bar{\tau}_2;\bar{\omega}_1,\bar{\beta}_1,\bar{\beta}_2) \text{ has rank } 1, \tag{37b}$$

then, for sufficiently small ε, $W^U(S^1_\varepsilon)$ intersects $W^S(S^2_\varepsilon)$ in a two-dimensional surface near $(\bar{\tau}_1,\bar{\tau}_2;\bar{\omega}_1,\bar{\beta}_1,\bar{\beta}_2)$. Furthermore, this two-dimensional intersection surface varies smoothly with ε for $0<\varepsilon\ll 1$, and at every point on it, the intersection of $W^U(S^1_\varepsilon)$ and $W^S(S^2_\varepsilon)$ is transverse, with the angle between the tangent planes being $\mathcal{O}(\varepsilon)$. See [35, Thm. 4.1.14], where our example corresponds to the case of $m=2$, $n=1$, and $l=0$.

For (30), an explicit computation of (36) yields

$$M(\tau_1,\tau_2;\omega_1,\beta_1,\beta_2) = -\pi\left[2\beta_1\omega_1\cos(2\pi\omega_1\tau_1)+\beta_2\sin(4\pi\tau_2)\right]\int_{-\infty}^{\infty}\left(1+\cos(q_h^{\text{up}}(t)\right)dt. \tag{38}$$

Since the improper integral in (38) is a finite positive number, one readily verifies that, for fixed ω_1,β_1, and β_2, the Melnikov function M has two smooth curves, η^{11} and η^{12}, of simple zeroes on T^2. For β_2 sufficiently small, the curves η^{11} and η^{12} stay bounded away from each other. The curve η^{11} passes through the points $(\tau_1=\frac{1}{4\omega_1},\tau_2=0)$, $(\tau_1=\frac{1}{4\omega_1},\tau_2=\frac{1}{4})$, $(\tau_1=\frac{1}{4\omega_1},\tau_2=\frac{1}{2})$, and $(\tau_1=\frac{1}{4\omega_1},\tau_2=\frac{3}{4})$. Similarly, the curve η^{12} passes through the points $(\tau_1=\frac{3}{4\omega_1},\tau_2=0)$, $(\tau_1=\frac{3}{4\omega_1},\tau_2=\frac{1}{4})$, $(\tau_1=\frac{3}{4\omega_1},\tau_2=\frac{1}{2})$, and $(\tau_1=\frac{3}{4\omega_1},\tau_2=\frac{3}{4})$. The zero set on S^1_0 is shown schematically in Fig. 2.

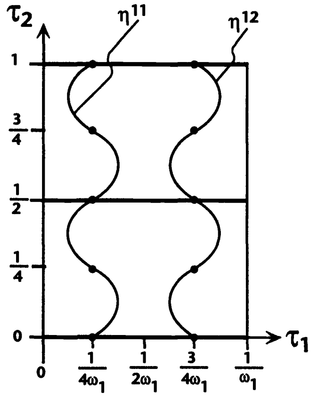


FIG. 2. *The zero set of the Melinkov function* (38) *on* S_0^1.

Hence, Melnikov theory implies that there exist two families of surviving one-pulse heteroclinic orbits. Each orbit is asymptotic to S_ε^1 as $t \to -\infty$ and to S_ε^2 as $t \to \infty$.

The Melnikov function for the lower separatrix $\Gamma_{\tau_1,\tau_2}^{\mathrm{down}}$ is identical to the Melnikov function given by (38) for the upper separatrix since the integral in (38) with $q_h^{\mathrm{down}}(t)$ replacing $q_h^{\mathrm{up}}(t)$ is the same. Thus, $W^U(S_\varepsilon^2)$ and $W^S(S_\varepsilon^1)$ also intersect transversely at an angle of $\mathcal{O}(\varepsilon)$ near the simple zeroes of the Melnikov function (38), and there are two families of one-pulse heteroclinic orbits that asymptote to S_ε^2 as $t \to -\infty$ and to S_ε^1 as $t \to \infty$.

We observe that the curves η^{11} and η^{12} transversely intersect each of the periodic orbits $\gamma^{1a}, \gamma^{1r}, \gamma^{2a}$, and γ^{2r}. Therefore, among the surviving orbits in the abovementioned families, there exist eight distinguished one-pulse orbits that are actually heteroclinic to the periodic orbits γ^{1a} and γ^{2a}; and γ^{1r} and γ^{2r}. Each of these is the unique orbit close on finite-time intervals to one of the eight distinguished unperturbed heteroclinic orbits

$$\text{(39a)} \qquad \Gamma^1 \equiv \Gamma^{\mathrm{up}}_{\frac{1}{4\omega_1},\frac{1}{2}}, \qquad \Gamma^2 \equiv \Gamma^{\mathrm{up}}_{\frac{3}{4\omega_1},\frac{1}{2}},$$

$$\text{(39b)} \qquad \Gamma^3 \equiv \Gamma^{\mathrm{up}}_{\frac{1}{4\omega_1},0}, \qquad \Gamma^4 \equiv \Gamma^{\mathrm{up}}_{\frac{3}{4\omega_1},0},$$

and $\Gamma^5 - \Gamma^8$, which are defined as $\Gamma^1 - \Gamma^4$ are, respectively, but with the superscript down replacing up. These are distinguished because their takeoff and touchdown points (limits as $t \to \mp\infty$) are the eight points on S_0^1 and S_0^2 in which the zero curves η^{11} and η^{22} intersect the periodic orbits $\gamma^{1a}, \gamma^{1r}, \gamma^{2a}$, and γ^{2r}.

This concludes our brief review of the one-pulse results.

We now turn our attention to multipulse orbits and establish the existence of many heteroclinic orbits connecting the periodic orbits γ^{1a}, γ^{1r}, γ^{2a}, and γ^{2r} to each other which have finitely many fast pulses interspersed with slow segments. The main idea in proving their existence is to track the unstable manifolds of these periodic orbits through each passage near the slow manifolds S_ε^1 and S_ε^2 using ELESE.

We begin by constructing *singular* multipulse heteroclinic orbits. Two examples of three-pulse singular heteroclinic orbits (see Fig. 3) are

$$\text{(40)} \qquad \gamma^{2a} \bigcup \Gamma^5 \bigcup \tilde{\gamma}^{1a} \bigcup \Gamma^2 \bigcup \tilde{\tilde{\gamma}}^{2a} \bigcup \Gamma^5 \bigcup \gamma^{1a}$$

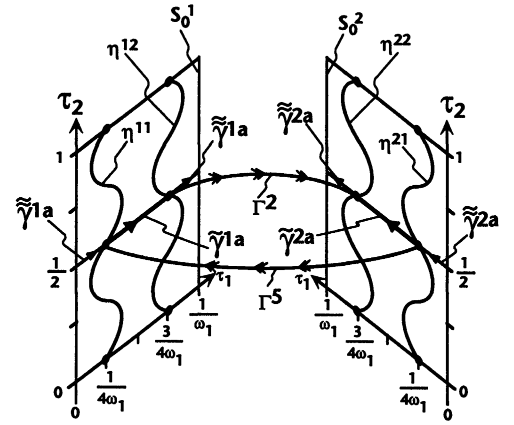


FIG. 3. *The geometry of the triple-pulse singular heteroclinic orbit given by* (40) : $\gamma^{2a} \cup \Gamma^5 \cup \tilde{\gamma}^{1a} \cup \Gamma^2 \cup \tilde{\tilde{\gamma}}^{2a} \cup \Gamma^5 \cup \gamma^{1a}$.

and

$$\gamma^{1r} \bigcup \Gamma^3 \bigcup \tilde{\tilde{\gamma}}^{2r} \bigcup \Gamma^8 \bigcup \gamma^{1r} \bigcup \Gamma^4 \bigcup \gamma^{2r}, \tag{41}$$

where $\tilde{\gamma}^{ij}$ denotes the segment of the periodic orbit γ^{ij} with $\frac{1}{4\omega_1} \leq \tau_1 \leq \frac{3}{4\omega_1}$ and $\tilde{\tilde{\gamma}}^{ij}$ denotes the remaining piece of the periodic orbit γ^{ij}.

In general, an N-pulse singular orbit hetero- or homoclinic to the attracting periodic orbits consists of the following:

A. the slow attracting periodic orbits, γ^{ia} and γ^{ja}, to which the orbit will be hetero- or homoclinic
and

B. the N unperturbed heteroclinic orbits chosen from $\{\Gamma^1, \Gamma^2, \Gamma^5, \Gamma^6\}$ that are alternatingly "up" and "down" and that connect points on γ^{1a} and γ^{2a},
which are interspersed with

C. segments of γ^{1a} and γ^{2a}, where a segment of γ^{ia} may consist of going around γ^{ia} finitely many (including fractional) times, i.e., any finite string of alternating segments $\tilde{\gamma}^{ia}$ and $\tilde{\tilde{\gamma}}^{ia}$.

In the same fashion, one constructs N-pulse singular orbits hetero- and homoclinic to the repelling periodic orbits, except that the unperturbed heteroclinics are chosen from the set $\{\Gamma^3, \Gamma^4, \Gamma^7, \Gamma^8\}$.

Remark. Of course, in A, $i = j$ if N is even, and $i \neq j$ if N is odd.

We now prove the following theorem.

THEOREM. *For all ε sufficiently small, there exists a unique true multipulse heteroclinic orbit $\mathcal{O}(\varepsilon)$ close to every singular multipulse heteroclinic orbit.*

Remark. The true and singular multipulse orbits are only $\mathcal{O}(\varepsilon)$ close because the stable and unstable manifolds of the perturbed tori are $\mathcal{O}(\varepsilon)$ close to their unperturbed counterparts. Furthermore, for general problems of the form (1), $\mathcal{O}(\varepsilon)$ closeness of these orbits is the best that can be expected since the perturbed and unperturbed slow manifolds in (1) are $\mathcal{O}(\varepsilon)$ apart, in contrast to the situation here where they coincide.

Proof of the theorem. For the sake of exposition, we prove the existence of the unique three-pulse heteroclinic orbit connecting γ^{2a} and γ^{1a} that lies near the singular heteroclinic orbit (40):

$$\gamma^{2a}\bigcup\Gamma^5\bigcup\tilde{\gamma}^{1a}\bigcup\Gamma^2\bigcup\tilde{\tilde{\gamma}}^{2a}\bigcup\Gamma^5\bigcup\gamma^{1a}. \tag{42}$$

The generalization to all of the other orbits in the theorem is given at the end of the proof.

Let $\mathcal{M}_\varepsilon$ denote the two-dimensional manifold $W^U(S^2_\varepsilon)|_{\gamma^{2a}}$ that we will track in this proof using ELESE with $k = 1$.

First, Melnikov theory immediately implies that

(43) $W^U(S^2_\varepsilon)|_{\gamma^{2a}}$ transversely intersects $W^S(S^1_\varepsilon)$ at an angle of $\mathcal{O}(\varepsilon)$

since this is a one-dimensional subset of the one-pulse intersection surface. Part of this intersection is near Γ^5.

Next, due to the existence of this transverse intersection, it follows by a continuous-dependence argument that there exist initial conditions on $\mathcal{M}_\varepsilon$ whose trajectories satisfy the hypotheses of ELESE, i.e., initial conditions whose trajectories spend $\mathcal{O}(\frac{1}{\varepsilon})$ time inside a Δ neighborhood of S^1_ε. Namely, the initial conditions actually in the intersection never leave the neighborhood, and there are initial conditions that leave the neighborhood in $\mathcal{O}(\ln\frac{1}{\varepsilon})$ time, due to the exponential expansion in the a direction. Hence, by continuity of the manifold, there exist initial conditions which spend the desired amount of time near S^1_ε. More specifically, there exists a point q on $\mathcal{M}_\varepsilon \cap \{|b| = \Delta\}$ such that the trajectory through q exits the Δ-neighborhood of S^1_ε at a point $\bar{q} \in \mathcal{M}_\varepsilon \cap \{|a| = \Delta\}$ after a time $(\frac{1}{2\omega_1\varepsilon}) + \mathcal{O}(1)$.

Therefore, the hypotheses of ELESE are satisfied, and ELESE implies that at the point $\bar{q}$, the manifold $\mathcal{M}_\varepsilon$ and its tangent space are $\mathcal{O}(e^{-\frac{c}{\varepsilon}})$ close to the local unstable manifold $W^U_{\text{loc}}(S^1_\varepsilon)|_{\gamma^{1a}}$ and its tangent space.

Now, the process repeats, but in the other direction. First, since $\mathcal{M}_\varepsilon$ and $W^U_{\text{loc}}(S^1_\varepsilon)|_{\gamma^{1a}}$ are so close by the previous step, and since $W^U_{\text{loc}}(S^1_\varepsilon)|_{\gamma^{1a}}$ transversely intersects $W^S(S^2_\varepsilon)$ at an angle of $\mathcal{O}(\varepsilon)$ by Melnikov theory (part of this intersection being close to Γ^2), it follows immediately that the tracked manifold $\mathcal{M}_\varepsilon$ transversely intersects $W^S(S^2_\varepsilon)$ at an angle of $\mathcal{O}(\varepsilon)$ as well. Then, as above, there exist orbits which spend the desired amount of time, $(\frac{1}{2\omega_1\varepsilon}) + \mathcal{O}(1)$ in the case of this example, in a Δ-neighborhood of S^2_ε and which satisfy the hypotheses of ELESE. Thus, ELESE implies that, at the point where these trajectories exit that neighborhood, $\mathcal{M}_\varepsilon$ and its tangent hyperplane are $\mathcal{O}(e^{-\frac{c}{\varepsilon}})$ close to $W^U_{\text{loc}}(S^2_\varepsilon)|_{\gamma^{2a}}$.

Finally, we use once more the first step in the process: since $\mathcal{M}_\varepsilon$ and $W^U_{\text{loc}}(S^2_\varepsilon)|_{\gamma^{2a}}$ are so close by the previous step, and since $W^U(S^2_\varepsilon)|_{\gamma^{2a}}$ transversely intersects $W^S(S^1_\varepsilon)$ at an angle of $\mathcal{O}(\varepsilon)$ by Melnikov theory (part of this intersection being near Γ^5), it follows immediately that the tracked manifold $\mathcal{M}_\varepsilon$ transversely intersects $W^S(S^1_\varepsilon)$ at an angle of $\mathcal{O}(\varepsilon)$, as well. Hence, we have constructed a three-pulse orbit backwards asymptotic to γ^{2a} and forwards asymptotic to γ^{1a}. Furthermore, this orbit is unique due to the transversality of the above intersections, and it is $\mathcal{O}(\varepsilon)$ close to the singular heteroclinic orbit (40).

The proof just given for the three-pulse example readily extends to a proof of the existence of all of the orbits in the theorem. Let $\mathcal{M}_\varepsilon$ denote the two-dimensional manifold $W^U(S^i_\varepsilon)|_{\gamma^{ia}}$ (where $i = 1$, or $i = 2$ is the index of the slow periodic orbit to which the desired hetero- or homoclinic orbit is backwards asymptotic). The above proof explicitly shows the existence of the first pulse near the first fast piece of the singular orbit. Then, by transversality of the intersections and by the same continuous-dependence argument given above, there exist initial conditions on $\mathcal{M}_\varepsilon$ whose trajectories spend long time intervals near the other slow manifold S^j_ε (i.e., $i \neq j$) and whose trajectories satisfy the hypotheses of ELESE. We are especially interested in those trajectories which spend time intervals of length $(\frac{k}{2\omega_1})\frac{1}{\varepsilon} + \mathcal{O}(1)$ near S^j_ε for any fixed positive integer k, since these leave the neighborhood of S^j_ε near the valid takeoff points given by the curves η^{j1} and η^{j2} of the Melnikov function zeroes.

Then, by ELESE, when these trajectories exit the neighborhood of S^j_ε, $\mathcal{M}_\varepsilon$ is C^1-$\mathcal{O}(e^{-\frac{c}{\varepsilon}})$ close to the fast unstable foliation $W^U_{\text{loc}}(S^j_\varepsilon)|_{\gamma^{ja}}$. Hence, a portion of $\mathcal{M}_\varepsilon$ is brought back into the neighborhood of the first slow manifold, S^i_ε, due to the fact that $W^U(S^j_\varepsilon)$ and $W^S(S^i_\varepsilon)$ intersect transversely at an angle of $\mathcal{O}(\varepsilon)$. By induction, this process can be repeated finitely many times until one completes the hetero- or homoclinic orbit with the desired finite number of pulses and slow segments in between. □

Remark. In addition to the heteroclinic multipulse orbits found above, there exist other types of multipulse orbits in this model which may be treated with similar methods. Above, we found orbits that are either backward and forward asymptotic to the periodic orbits that are attracting on the slow tori or orbits that are backward and forward asymptotic to the periodic orbits that are repelling on the slow tori. Another type of multipulse orbit consists of those that are backward asymptotic to a repelling periodic orbit and forward asymptotic to an attracting periodic orbit. The singular orbits for these consist of the union of finitely many segments of slow orbits on the tori (orbit segments that are disjoint from the periodic orbits) interspersed with the appropriate up and down heteroclinic connections. The segments of slow orbits start at the touchdown point $(\tau_1^{\text{touchdown}}, \tau_2^{\text{touchdown}})$ of the heteroclinic orbit $\Gamma^{\text{up or down}}_{\tau_1^{\text{touchdown}}, \tau_2^{\text{touchdown}}}$ and end at the takeoff point $(\tau_1^{\text{takeoff}}, \tau_2^{\text{takeoff}})$ of the next heteroclinic encountered in the singular orbit.

In contrast to the situation of the theorem in this section, the singular orbits discussed in the previous paragraph are not locally unique; there is a two-dimensional surface of each type of multipulse heteroclinic orbit, parametrized by the value of the τ_2 coordinate at their first takeoff points. The existence of real multipulse heteroclinic orbits near these singular ones may be established using the $k+2$ ELESE, where $(k+2)$-dimensional manifolds are tracked and where 2 is the number of center directions that are followed. Note, of course, that there are only two center directions. Equivalently a λ-lemma argument may be employed where the entire slow manifold is treated as the fixed point of some map associated with the vector field, say the time-1 map of the flow.

Appendix. In this appendix, we briefly review the relevant Melnikov theory. For fixed $(t = 0, \tau_1, \tau_2) \in \mathbb{R} \times T^2$, let $r \equiv (q_h^{\text{up}}(0), p_h^{\text{up}}(0), \tau_1, \tau_2)$ denote the point on the unperturbed homoclinic manifold $\Xi^{\text{up}} - (S^1_0 \cup S^2_0)$. At the point r, there is a three-dimensional hyperplane

$$\Pi_r \equiv \text{span}\left\{(0,0,\tau_1 = 1, 0), (0,0,0,\tau_2 = 1), \left(\frac{\partial H}{\partial q}, \frac{\partial H}{\partial p}, 0, 0\right)\right\} \tag{44}$$

transverse to Ξ^{up}, where the derivatives of H are evaluated at the point r. We will measure the splitting distance between the perturbed unstable and stable manifolds $W^U(S^1_\varepsilon)$ and $W^S(S^2_\varepsilon)$ in the transverse hyperplane Π_r.

First, we observe that the coincident unperturbed manifolds $W^U(S^1_0)$ and $W^S(S^2_0)$ intersect Π_r transversely in a two-dimensional surface. Hence, we also know that the perturbed manifolds $W^U_{\text{loc}}(S^1_\varepsilon)$ and $W^S_{\text{loc}}(S^2_\varepsilon)$ intersect Π_r transversely in two-dimensional surfaces, $\Sigma^U_{r,\varepsilon}$ and $\Sigma^S_{r,\varepsilon}$, since they are $\mathcal{O}(\varepsilon)$ away from their unperturbed counterparts. Let r_ε^U and r_ε^S denote two points on the surfaces $\Sigma_{r,\varepsilon}^U$ and $\Sigma_{r,\varepsilon}^S$, respectively, that have the same values of τ_1 and τ_2. Geometrically, the distance between the points r_ε^U and r_ε^S is

$$(45)\quad d(\tau_1,\tau_2;\omega_1,\beta_1,\beta_2;\varepsilon) \equiv \frac{\left(\nabla_{(q,p)} H(q_h^{\mathrm{up}}(0), p_h^{\mathrm{up}}(0), \tau_1, \tau_2)\right) \cdot (q_\varepsilon^U - q_\varepsilon^S, p_\varepsilon^U - p_\varepsilon^S)^T}{\|\nabla_{(q,p)} H(q_h^{\mathrm{up}}(0), p_h^{\mathrm{up}}(0), \tau_1, \tau_2), \|},$$

where $\cdot$ denotes the usual inner product in $\mathbb{R}^2$. Finally, we use the fact that the leading order term $d(\tau_1,\tau_2;\omega_1,\beta_1,\beta_2;0)$ in the Taylor expansion of the numerator of $d(\tau_1,\tau_2;\omega_1,\beta_1,\beta_2;\varepsilon)$ about $\varepsilon = 0$ is identically zero since the unperturbed unstable and stable manifolds are coincident, and we use Melnikov's original idea. These lead to

$$(46)\qquad d(\tau_1,\tau_2;\omega_1,\beta_1,\beta_2;\varepsilon) = \varepsilon \frac{M(\tau_1,\tau_2;\omega_1,\beta_1,\beta_2)}{\|\nabla_{(q,p)} H(q_h^{\mathrm{up}}(0), p_h^{\mathrm{up}}(0), \tau_1, \tau_2)\|} + \mathcal{O}(\varepsilon^2)$$

as $\varepsilon \to 0$, where the Melnikov function (i.e., the numerator of $\frac{\partial d}{\partial \varepsilon}|_{\varepsilon=0}$) is

$$(47)\qquad M(\tau_1,\tau_2;\omega_1,\beta_1,\beta_2) = \int_{-\infty}^{\infty} \left(\nabla_{(\tau_1,\tau_2)} H \cdot \mathbf{g}^\tau\right) (q_h^{\mathrm{up}}(t), p_h^{\mathrm{up}}(t), \tau_1, \tau_2) dt,$$

as reported above in (36).

REFERENCES

[1] V. I. Arnold, V. V. Kozlov, and A. I. Neihstadt, eds., *Dynamical Systems* III, Encyclopedia of Mathematical Sciences, Springer-Verlag, New York, 1988.

[2] F. Battelli, *Heteroclinic orbits in singular systems: A unifying approach*, J. Dynamics Differential Equations, 6 (1994), pp. 147–173.

[3] A. Bose, *Existence and stability of travelling waves in coupled nerve axon equations*, Ph.D. thesis, Division of Applied Mathematics, Brown University, Providence, RI, 1994.

[4] A. Bose and C. K. R. T. Jones, *Stability of the in-phase travelling wave solution in a pair of coupled nerve fibers*, Indiana Univ. Math. J., 44 (1995), pp. 189–220.

[5] G. Carpenter, *A geometric approach to singular perturbation problems with applications to nerve impulse equations*, J. Differential Equations, 23 (1977), pp. 335–367.

[6] S. N. Chow and X. B. Lin, *Bifurcation of a homoclinic orbit with a saddle-center equilibrium*, Differential Integral Equations, 3 (1990), pp. 435–466.

[7] B. Deng, *The Silnikov problem, exponential expansion, strong λ-lemma, C^1 linearization, and homoclinic bifurcation*, J. Differential Equations, 79 (1989), pp. 189–231.

[8] ———, *Homoclinic bifurcation with nonhyperbolic equilibria*, SIAM J. Math. Anal., 21 (1990), pp.693–720.

[9] ———, *The existence of infinitely many traveling front and back waves in the Fitzhugh Nagumo equations*, SIAM J. Math. Anal., 22 (1991), pp. 1631–1650.

[10] Z. C. Feng and P. R. Sethna, *Symmetry-breaking bifurcations in resonant surface waves*, J. Fluid Mech., 199 (1989), pp. 495–518.

[11] ———, *Global bifurcation and chaos in parametrically forced systems with one-one resonance*, Dynamics Stability Systems, 5 (1990), p. 201.

[12] Z. C. Feng and S. Wiggins, *On the existence of chaos in a class of two-degree-of-freedom, damped, parametrically forced mechanical systems with broken $O(2)$ symmetry*, Z. Angew. Math. Phys., 44 (1993), pp. 201–248.

[13] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.

[14] R. Gardner and J. Smoller, *Travelling wave solutions of predator-prey systems with singularly-perturbed diffusion*, J. Differential Equations, 47 (1983), pp. 133–161.

[15] J. Hale and K. Sakamoto, *Existence and stability of transition layers*, Japan J. Appl. Math., 5 (1988), pp. 367–405.

[16] G. Haller and S. Wiggins, *Orbits homoclinic to resonances: The Hamiltonian case*, Phys. D, 66 (1993), pp. 298–346.

[17] S. Hastings, *On the travelling waves of the Hodgkin-Huxley equations*, Arch. Rational Mech. Anal., 60 (1976), pp. 229–257.

[18] M. W. Hirsch, C. C. Pugh, and M. Shub, *Invariant Manifolds*, Lecture Notes in Mathematics, Vol. 583, Springer-Verlag, New York, 1983.

[19] C. K. R. T. Jones, *Stability of the travelling wave solutions of the Fitzhugh-Nagumo system*, Trans. Amer. Math. Soc., 286 (1984), pp. 431–469.

[20] C. K. R. T. Jones, N. Kopell, and R. Langer, *Construction of the Fitzhugh-Nagumo pulse using differential forms*, in Patterns and Dynamics in Reactive Media, H. Swinney, G. Aris, and D. Aronson, eds., IMA Volumes in Mathematics and its Applications, Vol. 37, Springer-Verlag, New York, 1991, pp. 101–116.

[21] C. K. R. T. Jones and N. Kopell, *Tracking invariant manifolds with differential forms*, J. Differential Equations, 108 (1994), pp. 64–88.

[22] T. J. Kaper and S. Wiggins, *On the structure of separatrix-swept regions in singularly-perturbed Hamiltonian systems*, Differential Integral Equations, 5 (1992), pp. 1363–1381.

[23] T. J. Kaper and G. Kovačič, *Multi-bump orbits homoclinic to resonance bands*, Trans. Amer. Math. Soc, to appear.

[24] G. Kovačič, *Hamiltonian dynamics of orbits homoclinic to a resonance band*, Phys. Lett. A, 167 (1992), pp. 137–142.

[25] ———, *Dissipative dynamics of orbits homoclinic to a resonance band*, Phys. Lett. A, 167 (1992), pp. 143–150.

[26] ———, *Singular perturbation theory for homoclinic orbits in a class of near-integrable Hamiltonian systems*, J. Dynamics Differential Equations, 5 (1993), pp. 559–597.

[27] G. Kovačič and S. Wiggins, *Orbits homoclinic to resonances with an application to chaos in a model of the forced and damped Sine-Gordon equation*, Phys. D, 57 (1992), pp. 185–225.

[28] X. B. Lin, *Shadowing lemma and singularly-perturbed boundary value problems*, SIAM J. Appl. Math., 49 (1989), pp. 26–54.

[29] H. Oka, *Singular perturbations of autonomous ordinary differential equations and heteroclinic bifurcations*, in Dynamical Systems, Longman Scientific and Technical, John Wiley, New York, 1990, pp. 159–194.

[30] K. Palmer, *Transversal heteroclinic points and Cherry's example of a nonintegrable Hamiltonian system*, J. Differential Equations, 65 (1986), pp. 321–360.

[31] C. Robinson, *Sustained resonance for a nonlinear system with slowly varying coefficients*, SIAM J. Math. Anal., 14 (1983), pp. 847–860.

[32] K. Sakamoto, *Invariant manifolds in singular perturbation problems for ordinary differential equations*, Proc. Roy. Soc. Edinburgh, Sect. A, 116 (1990), pp. 45–78.

[33] P. Szmolyan, *Heteroclinic and homoclinic orbits in singular perturbation problems*, J. Differential Equations, 92 (1991), pp. 255–281.

[34] S. K. Tin, N. Kopell, and C. K. R. T. Jones, *Invariant manifolds and singularly perturbed boundary value problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1558–1576.

[35] S. Wiggins, *Global Bifurcations and Chaos: Analytical Methods*, Springer-Verlag, New York, 1988.

SIAM J. MATH. ANAL.
Vol. 27, No. 1, pp. 578–593, January 1996

ON SUBDIVISION INTERPOLATION SCHEMES*

GUSTAF GRIPENBERG†

Abstract. Subdivision cardinal interpolation schemes that preserve functions of positive type are shown to be related to orthonormal multiresolutions. The interpolating function is the solution to a certain optimization problem, and this makes it possible to derive error estimates, in particular for Lagrange iterative interpolation schemes.

Key words. interpolation, subdivision, error estimate, wavelet, positive type, Lagrange

AMS subject classifications. 65D05, 42C15

1. Introduction.

The purpose of this paper is to study subdivision cardinal interpolation schemes of the form

$$f(2^{-j-1}m) = 2\sum_{k\in\mathbb{Z}}\gamma(m-2k)f(2^{-j}k), \quad m\in\mathbb{Z}, \quad j\geq 0, \tag{1}$$

where the sequence $\{\gamma(k)\}_{k\in\mathbb{Z}}$ is the mask that determines the interpolation scheme in question. (The normalizing factor 2 is for convenience introduced here so that it does not appear in the equations one obtains after taking Fourier transforms.) Such schemes have been studied by several authors, and various conditions for the mask to generate an interpolation scheme are known; see e.g., [1], [5], [6], [16], and the references mentioned there. It is clear that if the restriction $F = f_{|\mathbb{Z}}$ of f to the integers is known, then one finds from (1) the values of f at the half-integer values $\mathbb{Z}+\frac{1}{2}$ (by taking $j=0$), then the values at $\mathbb{Z}+\frac{1}{4}$ and $\mathbb{Z}+\frac{3}{4}$ (by taking $j=1$), and so on. (If the values of f are not given on $\mathbb{Z}$ but on some other set of evenly spaced points, one can use a simple transformation of the argument to reduce the problem to the one considered here.) If $m=2p$ in (1) is even, then one has

$$f(2^{-j}p) = 2\gamma(0)f(2^{-j}p) + 2\sum_{\substack{k\in\mathbb{Z}\\ k\neq p}}\gamma(2(p-k))f(2^{-j}k).$$

Since we are studying an interpolation and not a refinement scheme (that is, we do not want to change values of f already calculated), we have to require that

$$\gamma(2k) = \frac{1}{2}\delta_{0,k}, \quad k\in\mathbb{Z}$$

(where $\delta_{i,j}=1$ if $i=j$ and 0 otherwise).

Observe that the interpolation scheme in (1) is linear and translation invariant. Therefore, it suffices to find the fundamental interpolation function Φ that interpolates the sequence $\{\delta_{k,0}\}_{k\in\mathbb{Z}}$. Provided the sequence $\{F(n)\}_{n\in\mathbb{Z}}$ is known, one can then take as the desired interpolant $I_\Phi(F)$ the function

$$I_\Phi(F)(\underline{t}) \overset{\text{def}}{=} \sum_{k\in\mathbb{Z}}F(k)\Phi(\underline{t}-k).$$

* Received by the editors August 23, 1993; accepted for publication (in revised form) June 6, 1994.

† Department of Mathematics, University of Helsinki, P.O. Box 4 (Regeringsgatan 15), 00014 University of Helsinki, Finland (gustaf.gripenberg@helsinki.fi).

(Here and below we use underlined arguments like $\underline{t}$ to denote arguments used for defining functions so that, for example, $\underline{t}^2$ denotes the function that maps each real number t to the real number t^2.) It turns out (see [4, pp. 208–209]) that Φ should satisfy the equation

$$\Phi(\underline{t}) = 2\sum_{k\in\mathbb{Z}} \gamma(k)\Phi(2\underline{t}-k), \quad \Phi(\underline{k}) = \delta_{\underline{k},0}, \tag{2}$$

and we let Φ be defined by this relation. From (2), one immediately sees that

$$\gamma(\underline{k}) = \tfrac{1}{2}\Phi(\tfrac{\underline{k}}{2}). \tag{3}$$

We will only consider the case where Φ and hence γ are real functions, and the main restriction we put on Φ (in some of the results below) is that it is of positive type (and continuous), which just means that $\hat{\Phi}$ and $\hat{\gamma}$ are nonnegative functions.

Recall that a function $g : \mathbb{A} \to \mathbb{C}$ (where $\mathbb{A}$ is some additive group) is said to be of positive type if

$$\sum_{i=1}^{M}\sum_{j=1}^{M} g(t_i - t_j)c_i\overline{c_j} \geq 0, \quad t_i \in \mathbb{A}, \quad c_i \in \mathbb{C}, \quad i = 1, 2, \ldots, M$$

(see, e.g., [12]). We say that the subdivision interpolation scheme given by (1) is of positive type if the fundamental interpolation function Φ given by (2) is of positive type. It is straightforward to check that if this is the case, then $I_\Phi(F)$ is of positive type whenever the sequence F is of positive type.

One way to construct a subdivision interpolation scheme of positive type is to take

$$\Phi(\underline{t}) = \int_{\mathbb{R}} \varphi(s+\underline{t})\varphi(s)\,\mathrm{d}s, \tag{4}$$

where φ is the scaling function of an orthonormal multiresolution (or a multiresolution analysis) of $L^2(\mathbb{R};\mathbb{R})$ (for simplicity we consider only the real case), i.e., the pair $(\{V_m\}_{m\in\mathbb{Z}}, \varphi)$ satisfies the following properties:

(5) $\varphi \in L^2(\mathbb{R};\mathbb{R})$ and V_m is, for each $m \in \mathbb{Z}$, the closed subspace of $L^2(\mathbb{R};\mathbb{R})$ spanned by $\{\varphi(2^{-m}\underline{t} - k)\}_{k\in\mathbb{Z}}$;

(6) $V_m \subset V_{m-1}$, $m \in \mathbb{Z}$;

(7) $\lim_{m\to-\infty} V_m = L^2(\mathbb{R};\mathbb{R})$, i.e., $\lim_{m\to-\infty} P_m f = f$ for every $f \in L^2(\mathbb{R};\mathbb{R})$, where P_m is the orthogonal projection of $L^2(\mathbb{R};\mathbb{R})$ onto V_m;

(8) $\{\varphi(\underline{t} - k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis in V_0.

(Concerning (7), note that it follows from the other assumptions that for all f one has $\lim_{m\to\infty} P_m f = 0$ as well; see, e.g., [2].) By (6) and (8), we have

$$\varphi(\underline{t}) = 2\sum_{k\in\mathbb{Z}} \alpha(k)\varphi(2\underline{t} - k), \tag{9}$$

where the filter α is given by

$$\alpha(\underline{k}) = \int_{\mathbb{R}} \varphi(s)\varphi(2s-\underline{k})\,\mathrm{d}s.$$

It is straightforward to check that if (4) and (8) hold, then Φ is a continuous function of positive type. Combining (4) and (9), we conclude that

$$\gamma(\underline{k}) = \sum_{j\in\mathbb{Z}} \alpha(\underline{k}+j)\alpha(j). \tag{10}$$

Note that in practice, one often follows the reverse reasoning: starting from Φ and γ, one constructs φ and α such that (4) and (10) hold.

If we were to consider a complex multiresolution (of $L^2(\mathbb{R};\mathbb{C})$), then the only difference is that we would introduce complex conjugates at appropriate places in the formulas above. For further results on multiresolutions, see e.g., [4], [10], and [11].

We shall first give necessary and sufficient conditions for a sequence γ to be the mask of an interpolation scheme of positive type, and we shall prove (the not very surprising result) that essentially every subdivision interpolation scheme of positive type is associated with a multiresolution scheme by means of (4).

It is possible to prove that $\Phi \in L^2(\mathbb{R};\mathbb{R})$ and $\gamma \in \ell^2(\mathbb{Z};\mathbb{R})$. In general, one does not have $\Phi \in L^1(\mathbb{R};\mathbb{R})$ and $\gamma \in \ell^1(\mathbb{Z};\mathbb{R})$, which can be seen by considering the example $\varphi(\underline{t}) = \sin(\pi\underline{t})/(\pi\underline{t})$, in which case $\Phi = \varphi$ and $\alpha(\underline{k}) = \gamma(\underline{k}) = \sin(\pi\underline{k}/2)/(\pi\underline{k})$. Below, however, we shall make the stronger assumption that $\sum_{k\in\mathbb{Z}} \log(|k|+1)|\gamma(k)| < \infty$. Observe that if one wants the interpolation scheme to reproduce constants, then one must have $\sum_{k\in\mathbb{Z}} \gamma(k) = 1$, and this is also a necessary condition for (2) to have a nontrivial integrable solution.

The second and main problem is to find error estimates. To accomplish this, we first consider another problem. Find a norm $\|\bullet\|$ so that

$$\|I_\Phi(F)\| \le \|g\|, \quad \text{for all functions } g \text{ with } \quad g(n) = F(n), \quad n \in \mathbb{Z}.$$

It turns out that one should choose the norm to be of the form $\|\underline{g}\|^2 = \int_{\mathbb{R}} |\hat{\underline{g}}(\omega)|^2 \eta(\omega)\,\mathrm{d}\omega$ (where η is such that $\eta(\underline{\omega})\hat{\Phi}(\underline{\omega})$ is bounded and periodic), and one gets an error estimate of the form $\|f - I_\Phi(f_{|\mathbb{Z}})\| \le \|f\|$. From this inequality one can derive more useful results where one gets inequalities involving other norms (in, e.g., Sobolev spaces) of the error, but in any case, one must have enough information about the function η. These estimates become more precise if Φ becomes more regular. Observe that these results do not assume that the fundamental interpolation function Φ is of positive type.

For other approaches to error estimates involving wavelets, see, e.g., [14] and the references mentioned there.

2. Statement of results.

It is clear that the interpolation scheme (1) is easiest to use if γ has compact support, but nevertheless we state some of our results in greater generality. First, we study what the necessary assumptions on γ are. We define the Fourier transform of a function in $L^1(\mathbb{R};\mathbb{C})$ to be $\hat{f}(\underline{\omega}) = \int_{\mathbb{R}} \mathrm{e}^{-2\pi\mathrm{i}\underline{\omega}t} f(t)\,\mathrm{d}t$, and we use corresponding definitions for periodic functions and sequences, that is, we have the factor 2π in the exponent and not somewhere else. By $\widehat{L^1}(\mathbb{R};\mathbb{R})$, we denote the set of those functions in $C_0(\mathbb{R};\mathbb{R})$ (the set of continuous functions with limit 0 at $\pm\infty$) that are Fourier transforms of functions in $L^1(\mathbb{R};\mathbb{C})$.

PROPOSITION 1. *Assume that* $\Phi \in C(\mathbb{R};\mathbb{R})$ *is of positive type with* $\Phi(\underline{k}) = \delta_{0,\underline{k}}$*, and let* γ *be given by* (3). *Then* $\Phi \in L^2(\mathbb{R};\mathbb{R}) \cap \widehat{L^1}(\mathbb{R};\mathbb{R})$, $\gamma \in \ell^2(\mathbb{Z};\mathbb{R})$,

$$\hat{\gamma}(\underline{\omega}) + \hat{\gamma}(\underline{\omega} + \tfrac{1}{2}) \overset{a.e.}{=} 1, \tag{11}$$

and

$$\gamma \textit{ is of positive type.} \tag{12}$$

If, in addition,

$$\sup_{|s|\geq \underline{t}} |\Phi(s)| \in L^1(\mathbb{R}^+;\mathbb{R}) \tag{13}$$

and (2) *holds, then* $\gamma \in \ell^1(\mathbb{Z};\mathbb{R})$ *and*

$$\sum_{k\in\mathbb{Z}} \gamma(k) = 1, \tag{14}$$

and

(15) *there is a compact set* $C \subset \mathbb{R}$ *with* $\mathrm{m}(C) = 1$ *such that* $\hat{\gamma}$ *does not vanish on* $\bigcup_{k=1}^{\infty} 2^{-k}C$*, and for each* $\omega \in [0,1]$*, there is an integer* j *such that* $\omega + j \in C$*,*

The following result is closely related to corresponding results for multiresolutions where an assumption of the form (15) is seen to be of crucial importance; see [2].

THEOREM 2. *Let* $\gamma \in \ell^1(\mathbb{Z};\mathbb{R})$ *be such that* (11), (12), (14), *and* (15) *hold and*

$$\sum_{k\in\mathbb{Z}} \log(|k|+1)|\gamma(k)| < \infty. \tag{16}$$

Then there exists a function $\Phi \in C_0(\mathbb{R};\mathbb{R})$ *of positive type such that* (2) *holds. Moreover, there is a multiresolution* $(\{V_m\}_{m\in\mathbb{Z}}, \varphi)$ *of* $L^2(\mathbb{R};\mathbb{R})$ *such that* (4) *holds.*

If γ *has compact support, so do* Φ *and* φ.

Condition (15) can be given in other, equivalent forms; see, e.g., [4, Thm. 6.3.6] for the corresponding multiresolution case. It follows from the argument in [2, p. 452] that it is satisfied provided $\hat{\gamma}$ is continuous and $\hat{\gamma}(\omega) \neq 0$ for $|\omega| \leq \frac{1}{6}$. On the other hand, the standard example of a mask γ that satisfies all required conditions of Theorem 2 except (15) is the one where $\gamma(0) = \frac{1}{2}$, $\gamma(\pm 3) = \frac{1}{4}$, and $\gamma(k) = 0$ for all other k, in which case we have $\hat{\gamma}(\underline{\omega}) = (1 + \cos(6\pi\underline{\omega}))/2$.

The main problem considered in this paper is, however, what error estimates one can derive for subdivision interpolation schemes, and a related question is to what problem, if any, the subdivision interpolation scheme is an extremal solution. For example, the cubic spline interpolant minimizes the L^2-norm of the second derivative. In general, we cannot, of course, expect to get anything as simple as this, but we can get something that enables us to derive error estimates.

We use the notation $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, that is, a function with domain $\mathbb{T}$ can equivalently be considered to be a function defined on $\mathbb{R}$ which is periodic with period 1. For example, when we below assume that $\eta(\underline{\omega})\hat{\Phi}(\underline{\omega}) \in L^\infty(\mathbb{T};\mathbb{R})$, this means that $\eta(\underline{\omega})\hat{\Phi}(\underline{\omega}) \in L^\infty(\mathbb{R};\mathbb{R})$ and $\eta(\underline{\omega}+1)\hat{\Phi}(\underline{\omega}+1) \overset{\text{a.e.}}{=} \eta(\underline{\omega})\hat{\Phi}(\underline{\omega})$.

THEOREM 3. *Let* $\Phi \in L^2(\mathbb{R};\mathbb{R}) \cap \widehat{L^1}(\mathbb{R};\mathbb{R})$ *be a fundamental interpolation function, i.e.,* $\Phi(\underline{k}) = \delta_{0,\underline{k}}$. *Let* $\eta : \mathbb{R} \to \mathbb{R}^+$ *be a nonnegative measurable function* *such that* $\eta(\underline{\omega})\hat{\Phi}(\underline{\omega}) \in L^{\infty}(\mathbb{T};\mathbb{R})$. *If* $F \in \ell^1(\mathbb{Z};\mathbb{R})$, *then*

$$\int_{\mathbb{R}} \left|\widehat{I_{\Phi}(F)}(\omega)\right|^2 \eta(\omega)\,\mathrm{d}\omega + \int_{\mathbb{R}} \left|\hat{g}(\omega) - \widehat{I_{\Phi}(F)}(\omega)\right|^2 \eta(\omega)\,\mathrm{d}\omega = \int_{\mathbb{R}} |\hat{g}(\omega)|^2 \eta(\omega)\,\mathrm{d}\omega, \tag{17}$$

for all $g \in L^2(\mathbb{R};\mathbb{R}) \cap \widehat{L^1}(\mathbb{R};\mathbb{R})$ *satisfying* $\int_{\mathbb{R}} |\hat{g}(\omega)|^2 \eta(\omega)\,\mathrm{d}\omega < \infty$ *and* $g(n) = F(n)$ *for* $n \in \mathbb{Z}$. *In particular,*

$$\int_{\mathbb{R}} \left|\widehat{I_{\Phi}(F)}(\omega)\right|^2 \eta(\omega)\,\mathrm{d}\omega \leq \int_{\mathbb{R}} |\hat{g}(\omega)|^2 \eta(\omega)\,\mathrm{d}\omega,$$

and if $f \in L^2(\mathbb{R};\mathbb{R}) \cap \widehat{L^1}(\mathbb{R};\mathbb{R})$, $\int_{\mathbb{R}} \left|\hat{f}(\omega)\right|^2 \eta(\omega)\,\mathrm{d}\omega < \infty$, *and* $f_{|\mathbb{Z}} \in \ell^1(\mathbb{Z};\mathbb{R})$, *then*

$$\int_{\mathbb{R}} \left|\hat{f}(\omega) - \widehat{I_{\Phi}(f_{|\mathbb{Z}})}(\omega)\right|^2 \eta(\omega)\,\mathrm{d}\omega \leq \int_{\mathbb{R}} \left|\hat{f}(\omega)\right|^2 \eta(\omega)\,\mathrm{d}\omega. \tag{18}$$

Note that we do not assume here that the interpolation scheme is of positive type and that if η grows sufficiently rapidly, then it follows from the assumption $\int_{\mathbb{R}} |\hat{g}(\omega)|^2 \eta(\omega)\,\mathrm{d}\omega < \infty$ that $\hat{g} \in L^1(\mathbb{R};\mathbb{C})$. Recall also that by scaling these error estimates (as well as those given below), one gets results for the interpolation of a function at nodes $2^{-j}\underline{k}$, where one can then let $j \to \infty$.

Let us as an example consider the well-known case of cubic splines. In this case, the basic interpolation function Φ satisfies (4), where φ is the scaling function or father wavelet of the Battle–Lemarié wavelets constructed from the piecewise linear splines by orthonormalization, that is,

$$\hat{\varphi}(\underline{\omega}) = \frac{\sqrt{3}}{\pi^2} \frac{\sin(\pi\underline{\omega})^2}{\underline{\omega}^2\sqrt{1 + 2\cos(\pi\underline{\omega})^2}};$$

see [4, §5.4]. Since $\hat{\Phi}(\underline{\omega}) = |\hat{\varphi}(\underline{\omega})|^2$, we see that we can take $\eta(\underline{\omega}) = \underline{\omega}^4$, that is, the cubic splines minimize the L^2-norm of the second derivative. It is straightforward to extend this result to the other odd splines, that is, interpolation with splines of order $2N-1$ leads to $\eta(\underline{\omega}) = \underline{\omega}^{2N}$.

Next we consider the so-called Lagrange iterative interpolation schemes (see, e.g., [5]) where the filter γ is chosen so that $\gamma(k) = 0$ when $|k| \geq 2N$ and (1) holds for all polynomials of degree at most $2N-1$, i.e., these polynomials are reconstructed exactly in the interpolation scheme. As shown in [13, Thm. 5.2], the fundamental interpolation function Φ_N of such a Lagrange subdivision interpolation scheme of order $2N-1$, and the scaling function φ constructed in [3] with support width $2N-1$ such that the corresponding wavelets have N vanishing moments, are linked via (4). Thus we have, by [4, Prop. 6.1.2, p. 216],

$$\widehat{\Phi_N}(\underline{\omega}) = \left(\frac{\sin(\pi\underline{\omega})}{\pi\underline{\omega}}\right)^{2N} \Pi_{j=1}^{\infty} L_N(2^{-j}\underline{\omega}),$$

where

$$L_N(\underline{\omega}) = P_N\left(\sin(\pi\underline{\omega})^2\right) \quad \text{and} \quad P_N(\underline{y}) = \sum_{k=0}^{N-1} \binom{N-1+k}{k} \underline{y}^k.$$

Thus we see that the function η in Theorem 3 can be chosen to be

$$\eta_N(\underline{\omega}) = \frac{\underline{\omega}^{2N}}{\Pi_{j=1}^{\infty} L_N(2^{-j}\underline{\omega})}, \tag{19}$$

and we have the following estimate that follows easily from the ones in [4, §7.1].

PROPOSITION 4. *If $N \geq 1$ and η_N is given by* (19), *then*

$$C_N \frac{\omega^{2N}}{(1+|\omega|)^{\log_2(P_N(3/4))}} \leq \eta_N(\omega) \leq \omega^{2N}, \quad \omega \in \mathbb{R}, \tag{20}$$

where

$$C_N > \frac{N}{16^N e^{2N}}.$$

By a result in [4, p. 226], we have

$$\frac{1}{\sqrt{N}} 3^{N-1} \leq P_N\big(\tfrac{3}{4}\big) \leq 3^{N-1}.$$

It is not difficult to see that the exponents in the estimate (20) are the best possible ones; see [4, Lem. 7.1.3] for the lower bound and take $\omega = 2^k$ in the upper bound.

It is clear that (18) as such is not very useful, and therefore we record some quite easy consequences of it.

THEOREM 5. *Let the assumptions of Theorem* 3 *hold and let $f \in L^1(\mathbb{R};\mathbb{R}) \cap \widehat{L^1}(\mathbb{R};\mathbb{R})$ be such that $f_{|\mathbb{Z}} \in \ell^1(\mathbb{Z};\mathbb{R})$ and $\int_{\mathbb{R}} |\hat{f}(\omega)|^2 \eta(\omega)\,\mathrm{d}\omega < \infty$. Then for every $q \geq 0$, we have*

$$\int_{|\omega|\geq 1/2} |\omega|^q \big|\hat{f}(\omega) - \widehat{I_\Phi(f_{|\mathbb{Z}})}(\omega)\big|^2 \,\mathrm{d}\omega \leq \operatorname*{ess\,sup}_{|\omega|\geq 1/2} \frac{|\omega|^q}{\eta(\omega)} \int_{\mathbb{R}} \big|\hat{f}(\omega)\big|^2 \eta(\omega)\,\mathrm{d}\omega \tag{21}$$

and

$$\int_{|\omega|\geq 1/2} |\omega|^q \big|\hat{f}(\omega) - \widehat{I_\Phi(f_{|\mathbb{Z}})}(\omega)\big| \,\mathrm{d}\omega \leq \sqrt{\int_{|\omega|\geq 1/2} \frac{|\omega|^{2q}}{\eta(\omega)}\,\mathrm{d}\omega \int_{\mathbb{R}} \big|\hat{f}(\omega)\big|^2 \eta(\omega)\,\mathrm{d}\omega}. \tag{22}$$

If

$$C_\eta \overset{\text{def}}{=} \operatorname*{ess\,sup}_{|\omega|\leq 1/2} \sum_{\substack{k\in\mathbb{Z}\\ k\neq 0}} \frac{1}{\eta(\omega+k)} < \infty, \tag{23}$$

then

$$\int_{-1/2}^{1/2} \big|\hat{f}(\omega) - \widehat{I_\Phi(f_{|\mathbb{Z}})}(\omega)\big|^2 \,\mathrm{d}\omega \leq C_\eta \int_{\mathbb{R}} \big|\hat{f}(\omega)\big|^2 \eta(\omega)\,\mathrm{d}\omega \tag{24}$$

and

$$\int_{-1/2}^{1/2} \big|\hat{f}(\omega) - \widehat{I_\Phi(f_{|\mathbb{Z}})}(\omega)\big| \,\mathrm{d}\omega \leq \sqrt{C_\eta \int_{\mathbb{R}} \big|\hat{f}(\omega)\big|^2 \eta(\omega)\,\mathrm{d}\omega}. \tag{25}$$

Note that by Proposition 4, we have (23) at least for Lagrange iterative interpolation schemes and one gets an upper bound for the constant $\operatorname{ess\,sup}_{|\omega|\geq 1/2}(|\omega|^q/\eta(\omega))$ in (21). From Proposition 4, one can get some results about when the crucial constant $\int_{|\omega|\geq 1/2}(|\omega|^{2q}/\eta(\omega))\,\mathrm{d}\omega$ in (22) is finite as well. It is possible, however, to get better and more general results for this term, and we shall consider that problem below.

Observe also that if $\kappa = n + \nu$, where $n \in \mathbb{N}$ and $0 \leq \nu < 1$, and we define

$$C^\kappa(\mathbb{R};\mathbb{R}) \overset{\text{def}}{=} \Big\{\, f \in C^n(\mathbb{R};\mathbb{R}) \;\Big|\; \|f\|_{C^\kappa(\mathbb{R})} \overset{\text{def}}{=} \sup_{t\in\mathbb{R}} |f(t)| + \sup_{t,s\in\mathbb{R},t\neq s} \frac{|f^{(n)}(t)-f^{(n)}(s)|}{|t-s|^\nu} < \infty \,\Big\},$$

then it is easy to check that

$$\|f\|_{C^{\kappa}(\mathbb{R})} \leq \sup_{t\in\mathbb{R}}|f(t)| + 2^{\lceil\kappa\rceil}\pi^{\kappa}\int_{\mathbb{R}}|\omega|^{\kappa}|\hat{f}(\omega)|\,\mathrm{d}\omega$$

when $f \in \widehat{L^1}(\mathbb{R};\mathbb{R})$, for example. As a partial converse, one can prove that if $\hat{f}(\omega) \geq 0$ for $\omega \in \mathbb{R}$, and $f \in C^{\kappa}(\mathbb{R};\mathbb{R})$ with $\kappa > 0$, then $\int_{\mathbb{R}}|\omega|^{\kappa'}\hat{f}(\omega)\,\mathrm{d}\omega < \infty$ for all $\kappa' \in [0,\kappa)$. This is one reason for studying error estimates of the form (22).

THEOREM 6. *Assume that $\Phi \in C(\mathbb{R};\mathbb{R})$ is of positive type such that* (2) *and* (13) *hold, and let γ be given* (3). *Suppose there is a positive integer M such that $\underline{k}^{2M+1}\gamma(\underline{k}) \in \ell^1(\mathbb{Z};\mathbb{R})$,*

$$\hat{\gamma}^{(j)}\left(\tfrac{1}{2}\right) = 0, \quad j = 0,1,\dots,2M-1, \tag{26}$$

and

$$\hat{\gamma}^{(2M)}\left(\tfrac{1}{2}\right) \neq 0. \tag{27}$$

If one takes $\eta(\underline{\omega}) \overset{\text{def}}{=} \sin(\pi\underline{\omega})^{2M}/\hat{\Phi}(\underline{\omega})$ and $r \geq 0$, then the following three conditions are equivalent:

(i) $\displaystyle\int_{|\omega|\geq 1/2}\frac{|\omega|^r}{\eta(\omega)}\,\mathrm{d}\omega < \infty$;

(ii) $\displaystyle\int_{\mathbb{R}}|\omega|^r|\hat{\Phi}(\omega)|\,\mathrm{d}\omega < \infty$;

(iii) *$r < 2M - \log_2(\rho)$, where ρ is the spectral radius of the operator $A : C([0,1];\mathbb{C}) \to C([0,1];\mathbb{C})$ defined by*

$$(Af)(\underline{\omega}) = a\left(\tfrac{\underline{\omega}}{2}\right)f\left(\tfrac{\underline{\omega}}{2}\right) + a\left(\tfrac{\underline{\omega}+1}{2}\right)f\left(\tfrac{\underline{\omega}+1}{2}\right),$$

where

$$a(\underline{\omega}) \overset{\text{def}}{=} \frac{\hat{\gamma}(\underline{\omega})}{\cos(\pi\underline{\omega})^{2M}}.$$

This is not really a new result because equivalences similar to (ii) $\Leftrightarrow$ (iii) can be found in, e.g., [7]–[9] and [15].

Observe that if γ has compact support, then the spectral radius ρ of A is the spectral radius of a matrix and can easily be calculated; see the proof below. For a method to compute ρ when γ does not have compact support, see [9].

3. Proofs.

Proof of Proposition 1. Since we assume that $\Phi \in C(\mathbb{R};\mathbb{R})$ is of positive type and $\Phi(0) = 1$, there exists by Bochner's theorem (see, e.g., [12, p. 19]) a nonnegative Borel measure μ with $\mu(\mathbb{R}) = 1$ such that

$$\Phi(\underline{t}) = \int_{\mathbb{R}} \mathrm{e}^{2\pi\mathrm{i}\underline{t}\omega}\mu(\mathrm{d}\omega).$$

Define another nonnegative measure μ_* by $\mu_*(E) = \sum_{k\in\mathbb{Z}}\mu(E+k)$ for every Borel set E. It follows that

$$\int_{[0,1)} \mathrm{e}^{-2\pi\mathrm{i}\underline{k}\omega}\mu_*(\mathrm{d}\omega) = \Phi(-\underline{k}) = \delta_{0,\underline{k}},$$

and this in turn implies that μ_* is the Lebesgue measure. This means that μ is absolutely continuous with respect to Lebesgue measure, and then we can use the

Radon–Nikodym theorem and the uniqueness of the Fourier transform to show that $\hat{\Phi} \in L^1(\mathbb{R};\mathbb{R}^+)$ and that

$$\sum_{k\in\mathbb{Z}} \hat{\Phi}(\underline{\omega}+k) \overset{\text{a.e.}}{=} 1. \tag{28}$$

In particular, it follows that

$$\hat{\Phi} \in L^p(\mathbb{R};\mathbb{R}), \quad 1 \le p \le \infty, \tag{29}$$

and we conclude that $\Phi \in \widehat{L^1}(\mathbb{R};\mathbb{R}) \cap L^2(\mathbb{R};\mathbb{R})$.

By (3), it follows that γ is of positive type as well, which again by Bochner's theorem implies that γ is the Fourier transform of a positive measure on $[0,1)$. But from (3) and (29), we can easily conclude that this measure is in fact absolutely continuous with respect to Lebesgue measure (so that $\hat{\gamma}$ is a function), and we have

$$\hat{\gamma}(\underline{\omega}) = \sum_{k\in\mathbb{Z}} \hat{\Phi}(2\underline{\omega}+2k),$$

and therefore (11) is a direct consequence of (28). Because $0 \overset{\text{a.e.}}{\le} \hat{\gamma}(\underline{\omega}) \overset{\text{a.e.}}{\le} 1$, we have $\gamma \in \ell^2(\mathbb{Z};\mathbb{R})$ by Plancherel's theorem.

Let us now assume that (13) holds. It follows that we have $\gamma \in \ell^1(\mathbb{Z};\mathbb{R})$, and from (2), we get

$$\hat{\Phi}(2\underline{\omega}) = \hat{\gamma}(\underline{\omega})\hat{\Phi}(\underline{\omega}). \tag{30}$$

From (11) and (12), it follows that we must have $0 \le \hat{\gamma}(0) \le 1$, but if $\hat{\gamma}(0) < 1$, then we see from (30) that $\hat{\Phi}(\omega) = 0$ for all $\omega \in \mathbb{R}$, which is impossible. Therefore, we have (14).

In order to prove (15), we first show that the series $\sum_{k\in\mathbb{Z}} \hat{\Phi}(\underline{\omega}+k)$ that appears in (28) converges uniformly on $[0,1]$. Therefore, we consider the function $g_n(\underline{t}) \overset{\text{def}}{=} \max\{0, \min\{n-n\underline{t}, n+n\underline{t}\}\}$ (the graph of g_n and the t-axis thus form a triangle with corners at $(\pm 1/n, 0)$ and $(0,n)$) and note that it is a standard result that $\widehat{g_n}(\underline{\omega}) = \sin(\pi\underline{\omega}/n)^2/(\pi\underline{\omega}/n)^2$. It follows that the Fourier transform of the sequence $\{c_n(k)\}_{k\in\mathbb{Z}}$, where

$$c_n(\underline{k}) = \Phi(\underline{k}) - \int_{\mathbb{R}} g_n(\underline{k}-t)\Phi(t)\,\mathrm{d}t, \tag{31}$$

is given by

$$\widehat{c_n}(\underline{\omega}) = \sum_{k\in\mathbb{Z}} (1-\widehat{g_n}(\underline{\omega}+k))\hat{\Phi}(\underline{\omega}+k). \tag{32}$$

From (13), (31), and from the uniform continuity of Φ, we are able to deduce that $\lim_{n\to\infty} \|c_n\|_{\ell^1(\mathbb{Z})} = 0$, and this gives by (32) the uniform convergence of the series in (28) because all terms are nonnegative.

It follows that $\sum_{k\in\mathbb{Z}} \hat{\Phi}(\omega+k) = 1$ for every $\omega \in \mathbb{R}$, and hence for each $\omega \in [0,1]$, there exist numbers $k_\omega \in \mathbb{Z}$ and $\epsilon_\omega > 0$ such that $\hat{\Phi}(\xi + k_\omega) > 0$ when $|\xi - \omega| < \epsilon_\omega$. Since $[0,1]$ is compact, we can choose finitely many of these points ω_j, $j = 1, 2, \dots, n$ such that $[0,1] \subset \bigcup_{j=1}^{n}(\omega_j - \epsilon_{\omega_j}, \omega_j + \epsilon_{\omega_j})$. But then we can construct the set C as the finite union of closed intervals on which Φ does not vanish. Since it follows from

(30) that

$$\hat{\Phi}(\underline{\omega}) = \hat{\Phi}(2^{-k}\underline{\omega}) \prod_{j=1}^{k} \hat{\gamma}(2^{-j}\underline{\omega}),$$

we see from (14) and the fact that $\hat{\Phi}$ does not vanish on $\mathcal{C}$ that $\hat{\gamma}$ cannot vanish on $\bigcup_{k=1}^{\infty}(2^{-k}\mathcal{C})$. Thus we have established (15). □

Proof of Theorem 2. We have $\hat{\gamma}(\underline{\omega}) - 1 = \sum_{k\in\mathbb{Z}}(\mathrm{e}^{-2\pi\mathrm{i}\underline{\omega}k} - 1)\gamma(k)$ since $\hat{\gamma}(0) = 1$ by (14), and hence

$$|\hat{\gamma}(\underline{\omega}) - 1| \le \sum_{k\in\mathbb{Z}} 2|\sin(\pi\underline{\omega}k)||\gamma(k)|.$$

Let m be a positive integer and let $\omega \in \mathbb{R}$. Now it is clear from the preceding inequality, Fubini's theorem, and the fact that $|\sin(\underline{t})| \le \min\{1, |\underline{t}|\}$ that

$$\begin{aligned}\sum_{j=m}^{\infty}|\hat{\gamma}(2^{-j}\omega) - 1| &\le 2\sum_{j=m}^{\infty}\sum_{k\in\mathbb{Z}}|\sin(2^{-j}\pi\omega k)||\gamma(k)| \\ &\le 2\sum_{k\in\mathbb{Z}}\left(\sum_{j=m}^{\lceil\log_2(\pi|\omega k|)\rceil} 1 + \sum_{j=|\lceil\log_2(\pi|\omega k|)\rceil+1-m|_+ +m}^{\infty} 2^{-j}\pi|\omega k|\right)|\gamma(k)| \qquad (33) \\ &\le 2\sum_{k\in\mathbb{Z}}\left(|\lceil\log_2(\pi|\omega k|)\rceil + 1 - m|_+ + 2^{-|m-\lceil\log_2(\pi|\omega k|)\rceil-1|_+}\right)|\gamma(k)|.\end{aligned}$$

(Here we used the notation $|\bullet|_+ = \max\{0, \bullet\}$.) We let

$$\begin{aligned}P_m(\underline{\omega}) &= \chi_{2^m\mathcal{C}}(\underline{\omega})\Pi_{j=1}^{m}\hat{\gamma}(2^{-j}\underline{\omega}), \\ P(\underline{\omega}) &= \Pi_{j=1}^{\infty}\hat{\gamma}(2^{-j}\underline{\omega}),\end{aligned}$$

and observe that because $\hat{\gamma}(0) = 1$, we may assume that 0 belongs to the interior of the set $\mathcal{C}$. From (16) and (33), we therefore conclude that the sequence $\{P_m\}_{m\in\mathbb{N}}$ converges uniformly on compact intervals toward P, and this shows that P is continuous.

An immediate consequence of assumption (15) is that if $f \in L^1(\mathbb{T};\mathbb{C})$ (that is, $f \in L^1_{\mathrm{loc}}(\mathbb{R};\mathbb{C})$ is periodic with period 1), then

$$\int_{\mathcal{C}} f(x)\,\mathrm{d}x = \int_0^1 f(x)\,\mathrm{d}x.$$

Using (11) and the fact that $\hat{\gamma}$ is periodic with period 1, we therefore get

$$\begin{aligned}\int_{\mathbb{R}} P_m(\omega)\mathrm{e}^{2\pi\mathrm{i}\underline{k}\omega}\,\mathrm{d}\omega &= \int_{2^m\mathcal{C}}\prod_{j=1}^{m}\hat{\gamma}\left(2^{-j}\omega\right)\mathrm{e}^{2\pi\mathrm{i}\underline{k}\omega}\,\mathrm{d}\omega = 2^m\int_{\mathcal{C}}\prod_{j=0}^{m-1}\hat{\gamma}(2^j\omega)\mathrm{e}^{2\pi\mathrm{i}2^m\underline{k}\omega}\,\mathrm{d}\omega \\ &= 2^m\int_0^1\prod_{j=0}^{m-1}\hat{\gamma}(2^j\omega)\mathrm{e}^{2\pi\mathrm{i}2^m\underline{k}\omega}\,\mathrm{d}\omega \\ &= 2^m\int_0^{1/2}\prod_{j=1}^{m-1}\hat{\gamma}(2^j\omega)\left(\hat{\gamma}(\omega) + \hat{\gamma}(\omega + \tfrac{1}{2})\right)\mathrm{e}^{2\pi\mathrm{i}2^m\underline{k}\omega}\,\mathrm{d}\omega \\ &= 2^{m-1}\int_0^1\prod_{j=0}^{m-2}\hat{\gamma}(2^j\omega)\mathrm{e}^{2\pi\mathrm{i}2^{m-1}\underline{k}\omega}\,\mathrm{d}\omega = \int_{\mathbb{R}} P_{m-1}(\omega)\mathrm{e}^{2\pi\mathrm{i}\underline{k}\omega}\,\mathrm{d}\omega.\end{aligned}$$

Since $\int_{\mathbb{R}} P_0(\omega)\mathrm{e}^{2\pi \mathrm{i}\underline{k}\omega}\,\mathrm{d}\omega = \int_{\mathcal{C}} \mathrm{e}^{2\pi \mathrm{i}\underline{k}\omega}\,\mathrm{d}\omega = \delta_{0,\underline{k}}$, it follows by induction that

$$\int_{\mathbb{R}} P_m(\omega)\mathrm{e}^{2\pi \mathrm{i}\underline{k}\omega}\,\mathrm{d}\omega = \delta_{0,\underline{k}}, \tag{34}$$

and in particular that

$$\|P_m\|_{L^1(\mathbb{R})} = 1,$$

(because $P_m(\underline{\omega}) \geq 0$). Thus we also conclude from Fatou's lemma that $P \in L^1(\mathbb{R};\mathbb{R})$ and $\|P\|_{L^1(\mathbb{R})} \leq 1$.

The function P is continuous and does not vanish on $\mathcal{C}$, hence there is a constant ϵ such that $P(\omega) \geq \epsilon > 0$ for all $\omega \in \mathcal{C}$. Because P_m vanishes outside $2^m\mathcal{C}$ and satisfies $P_m(\underline{\omega}) = P(\underline{\omega})/P(2^{-m}\underline{\omega})$ on $2^m\mathcal{C}$, it follows that

$$P_m(\omega) \leq \frac{P(\omega)}{\epsilon}, \quad \omega \in \mathbb{R}.$$

This inequality allows us to apply the dominated convergence theorem, and we conclude that

$$P_m \to P \quad \text{in} \quad L^1(\mathbb{R};\mathbb{R}). \tag{35}$$

Now we can choose Φ so that $\hat{\Phi} = P$. From the definition of P it follows that (30) holds. By (35), we see that (34) holds with P_m replaced by P, and this is equivalent to (28) or $\Phi(\underline{k}) = \delta_{0,\underline{k}}$, that is, Φ is a fundamental interpolation function which is of positive type because $\hat{\Phi}(\underline{\omega})$ is nonnegative.

If γ has compact support, we can use a lemma by Riesz (see [4, Lem. 6.1.3]) and the fact that γ is real and of positive type to find a real sequence α with compact support such that

$$\left|\hat{\alpha}(\underline{\omega})\right|^2 = \hat{\gamma}(\underline{\omega}).$$

We may, of course, assume that $\hat{\alpha}(0) = 1$, and we can define the function φ by $\hat{\varphi}(\underline{\omega}) = \Pi_{j=1}^{\infty}\hat{\alpha}(2^{-j}\underline{\omega})$. It follows from [4, Lem. 6.2.2] that Φ and φ have compact support as well. If γ does not have compact support, we define α and φ by $\hat{\alpha}(\underline{\omega}) = \sqrt{\hat{\gamma}(\underline{\omega})}$ and $\hat{\varphi}(\underline{\omega}) = \sqrt{\hat{\Phi}(\underline{\omega})}$, so that we in any case have

$$\left|\hat{\varphi}(\underline{\omega})\right|^2 = \hat{\Phi}(\underline{\omega})$$

and

$$\hat{\varphi}(2\underline{\omega}) = \hat{\alpha}(\underline{\omega})\hat{\varphi}(\underline{\omega}).$$

Because $\hat{\alpha} \in L^\infty(\mathbb{T};\mathbb{C})$, it is clear that $\alpha \in \ell^2(\mathbb{Z};\mathbb{R})$ and $\varphi \in L^2(\mathbb{R};\mathbb{R})$ because $\Phi \in \widehat{L^1}(\mathbb{R};\mathbb{R})$. Moreover, it follows from (28) that (8) holds (see [4, p. 132]), and hence (9) holds too. If one uses the notation in (5), this implies that (6) is satisfied. Finally, we get (7) from the fact that $\hat{\Phi}$ is continuous and $\hat{\Phi}(0) = 1$ by [4, Prop. 5.3.2]. This completes the proof. □

Proof of Theorem 3. It is clear that we have $I_\Phi(F) \in L^2(\mathbb{R};\mathbb{R}) \cap C(\mathbb{R};\mathbb{R})$ and that $\widehat{I_\Phi(F)}(\underline{\omega}) = \hat{F}(\underline{\omega})\hat{\Phi}(\underline{\omega})$. Moreover, we have $\int_{\mathbb{R}}|\widehat{I_\Phi(F)}(\omega)|^2\eta(\omega)\,\mathrm{d}\omega < \infty$. Let $h \overset{\text{def}}{=} g - I_\Phi(F)$. We observe that $h(n) = 0$ for all $n \in \mathbb{Z}$ and that $\hat{h} \in L^1(\mathbb{R};\mathbb{C})$.

Let $p(\underline{\omega}) \overset{\text{def}}{=} \eta(\underline{\omega})\hat{\Phi}(\underline{\omega})\hat{F}(\underline{\omega})$. A straightforward calculation shows that

$$\int_{\mathbb{R}} \widehat{I_\Phi(F)}(\omega)\overline{\hat{h}(\omega)}\eta(\omega)\,\mathrm{d}\omega = \int_{\mathbb{R}} \hat{F}(\omega)\hat{\Phi}(\omega)\overline{\hat{h}(\omega)}\eta(\omega)\,\mathrm{d}\omega = \int_{\mathbb{R}} p(\omega)\overline{\hat{h}(\omega)}\,\mathrm{d}\omega. \tag{36}$$

If p_* is a trigonometric polynomial, i.e., a finite linear combination of the functions $e^{2\pi i k \underline{\omega}}$, where $k \in \mathbb{Z}$, then it follows from the fact that $h(n) = 0$ for all $n \in \mathbb{Z}$ that $\int_{\mathbb{R}} p_*(\omega)\overline{\hat{h}(\omega)}\,\mathrm{d}\omega = 0$. Since $p \in L^\infty(\mathbb{T};\mathbb{C})$, it is possible to find a sequence $\{p_n(\underline{\omega})\}_{n\in\mathbb{N}}$ of trigonometric polynomials such that $p_n(\underline{\omega}) \to p(\underline{\omega})$ almost everywhere and $\sup_{n\in\mathbb{N}}\|p_n\|_{L^\infty(\mathbb{T})} < \infty$. But then we see that

$$\int_{\mathbb{R}} p(\omega)\overline{\hat{h}(\omega)}\,\mathrm{d}\omega = \lim_{n\to\infty}\int_{\mathbb{R}} p_n(\omega)\overline{\hat{h}(\omega)}\,\mathrm{d}\omega = 0.$$

If we combine this result with (36) and recall the definition of h, then we easily see that (17) holds. The remaining claims now follow immediately. □

Proof of Proposition 4. Since we have $L_N(\underline{\omega}) \geq 1$, the upper bound in (20) is obvious. It follows from [4, Lem. 7.1.8] and an argument similar to the one used in [4, Lem. 7.1.6] that for every $m \geq 1$, we get

$$\Pi_{j=1}^{\infty} L_N(2^{-j}\underline{\omega}) \leq P_N\left(\tfrac{3}{4}\right)^{m-1} P_N(1)\Pi_{j=m+1}^{\infty} L_N(2^{-j}\underline{\omega}). \tag{37}$$

We observe that by the convexity of P_N, we have on the interval $[0, \frac{1}{2}]$ the inequality $P_N(\underline{y}) \leq 1 - 2\underline{y} + 2\underline{y}P_N(\frac{1}{2}) = 1 + (2^N - 2)\underline{y}$ since $P_N(\frac{1}{2}) = 2^{N-1}$ (see [4, p. 219]), and therefore, we have for every $\omega \in \mathbb{R}$ and $m \geq 1$ satisfying $2^{-m}|\omega| \leq 1$ that

$$\Pi_{j=m+1}^{\infty}\left|L_N(2^{-j}\omega)\right| \leq P_N(1)\Pi_{j=m+2}^{\infty}\left(1 + (2^N - 2)\pi^2 4^{-j}\omega^2\right) < P_N(1)e^{2^N}. \tag{38}$$

We can choose the integer m such that $2^{-m}|\omega| \leq 1$ and $2^{m-1} \leq |\omega| + 1$ so that we get

$$P_N\left(\tfrac{3}{4}\right)^{m-1} \leq (|\omega| + 1)^{\log_2(P_N(\frac{3}{4}))}.$$

The claimed result now follows by (37) and (38) because we have

$$P_N(1) = \binom{2N-1}{N} < \frac{4^N}{\sqrt{N}};$$

see [4, p. 223]. □

Proof of Theorem 5. Inequality (21) follows directly from (18), and in order to get (22), one must first invoke Hölder's inequality. Moreover, (25) follows from (24), again by Hölder's inequality. Thus it remains to prove (24).

We let $e \overset{\mathrm{def}}{=} f - I_\Phi(f_{|\mathbb{Z}})$, and we define the function $p \in L^\infty(\mathbb{T};\mathbb{C})$ by requiring that $p(\omega) = \overline{\hat{e}(\omega)}$ when $-\frac{1}{2} < \omega \leq \frac{1}{2}$. By the same argument that was used in the proof of Theorem 3, we find that

$$\int_{\mathbb{R}} \hat{e}(\omega)p(\omega)\,\mathrm{d}\omega = 0.$$

Recalling the definition of p, we therefore get the following result with the aid of

Hölder's inequality:

$$\begin{aligned}\int_{-1/2}^{1/2} |\hat{e}(\omega)|^2 \,\mathrm{d}\omega &= -\int_{|\omega|\geq 1/2} \hat{e}(\omega)p(\omega)\,\mathrm{d}\omega \\ &\leq \sqrt{\int_{\mathbb{R}} |\hat{e}(\omega)|^2 \eta(\omega)\,\mathrm{d}\omega}\sqrt{\sum_{\substack{k\in\mathbb{Z}\\ k\neq 0}} \int_{-1/2}^{1/2} |\hat{e}(\omega)|^2 \frac{1}{\eta(\omega+k)}\,\mathrm{d}\omega} \\ &\leq \sqrt{\int_{\mathbb{R}} |\hat{e}(\omega)|^2 \eta(\omega)\,\mathrm{d}\omega}\sqrt{\int_{-1/2}^{1/2} |\hat{e}(\omega)|^2 \,\mathrm{d}\omega}\sqrt{\operatorname*{ess\,sup}_{|\omega|\leq 1/2} \sum_{\substack{k\in\mathbb{Z}\\ k\neq 0}} \frac{1}{\eta(\omega+k)}}.\end{aligned}$$

Inequality (24) now follows from Theorem 3, and the proof is completed. □

The proof of Theorem 6 is essentially the same as the proof of [8, Thm. 1], but for completeness we give it here. We begin by proving a lemma that could be derived from more general results as well.

LEMMA 7. *Let $\kappa > 0$ and let $a \in C^{\kappa}(\mathbb{T};\mathbb{R})$ be nonnegative. Then there exists an eigenvalue λ of the operator $A : C([0,1];\mathbb{C}) \to C([0,1];\mathbb{C})$, defined by*

$$Af(\underline{\omega}) = a(\tfrac{\underline{\omega}}{2})f(\tfrac{\underline{\omega}}{2}) + a(\tfrac{\underline{\omega}+1}{2})f(\tfrac{\underline{\omega}+1}{2})$$

such that $|\lambda|$ is the spectral radius of A.

Proof of Lemma 7. Assume, without loss of generality, that $\kappa \in (0,1)$. We choose a sequence $\{\epsilon_n\}_{n\in\mathbb{N}}$ of positive numbers such that $\lim_{n\to\infty} \epsilon_n = 0$ and so that the functions a_n, defined by

$$a_n(\underline{\omega}) \overset{\text{def}}{=} \frac{1}{n+1}\int_0^1 \left(\frac{\sin((n+1)\pi t)}{\sin(\pi t)}\right)^2 a(\underline{\omega} - t)\,\mathrm{d}t + \epsilon_n, \quad n \in \mathbb{N},$$

satisfy

$$a_n(\underline{\omega}) \geq a(\underline{\omega}), \quad n \in \mathbb{N}. \tag{39}$$

Furthermore, we see that the functions a_n belong to $C^{\kappa}(\mathbb{T};\mathbb{R})$ and

$$\sup_{n\in\mathbb{N}} \|a_n\|_{C^{\kappa}(\mathbb{T})} < \infty, \tag{40}$$

and

$$\lim_{n\to\infty} \|a_n - a\|_{L^{\infty}(\mathbb{T})} = 0. \tag{41}$$

We define the operator $A_n : C([0,1];\mathbb{C}) \to C([0,1];\mathbb{C})$ by

$$(A_n f)(\underline{\omega}) = a_n(\tfrac{\underline{\omega}}{2})f(\tfrac{\underline{\omega}}{2}) + a_n(\tfrac{\underline{\omega}+1}{2})f(\tfrac{\underline{\omega}+1}{2}).$$

Because $\widehat{a_n}(\underline{k}) = \max\{0, 1 - \frac{|\underline{k}|}{n+1}\}\hat{a}(\underline{k}) + \epsilon_n\delta_{0,\underline{k}}$, it follows that a_n is a trigonometric polynomial. Now if $f \in C([0,1];\mathbb{C})$ is a trigonometric polynomial, then

$$a_n(\underline{\omega})f(\underline{\omega}) = \sum_{k\in\mathbb{Z}} \mathrm{e}^{2\pi i k\underline{\omega}} \sum_{j=-n}^{n} \widehat{a_n}(j)\hat{f}(k-j),$$

where the sum is actually a finite one. Then

$$(A_nf)(\underline{\omega}) = 2\sum_{k\in\mathbb{Z}} e^{2\pi ik\underline{\omega}} \sum_{j=-n}^{n} \widehat{a_n}(j)\hat{f}(2k-j)$$

because the odd terms cancel. Thus we see that if the support of the Fourier transform of f is contained in $[-n+1, n-1]$, then the same holds true for the support of the Fourier transform of A_nf. Thus A_n maps a finite-dimensional space of trigonometric polynomials into itself, and therefore there is an eigenvalue λ_n of A_n such that $|\lambda_n| = \rho_n$, where ρ_n is the spectral radius of A_n restricted to this space. We denote the corresponding eigenfunction by v_n, and we normalize it so that $\|v_n\|_{C([0,1])} = 1$. We note that ρ_n is also equal to the spectral radius of A_n in the space $C([0,1];\mathbb{C})$ because by the nonnegativity of a_n, we have $\|A_n^m\| = \|A_n^m 1\|_{C([0,1])}$, where $\|\bullet\|$ is the operator norm in any one of these spaces. By (39), we have $\rho_n \geq \rho$ (where ρ is the spectral radius of A), and from (41), we conclude that $A_n \to A$ in operator norm as $n \to \infty$ and hence also that

$$\lim_{n\to\infty} \rho_n = \rho. \tag{42}$$

(Observe that the fact that $\rho_n \geq \rho$ is needed to derive (42) was not noted in [8].)

Define a new operator $B_n : C([0,1] \times [0,1];\mathbb{C}) \to C([0,1] \times [0,1];\mathbb{C})$ by

$$(B_nh)(\underline{\omega}, \underline{\eta}) = a_n\big(\tfrac{\underline{\omega}}{2}\big)h\big(\tfrac{\underline{\omega}}{2}, \tfrac{\underline{\eta}}{2}\big) + a_n\big(\tfrac{\underline{\omega}+1}{2}\big)h\big(\tfrac{\underline{\omega}+1}{2}, \tfrac{\underline{\eta}+1}{2}\big).$$

We can also define B in a similar way with a_n replaced by a. We note that B_n, applied to a function that does not depend on its second argument, gives the same result (as a function of its first argument) as A_n applied to the same function (with only one argument). Since $\|B_n^m\| = \|B_n^m 1\|_{C([0,1]^2)}$ and $\|A_n^m\| = \|A_n^m 1\|_{C([0,1])}$, we therefore conclude that $\|B_n^m\| = \|A_n^m\|$, and these operators have the same spectral radius. Define the function $g_n \in C([0,1] \times [0,1];\mathbb{C})$ by

$$g_n(\underline{\omega}, \underline{\eta}) = \frac{v_n(\underline{\omega}) - v_n(\underline{\eta})}{|\underline{\omega} - \underline{\eta}|^{\kappa}}.$$

Then we get

$$2^{-\kappa}B_ng_n - \rho_ng_n = b_n,$$

where

$$b_n(\underline{\omega}, \underline{\eta}) = \frac{\big(a_n\big(\tfrac{\underline{\eta}}{2}\big) - a_n\big(\tfrac{\underline{\omega}}{2}\big)\big)v_n\big(\tfrac{\underline{\eta}}{2}\big) + \big(a_n\big(\tfrac{\underline{\eta}+1}{2}\big) - a_n\big(\tfrac{\underline{\omega}+1}{2}\big)\big)v_n\big(\tfrac{\underline{\eta}+1}{2}\big)}{|\underline{\omega} - \underline{\eta}|^{\kappa}}.$$

Because $B_n \to B$ as $n \to \infty$ by (41), $\sup_{\eta\in[0,1]}|v_n(\eta)| = 1$, and (40) and (42) hold, we conclude that

$$\sup_{n\geq 1}\ \sup_{\omega,\eta\in[0,1]} |g_n(\omega, \eta)| < \infty.$$

But this means that the functions v_n are uniformly Hölder continuous—in particular, equicontinuous—and we may pass to the limit and obtain a nontrivial function $v \in C([0,1];\mathbb{C})$ such that $Av = \lambda v$, where $|\lambda| = \rho$. □

Proof of Theorem 6. First, we consider the operator A. By changing variables and using the periodicity of a, we easily see that for all f, $g \in C(\mathbb{T};\mathbb{C})$, we have

$$\int_0^1 (Af)(\omega)g(\omega)\,\mathrm{d}\omega = \int_0^1 f(\omega)2a(\omega)g(2\omega)\,\mathrm{d}\omega. \tag{43}$$

In particular, this implies that if we define numbers $\sigma_{p,m}$ by

$$\sigma_{p,m} \overset{\text{def}}{=} 2^m \int_0^1 \sin(\pi 2^m \omega)^{2p} \prod_{k=0}^{m-1} a(2^k\omega)\,\mathrm{d}\omega, \quad p \in \{0, M\}, \quad m \geq 1, \tag{44}$$

then it follows from m applications of (43) that

$$\sigma_{p,m} = \int_0^1 (A^m 1)(\omega)\sin(\pi\omega)^{2p}\,\mathrm{d}\omega, \quad p \in \{0, M\}, \quad m \geq 1. \tag{45}$$

Thus we have

$$\sigma_{p,m} \leq \|A^m\|, \quad p \in \{0, M\}, \quad m \geq 1, \tag{46}$$

where $\|\bullet\|$ denotes the operator norm.

From the moment assumption on γ, it follows that $\hat{\gamma} \in C^{2M+1}(\mathbb{T};\mathbb{R})$, and hence it is clear that if we want to prove that $a \in C^1(\mathbb{T};\mathbb{R})$, then the only points where there may be problems are the points $\frac{1}{2} + \mathbb{Z}$, and by periodicity, it suffices to consider the point $\frac{1}{2}$. By (26) and Taylor's formula, we have

$$\hat{\gamma}(\underline{\omega}) = \frac{\hat{\gamma}^{(2M)}(\frac{1}{2})}{(2M)!}(\underline{\omega} - \tfrac{1}{2})^{2M} + \int_{1/2}^{\underline{\omega}} \hat{\gamma}^{(2M+1)}(\sigma)\frac{(\underline{\omega} - \sigma)^{2M}}{(2M)!}\,\mathrm{d}\sigma.$$

Now a straightforward calculation, where we use the fact that $\hat{\gamma} \in C^{2M+1}(\mathbb{T};\mathbb{R})$, shows that $\lim_{\omega\to 1/2} a(\omega)$ and $\lim_{\omega\to 1/2} a'(\omega)$ both exist, and this gives the desired conclusion that $a \in C^1(\mathbb{T};\mathbb{R})$.

Thus we can apply Lemma 7, and we conclude that there is a nontrivial function $v \in C([0,1];\mathbb{C})$ and a number $\lambda \in \mathbb{C}$ such that $Av(\underline{\omega}) = \lambda v(\underline{\omega})$ and $|\lambda| = \rho$, the spectral radius of A. Because we may assume that $\sup_{\omega\in[0,1]}|v(\omega)| = 1$, we therefore see by (45) that

$$\begin{aligned}\rho^m \int_0^1 |v(\omega)|^2 \sin(\pi\omega)^{2p}\,\mathrm{d}\omega &= \left|\int_0^1 (A^m v)(\omega)\overline{v(\omega)}\sin(\pi\omega)^{2p}\,\mathrm{d}\omega\right| \\ &\leq \int_0^1 (A^m 1)(\omega)\sin(\pi\omega)^{2p}\,\mathrm{d}\omega = \sigma_{p,m}, \quad p \in \{0, M\}, \quad m \geq 1.\end{aligned}$$

Combining this result with (46) and the fact that $\lim_{m\to\infty}\|A^m\|^{1/m} = \rho$, we conclude that for $p = 0$ or M, we have

$$\sum_{m=1}^{\infty} 2^{-sm}\sigma_{p,m} < \infty \quad \text{iff} \quad s > \log_2(\rho). \tag{47}$$

Let

$$\mu \overset{\text{def}}{=} \inf_{\omega\in\mathbb{R}}\Big(a(\omega) + a\big(\omega + \tfrac{1}{2}\big)\Big).$$

It is clear that $(A^m 1)(\omega) \geq \mu^m$, $\omega \in [0,1]$, hence $\|A^m\| \geq \mu^m$, and it follows that the spectral radius of A is at least μ. Since $|\cos(\pi\omega)| \leq 1$ for all ω, we conclude from (11) and the definition of a that $\mu \geq 1$. By (27), we have $a(\frac{1}{2}) \neq 0$, and we see that

$a(0) = 1$ because $\hat{\gamma}(0) = 1$ by (14). Since $|\cos(\pi\omega)| < 1$ when $\omega \notin \mathbb{Z}$, it therefore follows from (11) and the definition of a that $\mu > 1$, and hence $\rho > 1$.

Let us define the function Q by

$$Q(\underline{\omega}) = \frac{(\pi\underline{\omega})^{2M}\hat{\Phi}(\underline{\omega})}{\sin(\pi\underline{\omega})^{2M}\hat{\Phi}(0)}.$$

Our assumption (13) implies that $\Phi \in L^1(\mathbb{R};\mathbb{R})$, and hence we get (30) from (2) and we have $\hat{\Phi}(\underline{\omega}) = \hat{\Phi}(2^{-k}\underline{\omega})\Pi_{j=1}^{k}\hat{\gamma}(2^{-j}\underline{\omega})$, where $1 \le k \le \infty$. Therefore, it follows from (26) that Q is continuous. The important point, however, is that since $\prod_{k=1}^{\infty}\cos(\pi 2^{-k}\underline{\omega}) = \sin(\pi\underline{\omega})/(\pi\underline{\omega})$, we get in addition that

$$Q(\underline{\omega}) = \prod_{k=1}^{\infty} a(2^{-k}\underline{\omega}). \tag{48}$$

Let $s = 2M - r$ and assume that $s > 0$. It follows from the definition of Q that

$$\text{(i) holds iff } \int_{\mathbb{R}} (1+\omega^2)^{-s/2} Q(\omega)\,\mathrm{d}\omega < \infty, \tag{49}$$

and

$$\text{(ii) holds iff } \int_{\mathbb{R}} (1+\omega^2)^{-s/2} Q(\omega)\sin(\pi\omega)^{2M}\,\mathrm{d}\omega < \infty. \tag{50}$$

Because $s > 0$, after integrating by parts we get that

$$\int_{\mathbb{R}} (1+\omega^2)^{-s/2} Q(\omega)\sin(\pi\omega)^{2p}\,\mathrm{d}\omega$$
$$= \int_0^{\infty} s\omega(1+\omega^2)^{-1-s/2}\int_{-\omega}^{\omega} Q(\eta)\sin(\pi\eta)^{2p}\,\mathrm{d}\eta\,\mathrm{d}\omega, \quad p \in \{0, M\},$$

where equality holds in the case where one of the integrals diverges too. There are positive constants C_1 and C_2 such that $C_1 \le 2^{sm}\int_{2^{m-1}J}^{2^m J} s\omega(1+\omega^2)^{-1-s/2}\,\mathrm{d}\omega \le C_2$, and hence we see that for $p \in \{0, M\}$,

$$\begin{aligned} &\int_{\mathbb{R}} (1+\omega^2)^{-s/2} Q(\omega)\sin(\pi\omega)^{2p}\,\mathrm{d}\omega < \infty \\ &\qquad\qquad \text{iff} \quad \sum_{m=1}^{\infty} 2^{-sm}\int_{-2^m J}^{2^m J} \sin(\pi\omega)^{2p} Q(\omega)\,\mathrm{d}\omega < \infty. \end{aligned} \tag{51}$$

Using (48), changing variables, and invoking the periodicity of a, we get

$$\begin{aligned} &\int_{-2^m J}^{2^m J} \sin(\pi\omega)^{2p} Q(\omega)\,\mathrm{d}\omega \\ &\qquad = \int_{-2^m J}^{2^m J} \sin(\pi\omega)^{2p}\prod_{k=1}^{m} a(2^{-k}\omega) Q(2^{-m}\omega)\,\mathrm{d}\omega \\ &\qquad = 2^m \int_{-J}^{J} \sin(\pi 2^m\omega)^{2p}\prod_{k=0}^{m-1} a(2^k\omega) Q(\omega)\,\mathrm{d}\omega \\ &\qquad = 2^m \int_0^1 \sin(\pi 2^m\omega)^{2p}\prod_{k=0}^{m-1} a(2^k\omega)\sum_{j=-J}^{J-1} Q(\omega+j)\,\mathrm{d}\omega. \end{aligned} \tag{52}$$

In the proof of Proposition 1, we showed that the series in (28) converges uniformly and therefore there exists an integer J such that $\sum_{j=-J}^{J-1} \hat{\Phi}(\omega + j) \geq \frac{1}{2}$ for all $\omega \in [0,1]$. Moreover, from (28) we get $\hat{\Phi}(0) = 1$ because by (26) and (30) we have $\hat{\Phi}(n) = 0$ when $n \in \mathbb{Z} \setminus \{0\}$. Hence $Q(\underline{\omega}) \geq \hat{\Phi}(\underline{\omega})$, and we have for some constant C_3

$$\frac{1}{2} \leq \sum_{j=-J}^{J-1} \Phi(\omega + j) \leq C_3, \quad \omega \in [0,1].$$

If we use this result in (52), then we see from (44) and (51) that for $p \in \{0, M\}$,

$$\int_{\mathbb{R}} (1+\omega^2)^{-s/2} Q(\omega) \sin(\pi\omega)^{2p} \, \mathrm{d}\omega < \infty \qquad \text{iff} \qquad \sum_{m=1}^{\infty} 2^{-sm} \sigma_{p,m} < \infty.$$

Thus the claimed equivalence follows from (47), (49), and (50). □

REFERENCES

[1] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary subdivision*, Mem. Amer. Math. Soc., 93 (1991), pp. 1–186.

[2] A. Cohen, *Ondelettes, analyses multirésolutions et filtres miroirs en quadrature*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 439–459.

[3] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[4] ———, *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[5] G. Deslauriers and S. Dubuc, *Symmetric iterative interpolation*, Constr. Approx., 5 (1989), pp. 49–68.

[6] S. Dubuc, *Interpolation through an iterative scheme*, J. Math. Anal. Appl., 114 (1986), pp. 185–204.

[7] T. Eirola, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[8] G. Gripenberg, *Unconditional bases of wavelets for Sobolev spaces*, SIAM J. Math. Anal., 24 (1993), pp. 1030–1042.

[9] L. Herve, *Régularité et conditions de bases de Riesz pour les fonctions d'échelle*, C. R. Acad. Sci. Paris Sér. I Math., 315, pp. 1029–1032.

[10] S. G. Mallat, *Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.

[11] Y. Meyer, *Ondelettes et Opérateurs* I, Hermann, Paris, 1990.

[12] W. Rudin, *Fourier Analysis on Groups*, J. Wiley, New York, 1990.

[13] M. J. Shensa, *The discrete wavelet transform: Wedding the á trous and Mallat algorithms*, IEEE Trans. Signal Process., 40 (1992), pp. 2464–2482.

[14] W. Sweldens and R. Piessens, *Asymptotic error expansions of wavelets approximations of smooth functions* II, Numer. Math., 68 (1994), pp. 377–401.

[15] L. F. Villemoes, *Energy moments in time and frequency for two-scale difference equation solutions and wavelets*, SIAM J. Math. Anal., 23 (1992), pp. 1519–1543.

[16] L.F. Villemoes, *Wavelet analysis of refinement equations*, SIAM J. Math. Anal., 25 (1994), pp. 1433–1460.

SIAM J. MATH. ANAL.
Vol. 27, No. 2, pp. 594–608, March 1996

015

WAVELETS FROM SQUARE-INTEGRABLE REPRESENTATIONS*

DAVID BERNIER† AND KEITH F. TAYLOR†

Abstract. The continuous wavelet decompositions that arise from square-integrable representations of certain Lie groups on $L^2(\mathbb{R}^n)$ are investigated. The groups are formed as the semidirect product of $\mathbb{R}^n$ with an n-dimensional subgroup H of $GL_n(\mathbb{R})$. There is a natural "translation and dilation" representation of such groups on $L^2(\mathbb{R}^n)$. The basic formulas of Duflo and Moore, which lead to the resolution of the identity via a square-integrable representation, are given an elementary proof for this special case. Several two-dimensional examples are described. A method for discrete decompositions via frames is given using the representations under study.

Key words. wavelet, square-integrable representation, resolution of the identity, frame

AMS subject classifications. 42C15, 22D10

1. Introduction.

This paper is concerned with higher-dimensional analogs of the theory of continuous and discrete affine wavelets in $L^2(\mathbb{R})$. We emphasize the role of a locally compact group acting on $\mathbb{R}^n$ in such a way that square-integrable representations of the associated semidirect product group arise. These square-integrable representations are the source of existence of (at least) the continuous wavelets.

An introduction to the theory of wavelets in $L^2(\mathbb{R})$ can be found in the book by Daubechies [6] or the survey article by Heil and Walnut [16], which includes an explanation of the role of a certain pair of square-integrable representations of the affine group on $L^2(\mathbb{R})$. The affine group can be viewed as the semidirect product of $\mathbb{R}^+$, the multiplicative group of positive real numbers, acting on $\mathbb{R}$ as dilations. There is a natural representation of this semidirect product on $L^2(\mathbb{R})$ by "translations and dilations" that is the direct sum of two square-integrable representations. The abstract orthogonality relations of Duflo and Moore [9] then say that there exist vectors in $L^2(\mathbb{R})$ which can serve as continuous wavelets. The details are described in §1 below and can also be found in [15] and [16].

In order to generalize this to higher dimensions, we consider closed subgroups H of $GL_n(\mathbb{R})$ which act on $\mathbb{R}^n$ in such a manner that the natural representation of $\mathbb{R}^n \rtimes H$ on $L^2(\mathbb{R}^n)$ contains square-integrable representations. Here $\mathbb{R}^n \rtimes H$ denotes the semidirect product of $\mathbb{R}^n$ with H. The exact definition is given in §2, where we go on to give a very elementary proof of the basic Duflo–Moore theorem for the square-integrable representations in question. Our proof also makes the so-called admissibility condition very explicit.

In §3, examples in two dimensions are considered. Two of the examples have already found use in applications; one leads to wavelets which are tensor products of one-dimensional wavelets and the other gives wavelets for $L^2(\mathbb{R}^2)$ which are moved by translation, dilation, and rotation [2]. In addition to these two examples, a whole new class of examples is discussed.

In §4, we consider methods for discretizing the reconstruction formulas to obtain frames in $L^2(\mathbb{R}^n)$. In Theorem 3, we describe discrete subsets of $\mathbb{R}^n$ and H and conditions on a $g \in L^2(\mathbb{R}^n)$ which lead to the translates and dilates of g providing a frame

*Received by the editors September 28, 1993; accepted for publication (in revised form) June 7, 1994.

†Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, SK S7N 0W0, Canada. This research was supported by a National Sciences and Engineering Research Council of Canada grant.

in $L^2(\mathbb{R}^n)$. The general conditions are then made specific for the two-dimensional examples in §5. The new methods of constructing frames from a group representation are obtained in this paper by considering more general groups that still have square-integrable representations in the strict sense. Generalizations have also been studied by DeBièvre in [7] and Ali, Antoine, and Gazeau in [1], where the motivation is to generalize the concept of a coherent state. The representation that underlies their concept of a frame is not necessarily square integrable.

1. Foundational harmonic analysis.

The reader is referred to Chapter III of [11] for the general theory of locally compact groups; we present only what is necessary to establish the notation. Let G be a locally compact group with left Haar integral denoted $\int_G \cdots dx$. Let $C_c(G)$ denote the function space consisting of continuous compactly supported complex-valued functions on G. For $f \in C_c(G)$, let $\|f\|_p = \left(\int_G |f(x)|^p dx\right)^{1/p}$ for $1 \leq p < \infty$. Let $L^p(G)$ denote the completion of the normed linear space $\left(C_c(G), \|\cdot\|_p\right)$ for $1 \leq p < \infty$. We are most interested in $L^1(G)$ and $L^2(G)$. As usual, we identify elements in $L^p(G)$ with measurable functions.

The left Haar integral on G satisfies, for any $f \in C_c(G)$ and $y \in G$,

$$\int_G f(x)dx = \int_G f(yx)dx. \tag{1.1}$$

In this study, G will not usually be unimodular and the modular function on G plays a significant role (see [11, §III.8] for information on modular functions). The modular function is a continuous homomorphism $\triangle_G : G \to \mathbb{R}^+$ such that, for any $f \in C_c(G)$ and $y \in G$,

$$\int_G f(x)dx = \triangle_G(y) \int_G f(xy)dx. \tag{1.2}$$

A unitary representation of G is a homomorphism π of G into $\mathcal{U}(\mathcal{H}_\pi)$, the group of unitary operators on a Hilbert space $\mathcal{H}_\pi$. We will always require that π is continuous when $\mathcal{U}(\mathcal{H}_\pi)$ is equipped with the weak operator topology. That is, for any $\xi, \eta \in \mathcal{H}_\pi$, the coefficient function $v_{\xi,\eta}(x) = \langle \eta, \pi(x)\xi \rangle$ is a continuous function of $x \in G$. A good source of fundamental results on unitary representations is Dixmier's book [8]. For the rest of this paper, representation will mean unitary representation.

A representation π is called irreducible if $\{0\}$ and $\mathcal{H}_\pi$ are the only closed subspaces of $\mathcal{H}_\pi$ which are invariant under $\pi(x)$ for each $x \in G$. A vector $\xi \in \mathcal{H}_\pi$ is called a cyclic vector for π if $\{\pi(x)\xi : x \in G\}$ is a total set in $\mathcal{H}_\pi$. That is, for $\eta \in \mathcal{H}_\pi, \eta \perp \pi(x)\xi$, for all $x \in G$, implies $\eta = 0$. Equivalently, ξ is a cyclic vector for π if and only if $v_{\xi,\eta} = 0$ implies $\eta = 0$ for $\eta \in \mathcal{H}_\pi$. It is easy to see that π is irreducible if and only if every nonzero vector in $\mathcal{H}_\pi$ is a cyclic vector for π. In many cases, this is the most convenient method for establishing irreducibility of a representation.

From the point of view of wavelet analysis, the key concept involving representations is square integrability. A representation π of G is called square integrable if (i) π is irreducible and (ii) there exist $\xi, \eta \in \mathcal{H}_\pi$, both nonzero, such that $v_{\xi,\eta} \in L^2(G)$. The most important theorem concerning square-integrable representations was established by Duflo and Moore [9] and is stated here for easy reference. A vector $\xi \in \mathcal{H}_\pi$ is called admissible if there exists a nonzero $\eta \in \mathcal{H}_\pi$ such that $v_{\xi,\eta} \in L^2(G)$.

THEOREM (Duflo and Moore, [9, Thm. 3]). *Let π be a square-integrable representation of a locally compact group G on a Hilbert space $\mathcal{H}_\pi$. Then there exists a unique operator K on $\mathcal{H}_\pi$, self-adjoint positive and satisfying the following:*

(i) $\pi(x)K\pi(x)^{-1} = \triangle_G(x)^{-1}K$ *for all $x \in G$.*

(ii) dom$K^{-\frac{1}{2}} = \{\xi \in \mathcal{H}_\pi : \xi$ *is admissible*$\}$.

(iii) *If* ξ *is admissible, then* $v_{\xi,\eta} \in L^2(G)$ *for all* $\eta \in \mathcal{H}_\pi$.

(iv) *If* ξ_1 *and* ξ_2 *are admissible and* $\eta_1, \eta_2 \in \mathcal{H}_\pi$, *then*

$$\langle v_{\xi_1,\eta_1}, v_{\xi_2,\eta_2}\rangle_{L^2(G)} = \langle \eta_1, \eta_2\rangle_{\mathcal{H}_\pi}\langle K^{-\frac{1}{2}}\xi_2, K^{-\frac{1}{2}}\xi_1\rangle_{\mathcal{H}_\pi}. \tag{1.3}$$

The proof in [9] used details of the Mackey analysis to prove this theorem. Alternate proofs were also provided by Carey [5] using reproducing kernel Hilbert-space methods and by Phillips [17] using quasi-Hilbert algebras.

In [15], it was recognized that this theorem allows one to represent arbitrary elements of $\mathcal{H}_\pi$ in terms of a fixed admissible element and its images under the representation. In particular, if ξ is admissible and normalized so that $\|K^{-\frac{1}{2}}\xi\| = 1$, then we can define a linear mapping $V_\xi : \mathcal{H}_\pi \to L^2(G)$ by $V_\xi\eta = v_{\xi,\eta}$, for $\eta \in \mathcal{H}_\pi$. The notation V_ξ follows [10], where it is called the voice transform defined by ξ. Using (1.3), we get, for $\eta_1, \eta_2 \in \mathcal{H}_\pi$, since $\|K^{-\frac{1}{2}}\xi\| = 1$,

$$\langle V_\xi\eta_1, V_\xi\eta_2\rangle_{L^2(G)} = \langle \eta_1, \eta_2\rangle_{\mathcal{H}_\pi}. \tag{1.4}$$

Let $\mathcal{K}_\xi$ denote the range of V_ξ, a closed subspace of $L^2(G)$. Then $V_\xi : \mathcal{H}_\pi \to \mathcal{K}_\xi$ is a unitary map between Hilbert spaces. Let l_G denote the left regular representation of G on $L^2(G)$ defined by $[l_G(x)f](y) = f(x^{-1}y)$ for all $y \in G, f \in L^2(G)$, and $x \in G$. An easy calculation shows that $l_G(x)(V_\xi\eta) = V_\xi(\pi(x)\eta)$ for all $\eta \in \mathcal{H}_\pi$. Thus $\mathcal{K}_\xi$ is a l_G-invariant subspace of $L^2(G)$ and V_ξ defines a unitary equivalence of π with the subrepresentation of l_G formed by restricting $l_G(x)$ to $\mathcal{K}_\xi$ for each $x \in G$.

The reconstruction of $\eta \in \mathcal{H}_\pi$ from $V_\xi\eta$ is implicit in (1.4). If Hilbert-space-valued integrals are considered in the weak sense and $\eta \in \mathcal{H}_\pi$, then (1.4) implies

$$\int_G V_\xi\eta(x)\pi(x)\xi dx = \eta$$

or

$$\int_G \langle \eta, \pi(x)\xi\rangle_{\mathcal{H}_\pi}\pi(x)\xi\, dx = \eta. \tag{1.5}$$

This can be further refined to give a resolution of the identity operator I on $\mathcal{H}_\pi$. If $\mathcal{H}$ is a Hilbert space and $\nu \in \mathcal{H}, \nu \neq 0$, let $\nu \otimes \nu$ denote the rank-1 operator given by $(\nu\otimes\nu)(\mu) = \langle \mu, \nu\rangle_{\mathcal{H}}\nu$ for all $\mu \in \mathcal{H}$. With operator-valued integrals interpreted in the weak-operator topology sense, (1.5) becomes

$$\int_G (\pi(x)\xi) \otimes (\pi(x)\xi)dx = I. \tag{1.6}$$

This completes our summary of the known theory of square-integrable representations. Formula (1.5) is useful when the Hilbert space $\mathcal{H}_\pi$ is realized as a meaningful function space. One of the important questions is: which ξ are admissible, or, equivalently, what is the domain of the unbounded (if G is not unimodular) operator $K^{-\frac{1}{2}}$? In the next section, we will study a class of groups which have square-integrable representations for which the operator K has an easily understood form and, consequently, the admissibility condition is easily stated.

2. A class of semidirect products. Let $GL_n(\mathbb{R})$ denote the group of invertible $n \times n$ real matrices with the usual topology. Let H be a closed subgroup of

$GL_n(\mathbb{R})$. We will consider the elements $\underline{x}$ of $\mathbb{R}^n$ as column vectors and the elements h of H as $n \times n$ matrices. Then the matrix product $(h, \underline{x}) \rightarrow h\underline{x}$ gives a natural action of H on $\mathbb{R}^n$. For $h \in H$, let $\delta(h) = |\det(h)|$. For any integrable function g on $\mathbb{R}^n$,

$$(2.1) \qquad \int_{\mathbb{R}^n} g(\underline{x})d\underline{x} = \delta(h) \int_{\mathbb{R}^n} g(h\underline{x})d\underline{x}.$$

We will use $\widehat{\mathbb{R}^n}$ for n-dimensional Euclidean space with the elements written as row vectors. For $\underline{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_n) \in \widehat{\mathbb{R}^n}$ and

$$\underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n,$$

we have $\underline{\gamma}\underline{x} = \sum_{j=1}^{n} \gamma_j x_j$. This facilitates calculations with the Fourier transform. For $f \in L^1(\mathbb{R}^n)$, its Fourier transform, $\hat{f} : \widehat{\mathbb{R}^n} \rightarrow \mathbb{C}$, is given, for $\underline{\gamma} \in \widehat{\mathbb{R}^n}$, by

$$(2.2) \qquad \hat{f}(\underline{\gamma}) = \int_{\mathbb{R}^n} f(\underline{x}) e^{2\pi i \underline{\gamma}\underline{x}}\, d\underline{x}.$$

If $g \in L^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$, let $\mathcal{P}g = \hat{g}$. The Plancherel theorem says that $\mathcal{P}$ extends to a unitary map of $L^2(\mathbb{R}^n)$ onto $L^2(\widehat{\mathbb{R}^n})$.

The dual action of H on $\widehat{\mathbb{R}^n}$ is also very important to us and is given by the natural product $(h, \underline{\gamma}) \rightarrow \underline{\gamma}h$ for $h \in H$, $\underline{\gamma} \in \widehat{\mathbb{R}^n}$. Then, for any integrable ξ on $\widehat{\mathbb{R}^n}$ and $h \in H$,

$$(2.3) \qquad \int_{\widehat{\mathbb{R}^n}} \xi(\underline{\gamma})d\underline{\gamma} = \delta(h) \int_{\widehat{\mathbb{R}^n}} (\underline{\gamma}h)d\underline{\gamma}.$$

The group H and its action on $\mathbb{R}^n$ can be used to form a new locally compact group $G = \mathbb{R}^n \rtimes H$, the semidirect product of $\mathbb{R}^n$ and H (see [11, §§III.4.7 and III.9.4]). The elements of G are all the ordered pairs $(\underline{x}, h)$ with $\underline{x} \in \mathbb{R}^n$ and $h \in H$. The product rule for G is, for $(\underline{x}, h), (\underline{y}, k) \in G$,

$$(2.4) \qquad (\underline{x}, h)(\underline{y}, k) = (\underline{x} + h\underline{y}, hk).$$

We will make heavy use of left-invariant (Haar) integration on H and G. Let $\int_H \cdots dh$ denote a fixed left Haar integral on H. Then left Haar integration on G is given, for any $f \in C_c(G)$, by

$$(2.5) \qquad \int_G f(\underline{x}, h)d(\underline{x}, h) = \int_H \int_{\mathbb{R}^n} f(\underline{x}, h)\delta(h)^{-1} d\underline{x}\, dh.$$

It is easily checked that (2.5) defines a left-invariant integral on G. However, this integral will not usually be right invariant. If $\triangle_H$ denotes the modular function on H, then a short calculation shows that the modular function on G is given by

$$(2.6) \qquad \triangle_G(\underline{x}, h) = \triangle_H(h)/\delta(h), \quad \text{for all} \quad (\underline{x}, h) \in G.$$

There is a very natural and important representation of G on $L^2(\mathbb{R}^n)$ that combines translation by vectors in $\mathbb{R}^n$ with the action of H. For $(\underline{x}, h) \in G$ and $g \in L^2(\mathbb{R}^n)$, define

$$(2.7) \qquad \rho(\underline{x}, h)g(\underline{y}) = \delta(h)^{-\frac{1}{2}} g\big(h^{-1}(\underline{y} - \underline{x})\big)$$

for all $\underline{y} \in \mathbb{R}^n$. The basic representation of the affine group $\mathbb{R} \rtimes \mathbb{R}^+$ on $L^2(\mathbb{R})$ that plays a fundamental role in affine wavelets in $L^2(\mathbb{R})$ is essentially a special case of (2.7) (see [16, §3.3.2]). It is routine to show that ρ is a (continuous unitary) representation of G on $L^2(\mathbb{R}^n)$. It will be very useful to know the equivalent representation π on $L^2(\widehat{\mathbb{R}^n})$ obtained from using the Plancherel transform $\mathcal{P}$. That is,

$$\pi(\underline{x}, h) = \mathcal{P}\rho(\underline{x}, h)\mathcal{P}^{-1} \quad \text{for all} \quad (\underline{x}, h) \in G. \tag{2.8}$$

Proposition 1. *For $(\underline{x}, h) \in G$ and $\xi \in L^2(\widehat{\mathbb{R}^n})$,*

$$\pi(\underline{x}, h)\xi(\underline{\gamma}) = \delta(h)^{\frac{1}{2}} e^{2\pi i \underline{\gamma}\underline{x}} \xi(\underline{\gamma} h) \quad \textit{for all} \quad \underline{\gamma} \in \widehat{\mathbb{R}^n}. \tag{2.9}$$

Proof. It suffices to check (2.9) for $\xi = \hat{g}$ with $g \in L^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$. For such a $\xi = \hat{g}$ and $\underline{\gamma} \in \widehat{\mathbb{R}^n}$,

$$\begin{aligned}
\pi(\underline{x}, h)\xi(\underline{\gamma}) &= \big(\mathcal{P}\rho(\underline{x}, h)\mathcal{P}^{-1}\big)(\mathcal{P}g)(\underline{\gamma}) \\
&= \int_{\mathbb{R}^n} \rho(\underline{x}, h)g(\underline{y})e^{2\pi i \underline{\gamma}\,\underline{y}}\, d\underline{y} \\
&= \int_{\mathbb{R}^n} \delta(h)^{-\frac{1}{2}} g\big(h^{-1}(\underline{y} - \underline{x})\big)e^{2\pi i \underline{\gamma}\,\underline{y}}\, d\underline{y} \\
&= e^{2\pi i \underline{\gamma}\underline{x}}\delta(h)^{\frac{1}{2}} \int_{\mathbb{R}^n} g(\underline{y})e^{2\pi i \underline{\gamma} h \underline{y}}\, d\underline{y} \\
&= \delta(h)^{\frac{1}{2}} e^{2\pi i \underline{\gamma}\underline{x}} \xi(\underline{\gamma} h). \qquad \square
\end{aligned}$$

We denote the unitary equivalence of two representations with a $\sim$. Thus, $\pi \sim \rho$.

If U is an open subset of $\widehat{\mathbb{R}^n}$, let $L^2(U)$ denote the closed subspace of $L^2(\widehat{\mathbb{R}^n})$ consisting of elements supported on U and let $\mathcal{H}^2_U = \mathcal{P}^{-1}\big(L^2(U)\big)$. Then $\mathcal{H}^2_U$ can be thought of as a generalized Hardy space. If U is an H-invariant open subset of $\widehat{\mathbb{R}^n}$ (that is, $\underline{\gamma} \in U$ and $h \in H$ imply $\underline{\gamma} h \in U$), then $L^2(U)$ is seen to be a π-invariant subspace of $L^2(\widehat{\mathbb{R}^n})$ by (2.9) and $\mathcal{H}^2_U$ is ρ-invariant. In this case, let π_U and ρ_U denote the subrepresentations of π and ρ, respectively, formed by restricting to $L^2(U)$ and $\mathcal{H}^2_U$. Note that $\rho_U \sim \pi_U$. The open H-invariant subsets of $\widehat{\mathbb{R}^n}$ which are single H-orbits are of special interest.

An H-orbit in $\widehat{\mathbb{R}^n}$ is a set of the form $\underline{\gamma} H = \{\underline{\gamma} h : h \in H\}$ for $\underline{\gamma} \in \widehat{\mathbb{R}^n}$. Clearly, $\underline{\gamma}_1 H = \underline{\gamma}_2 H$ if and only if there exists $k \in H$ such that $\underline{\gamma}_1 = \underline{\gamma}_2 k$ and H-orbits are H-invariant sets. An H-orbit $\underline{\gamma} H$ is called free if $\underline{\gamma} h = \underline{\gamma}$ implies $h = e$, the identity element of H. It doesn't matter which element of the H-orbit is used to check for freeness. If $\underline{\gamma} \in \widehat{\mathbb{R}^n}$ and $\underline{\gamma} H$ is a free orbit, then $h \rightarrow \underline{\gamma} h$ is a continuous bijection of H with $\underline{\gamma} H$. In general, this bijection need not be a homeomorphism. However, if $\underline{\gamma} H$ is not only free but open in $\widehat{\mathbb{R}^n}$, then $h \rightarrow \underline{\gamma} h$ is a homeomorphism of H onto $\underline{\gamma} H$ by Theorem 1 of [12]. Most of the rest of this paper is based on situations where we have open free H-orbits in $\widehat{\mathbb{R}^n}$. The next theorem shows why.

Theorem 1. *Let H be a closed subgroup of $GL_n(\mathbb{R})$ and let $G = \mathbb{R}^n \rtimes H$. Let U be an open free H-orbit in $\widehat{\mathbb{R}^n}$. Then ρ_U is a square-integrable representation of G on $\mathcal{H}^2_U$.*

Remarks. Several examples of H and U satisfying the hypothesis of Theorem 1 will be given in §3. The proof given here for Theorem 1 is adapted from the proof for the affine group, $\mathbb{R} \rtimes \mathbb{R}^+$, given in [16, §§3.3.5 and 3.3.6]. Besides being elementary and giving a simple admissibility condition, the proof leads the way to a very simple proof of (1.3) for the representation ρ_U, as we will see in Theorem 2. Before proving Theorem 1, we study the relationship between the Lebesgue measure on the open free H-orbit U and the measure on U obtained by transferring the left Haar measure of H to U. Let λ^n denote the Lebesgue measure on $\widehat{\mathbb{R}^n}$, which, by restriction, we view as a regular Borel measure on U. For $\underline{\gamma} \in U$, define $m_{\underline{\gamma}}$, a measure on U, by transferring the left Haar measure of H to U via the homeomorphism $h \to \underline{\gamma} \cdot h$. That is, for a Borel subset B of U, let $\tilde{B}_{\underline{\gamma}} = \{h \in H : \underline{\gamma} h \in B\}$ and $m_{\underline{\gamma}}(B) = \int_H \chi_{\tilde{B}_{\underline{\gamma}}}(h)dh$.

LEMMA 1. *$m_{\underline{\gamma}}$ is independent of $\underline{\gamma} \in U$.*

Proof. If $\underline{\gamma}_1, \underline{\gamma}_2 \in U$, then there exists a $k \in H$ such that $\underline{\gamma}_1 = \underline{\gamma}_2 k$. Now, for $h \in H$, and any Borel set $B \subseteq U$, $h \in \tilde{B}_{\underline{\gamma}_1}$ iff $\underline{\gamma}_1 h \in B$ iff $\underline{\gamma}_2 kh \in B$ iff $h \in k^{-1}\tilde{B}_{\underline{\gamma}_2}$. By left invariance of the Haar integral on H, $m_{\underline{\gamma}_1}(B) = m_{\underline{\gamma}_2}(B)$ for any Borel $B \subseteq U$. □

Let $m = m_{\underline{\gamma}}$, for any $\underline{\gamma} \in U$.

LEMMA 2. *The measures λ^n and m are mutually absolutely continuous on U.*

Proof. Now, it is convenient to move λ^n and m to H. Fix a $\underline{\gamma} \in U$ and define, for a Borel subset A of H,

$$\lambda_H^n(A) = \lambda^n(\underline{\gamma} \cdot A) \quad \text{and} \quad m_H(A) = m(\underline{\gamma} \cdot A),$$

where $\underline{\gamma} \cdot A = \{\underline{\gamma} h : h \in A\}$. Of course, m_H is just the left Haar measure on H and, for $k \in H$,

$$\lambda_H^n(Ak) = \lambda^n(\underline{\gamma} \cdot Ak) = \lambda^n\big((\underline{\gamma} \cdot A)k\big) = \delta(k)\lambda^n(\underline{\gamma} \cdot A) = \delta(k)\lambda_H^n(A).$$

Thus λ_H^n is a (right) quasi-invariant measure on H. Of course, m_H is also quasi-invariant under right translations; so, by uniqueness of quasi-invariant measures on a coset space (in this case H, itself) (see [11, §III.4.9]), this means that λ_H^n is mutually absolutely continuous with m_H. Transferring back to U via the homeomorphism completes the proof. □

Let Ψ denote the Radon–Nikodym derivative of m with respect to λ^n. The properties of Ψ are collected in the following proposition.

PROPOSITION 2. *Let H be a closed subgroup of $GL_n(\mathbb{R})$ and let U be an open free H-orbit in $\widehat{\mathbb{R}^n}$. Then there exists a Borel measurable function Ψ on U with the following properties:*

(i) $0 < \Psi(\underline{\gamma}) < \infty$, *for all* $\underline{\gamma} \in U$.

(ii) *If η is either a continuous function or a nonnegative measurable function on U, then, for any $\underline{\gamma}_0 \in U$,*

$$\int_H \eta(\underline{\gamma}_0 h)dh = \int_U \eta(\underline{\gamma})\Psi(\underline{\gamma})d\underline{\gamma}. \tag{2.10}$$

(iii) *For any $k \in H$,*

$$\Psi(\underline{\gamma} k) = \frac{\Delta_H(k)}{\delta(k)}\Psi(\underline{\gamma}) \quad \textit{for a.e.} \quad \underline{\gamma} \in U. \tag{2.11}$$

Proof. Properties (i) and (ii) are easily checked from the properties of Radon–Nikodym derivatives. To verify (iii), let $\underline{\gamma}_0 \in U$ be fixed and let $\eta \in C_c(U)$. Then, by

(2.10), (2.3), and (1.2),

$$\begin{aligned}\int_U \eta(\underline{\gamma})\Psi(\underline{\gamma}k)d\underline{\gamma} &= \delta(k)^{-1}\int_U \eta(\underline{\gamma}k^{-1})\Psi(\underline{\gamma})d\underline{\gamma}\\ &= \delta(k)^{-1}\int_H \eta(\underline{\gamma}_0 hk^{-1})dh\\ &= \triangle_H(k)\delta(k)^{-1}\int_H \eta(\underline{\gamma}_0 h)dh\\ &= \triangle_H(k)\delta(k)^{-1}\int_U \eta(\underline{\gamma})\Psi(\gamma)d\gamma.\end{aligned}$$

Thus, $\Psi(\underline{\gamma}k) = \triangle_H(k)\delta(k)^{-1}\Psi(\underline{\gamma})$, for a.e. $\underline{\gamma} \in U$. □

We are now ready to prove Theorem 1.

Proof of Theorem 1. Let $f, g \in \mathcal{H}^2_U$. Then, using $\hat{f} = \mathcal{P}f$ and $\hat{g} = \mathcal{P}g$ for simplicity,

$$\begin{align*}&\int_G \left|\langle f, \rho_U(\underline{x}, h)g\rangle_{L^2(\mathbb{R}^n)}\right|^2 d(\underline{x}, h)\\ &\quad= \int_G \left|\langle \hat{f}, \pi_U(\underline{x}, h)\hat{g}\rangle_{L^2(\widehat{\mathbb{R}^n})}\right|^2 d(\underline{x}, h)\\ &\quad= \int_G \left|\int_{\widehat{\mathbb{R}^n}} \hat{f}(\underline{\gamma})\delta(h)^{\frac{1}{2}}e^{-2\pi i\underline{\gamma}\underline{x}}\,\overline{\hat{g}}(\underline{\gamma}h)d\underline{\gamma}\right|^2 d(\underline{x}, h). \tag{2.12}\end{align*}$$

For each $h \in H$, let $\phi_h(\underline{\gamma}) = \hat{f}(\underline{\gamma})\overline{\hat{g}}(\underline{\gamma}h)$ for all $\underline{\gamma} \in \widehat{\mathbb{R}^n}$. Then $\phi_h \in L^1(\widehat{\mathbb{R}^n})$ for each $h \in H$. Let $\phi_h^\vee$ denote its "inverse" Fourier transform on $\mathbb{R}^n$. Then (2.12) becomes

$$\begin{align*}&\int_G \left|\langle f, \rho_U(\underline{x}, h)g\rangle_{L^2(\mathbb{R}^n)}\right|^2 d(\underline{x}, h)\\ &\quad= \int_G \delta(h)\left|\phi_h^\vee(\underline{x})\right|^2 d(\underline{x}, h)\\ &\quad= \int_H\int_{\mathbb{R}^n} \left|\phi_h^\vee(\underline{x})\right|^2 d\underline{x}\,dh \tag{by 2.5}\\ &\quad= \int_H\int_{\widehat{\mathbb{R}^n}} \left|\phi_h(\underline{\gamma})\right|^2 d\underline{\gamma}dh\\ &\quad= \int_H\int_{\widehat{\mathbb{R}^n}} \left|\hat{f}(\gamma)\right|^2\left|\hat{g}(\underline{\gamma}h)\right|^2 d\underline{\gamma}\,dh\\ &\quad= \int_U \left|\hat{f}(\gamma)\right|^2\left(\int_H \left|\hat{g}(\underline{\gamma}h)\right|^2 dh\right) d\underline{\gamma}\\ &\quad= \int_U \left|\hat{f}(\underline{\gamma})\right|^2\left(\int_U \left|\hat{g}(\underline{\nu})\right|^2 \Psi(\underline{\nu})d\underline{\nu}\right) d\underline{\gamma} \tag{by 2.10}\\ &\quad= \|\hat{f}\|^2_{L^2(U)}\,\|\hat{g}\Psi^{\frac{1}{2}}\|_2^2\\ &\quad= \|f\|^2_{\mathcal{H}^2_U}\|\hat{g}\Psi^{\frac{1}{2}}\|_2^2. \tag{2.13}\end{align*}$$

Note that $\|\hat{g}\Psi^{\frac{1}{2}}\|_2^2$ may be infinite for $g \in \mathcal{H}^2_U$, but $\|\hat{g}\Psi^{\frac{1}{2}}\|_2^2 = 0$ if and only if $g = 0$.

If g is any nonzero element of $\mathcal{H}^2_U$ and $f \in \mathcal{H}^2_U$ satisfies $f \perp \rho_U(\underline{x}, h)g$ for all $(\underline{x}, h) \in G$, then (2.13) implies that $\|f\|^2_{\mathcal{H}^2_U}\,\|\hat{g}\psi^{\frac{1}{2}}\|_2^2 = 0$. Since $g \neq 0$, this implies $f = 0$. Therefore, any nonzero vector in $\mathcal{H}^2_U$ is a cyclic vector for ρ_U. Hence, ρ_U is an irreducible representation of G.

On the other hand, if $\xi \in C_c(U)$, $\xi \neq 0$, and $g \in \mathcal{H}^2_U$ is such that $\hat{g} = \xi$, then $\|\hat{g}\Psi^{\frac{1}{2}}\|_2^2 < \infty$. So (2.13) implies that, for any $f \in \mathcal{H}^2_U, v_{g,f} \in L^2(G)$. Therefore, ρ_U is a square-integrable representaiton of G. □

The following two corollaries are clear from the proof of Theorem 1.

COROLLARY 1. *For $g \in \mathcal{H}^2_U, g$ is admissible if and only if $\hat{g}\Psi^{\frac{1}{2}} \in L^2(U)$.*

COROLLARY 2. *If $f, g \in \mathcal{H}^2_U$ and $\phi_h(\underline{\gamma}) = \hat{f}(\underline{\gamma})\overline{\hat{g}}(\underline{\gamma}h)$ for all $\underline{\gamma} \in \widehat{\mathbb{R}^n}$ and $h \in H$, then g admissible implies $\phi_h \in L^2(\widehat{\mathbb{R}^n})$ for a.e. $h \in H$.*

We are now in a position to give an elementary proof the theorem of Duflo and Moore described in §1 for the square-integrable representation ρ_U.

THEOREM 2. *Let H be a closed subgroup of $GL_n(\mathbb{R})$ and let $G = \mathbb{R}^n \rtimes H$. Let U be an open free H-orbit in $\widehat{\mathbb{R}^n}$. Then there exists a self-adjoint positive operator K on $\mathcal{H}^2_U$ satisfying the following:*

(i) $\rho_U(\underline{x}, h)K\ \rho_U(\underline{x}, h)^{-1} = \triangle_G(\underline{x}, h)^{-1}K$ *for all* $(\underline{x}, h) \in G$.

(ii) $\operatorname{dom} K^{-\frac{1}{2}} = \{g \in \mathcal{H}^2_U : g$ *is admissible*$\}$.

(iii) *If g is an admissible element of $\mathcal{H}^2_U$, then $v_{g,f} \in L^2(G)$ for all $f \in \mathcal{H}^2_U$.*

(iv) *If g_1 and g_2 are admissible elements of $\mathcal{H}^2_U$ and $f_1, f_2 \in \mathcal{H}^2_U$, then*

$$\langle v_{g_1,f_1}, v_{g_2,f_2}\rangle_{L^2(G)} = \langle f_1, f_2\rangle_{\mathcal{H}^2_U}\ \langle K^{-\frac{1}{2}}g_2, K^{-\frac{1}{2}}g_1\rangle_{\mathcal{H}^2_U}.$$

Proof. Recall that Ψ denotes the Radon-Nikodym derivative of the left Haar measure on H, transfered to U, with respect to the Lebesgue measure of $\widehat{\mathbb{R}^n}$ restricted to U. Thus Ψ^{-1} is the Radon–Nikodym derivative of these measures in the other order. Let Ψ^{-1} also denote the operator on $L^2(U)$ defined by pointwise multiplication. Define K to be $\mathcal{P}^{-1}\Psi^{-1}\mathcal{P}$. Thus,

$$\operatorname{dom} K = \{g \in \mathcal{H}^2_U : \hat{g}/\Psi \in L^2(U)\}$$

and, for $g \in \operatorname{dom} K$, $\widehat{Kg} = \hat{g}/\Psi$.

For $\eta \in \operatorname{dom}\Psi^{-1}$ and $(\underline{x}, h) \in G$, using (2.9), we get

$$\pi_U(\underline{x}, h)\left[\Psi^{-1}\pi_U(\underline{x}, h)^{-1}\eta\right](\underline{\gamma}) = \Psi^{-1}(\underline{\gamma}h)\eta(\underline{\gamma}) \tag{2.14}$$

for all $\underline{\gamma} \in U$. By 2.11, $\Psi^{-1}(\underline{\gamma}h) = \frac{\delta(h)}{\triangle_H(h)}\Psi^{-1}(\underline{\gamma})$. Recall from (2.6) that $\triangle_G = \triangle_H/\delta$, and the definition of K establishes (i).

Now, for $g \in \mathcal{H}^2_U, g \in \operatorname{dom} K^{-\frac{1}{2}}$ iff $\hat{g}\Psi^{\frac{1}{2}} \in L^2(U)$ iff g is admissible by Corollary 1. Thus (ii) holds and (iii) follows immediately from the calculation (2.13). It remains to prove (iv).

Let g_1 and g_2 be admissible elements of $\mathcal{H}^2_U$ and $f_1, f_2 \in \mathcal{H}^2_U$. Let $\phi^j_h(\underline{\gamma}) = \hat{f}_j(\underline{\gamma})\overline{\hat{g}}_j(\underline{\gamma}h)$ for $\underline{\gamma} \in \widehat{\mathbb{R}^n}, h \in H$, and $j = 1, 2$. By Corollary 2, $\phi^1_h, \phi^2_h \in L^1(\widehat{\mathbb{R}^n}) \cap L^2(\widehat{\mathbb{R}^n})$ for almost all $h \in H$. Again let $\phi^{j\vee}_h = \mathcal{P}^{-1}\phi^j_h$, for $j = 1, 2$. The computation is similar to (2.13), so we leave some steps to the reader.

$$\begin{aligned}
&\langle v_{g_1,f_1}, v_{g_2,f_2}\rangle_{L^2(G)} = \int_G \langle \hat{f}_1, \pi_U(\underline{x}, h)\hat{g}_1\rangle_{L^2(U)}\overline{\langle \hat{f}_2, \pi_U(\underline{x}, h)\hat{g}_2\rangle}_{L^2(U)} d(\underline{x}, h)\\
&= \int_G \int_{\widehat{\mathbb{R}^n}} \hat{f}_1(\underline{\mu})\delta(h)^{\frac{1}{2}} e^{-2\pi i \underline{\mu}\underline{x}}\,\overline{\hat{g}}_1(\underline{\mu}h)d\underline{\mu} \int_{\widehat{\mathbb{R}^n}} \overline{\hat{f}}_2(\underline{\nu})\delta(h)^{\frac{1}{2}} e^{2\pi i \underline{\nu}\,\underline{x}}\hat{g}_2(\underline{\nu}h)d\underline{\nu}\, d(\underline{x}, h)\\
&= \int_H \int_{\mathbb{R}^n} \phi^{1\vee}_h(\underline{x})\overline{\phi^{2\vee}_h}(\underline{x})d\underline{x}dh
\end{aligned}$$

$$\begin{aligned}
&= \int_H \int_{\widehat{\mathbb{R}^n}} \phi_h^1(\underline{\gamma})\overline{\phi_h^2(\underline{\gamma})}d\underline{\gamma}dh \\
&= \int_H \int_U \hat{f}_1(\underline{\gamma})\overline{\hat{f}}_2(\underline{\gamma})\hat{g}_2(\underline{\gamma}h)\overline{\hat{g}}_1(\underline{\gamma}h)d\underline{\gamma}dh \\
&= \int_U \hat{f}_1(\underline{\gamma})\overline{\hat{f}}_2(\underline{\gamma}) \int_U \hat{g}_2(\underline{\mu})\overline{\hat{g}}_1(\underline{\mu})\Psi(\underline{\mu})\, d\underline{\mu}d\underline{\gamma} \qquad \text{(by 2.10)} \\
&= \langle \hat{f}_1, \hat{f}_2\rangle_{L^2(U)} \langle \hat{g}_2\Psi^{\frac{1}{2}}, \hat{g}_1\Psi^{\frac{1}{2}}\rangle_{L^2(U)} \\
&= \langle f_1, f_2\rangle_{\mathcal{H}_U^2} \langle K^{-\frac{1}{2}}g_2, K^{-\frac{1}{2}}g_1\rangle_{\mathcal{H}_U^2}. \qquad \square
\end{aligned}$$

Remarks. (1) The only part of the Duflo–Moore theorem that has not been established for ρ_U, by these elementary methods is the uniqueness of K.

(2) The forms that the reproducing formulas (1.5) and (1.6) take in the present setting are now obvious.

Our final remarks in this section concern the assumption that there exists an open free H-orbit in $\widehat{\mathbb{R}^n}$. This forces the dimension of H, as a Lie group, to be n. There are several questions which arise concerning such groups H and their action on $\mathbb{R}^n$(and $\widehat{\mathbb{R}^n}$).

(3) If H is a closed n-dimensional subgroup of $GL_n(\mathbb{R})$, must there exist an open H-orbit in $\widehat{\mathbb{R}^n}$? The answer is no; an example for $n = 2$ is easily found. Let

$$H = \left\{ \begin{pmatrix} a & 0 \\ b & 1 \end{pmatrix} : a, b \in \mathbb{R}, a > 0 \right\}.$$

Then, for $(\gamma_1, \gamma_2) \in \widehat{\mathbb{R}^2}$, $(\gamma_1, \gamma_2)\left(\begin{smallmatrix} a & 0 \\ b & 1 \end{smallmatrix}\right) = (a\gamma_1 + b\gamma_2, \gamma_2)$, from which it is easy to see that there are no open H-orbits in $\widehat{\mathbb{R}^2}$. Nevertheless, the answer seems to be yes for "most" H. It would be interesting to make this precise.

(4) If H is a closed n-dimensional subgroup of $GL_n(\mathbb{R})$ and if there exists at least one open free H-orbit in $\widehat{\mathbb{R}^n}$, then a multivariate calculus argument shows that the union of the open free H-orbits is dense in $\widehat{\mathbb{R}^n}$. Thus, if there is one open free H-orbit, then $L^2(\mathbb{R}^n)$ is a direct sum of ρ-invariant subspaces, the $\mathcal{H}_U^2$'s, on each of which ρ_U is square integrable. In such cases, it turns out that the regular representation of G is then a direct sum of irreducible subrepresentations, see [3] where several examples are given. Groups with this property were also studied in [4] and [18].

(5) In any dimension n, can one classify, up to inner equivalence in $GL_n(\mathbb{R})$, those closed n-dimensional subgroups H for which there exists an open orbit?

3. Two-dimensional examples. In this short section, we present some examples of two-dimensional subgroups of $GL_2(\mathbb{R})$, describe open orbits in $\widehat{\mathbb{R}^2}$ for these groups, and give admissibility conditions for vectors in the associated Hilbert spaces. We will return to these examples in §5, where we discuss frames arising from these group actions.

Example 1. Let A denote the diagonal subgroup of $GL_2(\mathbb{R})$. That is,

$$A = \left\{ \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix} : a_1, a_2 \in \mathbb{R}\backslash\{0\} \right\}.$$

Let $G = \mathbb{R}^2 \rtimes A$. Clearly, G factors as a direct product of two copies of the (disconnected) affine group of one dimension. The map $\left(\begin{smallmatrix} a_1 & 0 \\ 0 & a_2 \end{smallmatrix}\right) \to (1,1)\left(\begin{smallmatrix} a_1 & 0 \\ 0 & a_2 \end{smallmatrix}\right) = (a_1, a_2)$ is a homeomorphism of A onto $U = \left\{(\gamma_1, \gamma_2) \in \widehat{\mathbb{R}^2} : \gamma_1\gamma_2 \neq 0\right\}$. Since U is dense in

$\widehat{\mathbb{R}^2}, \mathcal{H}_U^2 = L^2(\mathbb{R}^2)$ and, by Theorem 1, ρ is a square-integrable representation of G on $L^2(\widehat{\mathbb{R}^2})$, where ρ is given by (2.7), which in this case becomes

$$\rho\left[\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}\right] g\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \frac{1}{\sqrt{|a_1 a_2|}}\, g\begin{pmatrix} (y_1 - x_1)/a_1 \\ (y_2 - x_2)/a_2 \end{pmatrix}$$

for all $\left(\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right) \in \mathbb{R}^2$, $g \in L^2(\mathbb{R}^2)$, and $[\left(\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right), \left(\begin{smallmatrix} a_1 & 0 \\ 0 & a_2 \end{smallmatrix}\right)] \in G$.

The Radon–Nikodym derivative of Proposition 2 is given by

$$\Psi(\gamma_1, \gamma_2) = \frac{1}{|\gamma_1 \gamma_2|} \quad \text{for} \quad (\gamma_1, \gamma_2) \in U$$

and $g \in L^2(\mathbb{R}^2)$ is an admissible vector for ρ if and only if

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left| \frac{\hat{g}(\gamma_1, \gamma_2)}{\sqrt{|\gamma_1 \gamma_2|}} \right|^2 d\gamma_1\, d\gamma_2 < \infty.$$

Example 2. Let

$$K_1 = \left\{ \begin{pmatrix} a & b \\ -b & a \end{pmatrix} : (a, b) \in \mathbb{R}^2 \backslash \{(0,0)\} \right\}.$$

Let $G = \mathbb{R}^2 \rtimes K_1$. The map $\left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right) \to (1,0)\left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right) = (a, b)$ is a homeomorphism of K_1 onto the open orbit $U = \widehat{\mathbb{R}^2} \backslash \{(0,0)\}$. Again $\mathcal{H}_U^2 = L^2(\mathbb{R}^2)$ and ρ, as given by (2.7), is square integrable. We leave it to the reader to calculate the explicit form of ρ in this case. We prefer to give the equivalent representation π on $L^2(\widehat{\mathbb{R}^2})$.

For $\eta \in L^2(\widehat{\mathbb{R}^2})$, $[\left(\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right), \left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right)] \in G$, and $(\gamma_1, \gamma_2) \in \widehat{\mathbb{R}^2}$,

$$\pi\left[\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} a & b \\ -b & a \end{pmatrix}\right] \eta(\gamma_1, \gamma_2) = \sqrt{a^2 + b^2}\; e^{2\pi i(\gamma_1 x_1 + \gamma_2 x_2)} \eta(a\gamma_1 - b\gamma_2,\ b\gamma_1 + a\gamma_2).$$

An $\eta \in L^2(\widehat{\mathbb{R}^2})$ is an admissible vector for π if and only if

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left| \frac{\eta(\gamma_1, \gamma_2)}{\sqrt{\gamma_1^2 + \gamma_2^2}} \right|^2 d\gamma_1 d\gamma_2 < \infty.$$

Remark. K_1 is just the direct product of the dilations and rotations. That is,

$$K_1 = \{aR_\theta : a > 0 \text{ and } 0 \leq \theta < 2\pi\},$$

where $R_\theta = \left(\begin{smallmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{smallmatrix}\right)$.

For any $r > 0$, define

$$K_r = \left\{ \begin{pmatrix} a & b \\ -rb & a \end{pmatrix} : (a, b) \in \mathbb{R}^2 \backslash \{(0,0)\} \right\}.$$

Then K_r has properties very similar to K_1 and, in fact, K_r is inner conjugate to K_1 in $GL_2(\mathbb{R})$ since

$$K_r = \begin{pmatrix} 1/\sqrt{r} & 0 \\ 0 & 1 \end{pmatrix} K_1 \begin{pmatrix} \sqrt{r} & 0 \\ 0 & 1 \end{pmatrix}.$$

Thus, any K_r, $r > 0$ is equivalent to K_1 via a change of scale on the horizontal axis (cf. Remark (5) in §2).

Example 3. This is a family of examples. For each $c \in \mathbb{R}$, let

$$H_c = \left\{ \begin{pmatrix} a & 0 \\ b & a^c \end{pmatrix} : a, b \in \mathbb{R},\ a > 0 \right\}.$$

If $c \neq 0$, there are two open orbits U_+ and U_- in $\widehat{\mathbb{R}^2}; U_+ = \{(\gamma_1, \gamma_2) : \gamma_2 > 0\}$ and $U_- = \{(\gamma_1, \gamma_2) : \gamma_2 < 0\}$. Note that when $c = 0$, we have the example in Remark (3) at the end of §2. Fix $c \neq 0$. Let $G = \mathbb{R}^2 \rtimes H_c$.

As with the group in example 2, the formula for π, the representation on $L^2(\widehat{\mathbb{R}^2})$ is simpler than that for ρ. For $[(\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}), (\begin{smallmatrix} a & 0 \\ b & a^c \end{smallmatrix})] \in G$, $\eta \in L^2(\widehat{\mathbb{R}^2})$, and $(\gamma_1, \gamma_2) \in \widehat{\mathbb{R}^2}$,

$$\pi\left[\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} a & 0 \\ b & a^c \end{pmatrix}\right] \eta(\gamma_1, \gamma_2) = a^{\frac{c+1}{2}} e^{2\pi i(\gamma_1 x_1 + \gamma_2 x_2)} \eta(a\gamma_1 + b\gamma_2,\ a^c\gamma_2).$$

Then $\pi = \pi_{U_+} \oplus \pi_{U_-}$, where π_{U_+} and π_{U_-} are the subrepresentations of π formed by restriction to $L^2(U_+)$ and $L^2(U_-)$, respectively.

For π_{U_+}, the admissibility condition for $\eta \in L^2(U_+)$ is

$$\int_0^\infty \int_{-\infty}^\infty \left| \frac{\eta(\gamma_1, \gamma_2)}{\gamma_2} \right|^2 d\gamma_1\, d\gamma_2 < \infty.$$

The condition for U_- is similar.

For this family of examples, the case of $c = 1$ leads to the simplest formulas and is representative of the general case in terms of the nature of the action on $\widehat{\mathbb{R}^2}$.

4. Discrete frames. We begin this section by recalling the definition of a frame. If $\mathcal{H}$ is a Hilbert space, a frame in $\mathcal{H}$ is a family of vectors $(\omega_j)_{j \in J}$ such that there exist $A > 0$ and $B < \infty$, called frame bounds, with

$$A\|\eta\|^2 \leq \sum_{j \in J} |\langle \eta, \omega_j \rangle|^2 \leq B\|\eta\|^2 \tag{4.1}$$

for all $\eta \in \mathcal{H}$. For a good discussion of frames constructed from the affine group, see [16, Chap. 3]. In this section, we show how to generate a frame in $\mathcal{H}^2_U$ for the groups $\mathbb{R}^n \rtimes H$ and the square-integrable representations ρ_U discussed in §2. The results are of a general nature and we do not attempt to obtain tight frames (where $A = B$ in (4.1)). For any particular group H that proves useful, detailed analyses can be carried out in later work.

For this section, fix a closed subgroup H of $GL_n(\mathbb{R})$ such that there exists an open free H-orbit U in $\widehat{\mathbb{R}^n}$. Let $\mathcal{H}^2_U$ denote the Hardy space of elements in $L^2(\mathbb{R}^n)$ with Fourier transform supported on U. Let ρ be defined as in (2.7) and ρ_U be the subrepresentation of ρ determined by restriction to $\mathcal{H}^2_U$ as in §2. Our aim is to describe guidelines for selecting an admissible $g \in \mathcal{H}^2_U$ and a discrete set $\left((\underline{x}_j, h_j)\right)_{j \in J}$ in $G = \mathbb{R}^n \rtimes H$ so that $\left(\rho(\underline{x}_j, h_j)g\right)_{j \in J}$ is a frame in $\mathcal{H}^2_U$. In what follows, we have been heavily influenced by [10] and [16].

DEFINITION. *A subset P of H is called* separated *if there exists a neighbourhood V of the identity e in H such that $l^{-1}V \cap k^{-1}V = \emptyset$ for $l \neq k$ and $l, k \in P$.*

It will be convenient for us to express this condition in terms of the action of H on U.

LEMMA 3. *A subset P of H is separated if and only if there exists a compact subset B of U with nonempty interior such that $Bl \cap Bk = \emptyset$ for $l, k \in P, l \neq k$.*

Proof. Fix $\underline{\gamma}_0 \in U$. Suppose P is separated by a neighbourhood V of e in H. We may assume that V is a compact neighbourhood of e. Let

$$B = \{\underline{\gamma}_0 h^{-1} : h \in V\}.$$

Then, the map, $h \to \underline{\gamma}_0 h^{-1}$, is a homeomorphism of H onto U; so B is a compact subset of U with nonempty interior. For $k \in P$, $Bk = \{\underline{\gamma}_0 h^{-1} k : h \in V\} = \{\underline{\gamma}_0 h^{-1} : h \in k^{-1}V\}$. Thus $l^{-1}V \cap k^{-1}V = \emptyset$ implies $Bl \cap Bk = \emptyset$. The converse is similar. □

We will say that P is separated by B in U if the condition of Lemma 3 holds.

LEMMA 4. *Let P be a separated subset of H and let D be a compact subset of U. Then $\sup_{k \in P} \left(\#\{l \in P : Dl \cap Dk \neq \emptyset\}\right) < \infty$.*

Proof. This follows from (i) ⇒ (iii) of Lemma 3.3 in [10], or the reader can show it directly with an easy compactness argument. □

DEFINITION. *A* frame generator *is a pair (P, F), where P is a separated subset of H and F is a compact subset of U, such that $\bigcup_{k \in P} Fk = U$.*

We now describe how to obtain a frame in $\mathcal{H}_U^2$ from a frame generator (P, F) and the square-integrable representation ρ_U. Let D be a compact subset of U with nonempty interior D^0 and $F \subseteq D^0$. Let R be an n-dimensional parallelepiped with $D \subseteq R$. By R being an n-dimensional parallelepiped, we mean there are vectors $\underline{\lambda}^1, \underline{\lambda}^2, \ldots, \underline{\lambda}^n \in \widehat{\mathbb{R}^n}$, linearly independent, and real numbers $a_1 < b_1, a_2 < b_2, \ldots, a_n < b_n$ such that

$$R = \left\{\underline{\gamma} = \sum_{j=1}^{n} \gamma_j \underline{\lambda}^j : a_j \leq \gamma_j \leq b_j,\ 1 \leq j \leq n\right\}.$$

Since $\{\underline{\lambda}^1, \underline{\lambda}^2, \ldots, \underline{\lambda}^n\}$ forms a basis of $\widehat{\mathbb{R}^n}$, we have a dual basis $\{\underline{y}^1, \underline{y}^2, \ldots, \underline{y}^n\}$ for $\mathbb{R}^n$. Let $I = \mathbb{Z}^n$ considered as an index set for the following discrete set in $\mathbb{R}^n$. For each $\underline{i} = (i_1, i_2, \ldots, i_n) \in I$, let

$$\underline{x}^{\underline{i}} = \sum_{j=1}^{n} \left(\frac{i_j}{b_j - a_j}\right) \underline{y}^j$$

and

$$e_{\underline{i}}(\underline{\gamma}) = \begin{cases} 0 & \text{if } \ \underline{\gamma} \in \widehat{\mathbb{R}^n} \backslash R, \\ \dfrac{1}{\sqrt{\lambda(R)}} e^{2\pi i \underline{\gamma} \underline{x}^{\underline{i}}} & \text{if } \ \underline{\gamma} \in R, \end{cases}$$

where $\lambda(R)$ denotes the Lebesgue measure of R. Identify $L^2(R)$ with $\{\eta \in L^2(\widehat{\mathbb{R}^n}) : \eta(\underline{\gamma}) = 0$ for almost all $\underline{\gamma} \in \widehat{\mathbb{R}^n} \backslash R\}$. Then $\{e_{\underline{i}} : \underline{i} \in I\}$ is an orthonormal basis for $L^2(R)$. Let

$$M = \sup_{k \in P} \left(\#\{l \in P : Dl \cap Dk \neq \emptyset\}\right).$$

which is finite by Lemma 4.

Now suppose $g \in \mathcal{H}_U^2$ satisfies the following conditions:

(i) Support of $\hat{g} \subseteq D$. (That is, $\hat{g}(\underline{\gamma}) = 0$ for almost all $\underline{\gamma} \in \widehat{\mathbb{R}^n} \backslash D$.)

(ii) $a = \inf\left\{|\hat{g}(\underline{\gamma})| : \underline{\gamma} \in F\right\} > 0$.

(iii) $b = \sup\left\{|\hat{g}(\underline{\gamma})| : \underline{\gamma} \in D\right\} < \infty$.

Let $A = \lambda(R)a^2$ and $B = \lambda(R)Mb^2$.

THEOREM 3. *With the notation which has been established above,* $\{\rho(\underline{x}^{\underline{i}},k)^{-1}g : \underline{i}\in I, k\in P\}$ *is a frame for* $\mathcal{H}^2_U$ *with frame bounds* A *and* B.

Proof. For $(\underline{x},h)\in G, (\underline{x},h)^{-1} = \big((\underline{x},e)(\underline{o},h)\big)^{-1} = (\underline{o},h^{-1})(-\underline{x},e)$. Thus, for $f\in\mathcal{H}^2_U$, $\underline{i}\in I$, and $k\in P$,

$$\begin{aligned}
\langle f,\rho(\underline{x}^{\underline{i}},k)^{-1}g\rangle_{\mathcal{H}^2_U} &= \langle\rho(\underline{o},k)f,\ \rho(-\underline{x}^{\underline{i}},e)g\rangle_{\mathcal{H}^2_U}\\
&= \langle\pi(\underline{o},k)\hat f,\pi(-\underline{x}^{\underline{i}},e)\hat g\rangle_{L^2(U)}\\
\text{(by 2.9)}\qquad &= \int_U \delta(k)^{\frac{1}{2}}\hat f(\underline{\gamma}k)e^{2\pi i\underline{\gamma}\underline{x}^{\underline{i}}}\,\overline{\hat g}(\underline{\gamma})d\underline{\gamma}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\sum_{k\in P}\sum_{\underline{i}\in I}\left|\langle f,\rho(\underline{x}^{\underline{i}},k)^{-1}g\rangle_{\mathcal{H}^2_U}\right|^2 &= \sum_{k\in P}\delta(k)\lambda(R)\sum_{\underline{i}\in I}\left|\int_R \overline{\hat f}(\underline{\gamma}k)\hat g(\underline{\gamma})\overline{e_{\underline{i}}}(\underline{\gamma})d\underline{\gamma}\right|^2\\
&= \lambda(R)\sum_{k\in P}\int_R \delta(k)\left|\hat f(\underline{\gamma}k)\right|^2\left|\hat g(\underline{\gamma})\right|^2 d\underline{\gamma}\\
\text{(by 2.3)}\qquad &= \lambda(R)\sum_{k\in P}\int_U \left|\hat f(\underline{\gamma})\right|^2\left|\hat g(\underline{\gamma}k^{-1})\right|^2 d\underline{\gamma}\\
\text{(4.2)}\qquad &= \lambda(R)\int_U \left|\hat f(\underline{\gamma})\right|^2\sum_{k\in P}\left|\hat g(\underline{\gamma}k^{-1}\right|^2 d\underline{\gamma}.
\end{aligned}$$

On the one hand, using the fact that (P,F) is a frame generator, we have $\sum_{k\in P}\left|\hat g(\underline{\gamma}k^{-1})\right|^2 \geq a^2$, for all $\underline{\gamma}\in U$. Thus, by (4.2),

$$\begin{aligned}
\sum_{k\in P}\sum_{\underline{i}\in I}\left|\langle f,\rho(\underline{x}^{\underline{i}},k)^{-1}g\rangle_{\mathcal{H}^2_U}\right|^2 &\geq \lambda(R)a^2\int_U\left|\hat f(\underline{\gamma})\right|^2 d\underline{\gamma}\\
\text{(4.3)}\qquad &= A\|f\|^2_{\mathcal{H}^2_U}.
\end{aligned}$$

On the other hand, for a given $\underline{\gamma}\in U$, $\hat g(\underline{\gamma}k^{-1})\neq 0$ for at most M values of $k\in P$ and $\left|\hat g(\underline{\gamma}k^{-1})\right|\leq b$, for any of those values. Thus,

$$\begin{aligned}
\sum_{k\in P}\sum_{\underline{i}\in I}\left|\langle f,\rho(\underline{x}^{\underline{i}},k)^{-1}g\rangle_{\mathcal{H}^2_U}\right|^2 &\leq \lambda(R)Mb^2\int_U\left|\hat f(\underline{\gamma})\right|^2 d\underline{\gamma}\\
\text{(4.4)}\qquad &= B\|f\|^2_{\mathcal{H}^2_U}.
\end{aligned}$$

Together, (4.3) and (4.4) mean that $\{\rho(\underline{x}^{\underline{i}},k)^{-1}g : \underline{i}\in I, k\in P\}$ is a frame for $\mathcal{H}^2_U$ with frame bounds A and B. □

Remark. The proof of Theorem 3 is clearly an easy adaptation of the standard arguments for the affine group (see, for example, [16, §5.1.2]).

5. Frame generators in two dimensions. We now use the two dimensional examples from §3 to illustrate how frame generators, as described in the previous section, can be easily found in particular examples.

Example 1. In this case, $A = \{\left(\begin{smallmatrix}a_1 & 0\\ 0 & a_2\end{smallmatrix}\right) : a_1,a_2\in\mathbb{R}\backslash\{0\}\}$ and $G = \mathbb{R}^2\rtimes A$. Note that, by including negatives of the dilations, there is a unique open A-orbit $U = \{(\gamma_1,\gamma_2)\in\widehat{\mathbb{R}^2} : \gamma_1\neq 0, \gamma_2\neq 0\}$ in $\widehat{\mathbb{R}^2}$. Let

$$P = \left\{\begin{pmatrix}\pm 2^m & 0\\ 0 & \pm 2^n\end{pmatrix} : m,n\in\mathbb{Z}\right\}$$

and let $F = [1,2] \times [1,2]$. Then P is separated (by $[1-\varepsilon, 1+\varepsilon] \times [1-\varepsilon, 1+\varepsilon]$ as long as $0 < \varepsilon < \frac{1}{3}$) and $\bigcup_{k \in P} Fk = U$. Thus (P, F) is a frame generator for $\mathbb{R}^2 \rtimes A$ and the associated square-integrable representation ρ on $L^2(\mathbb{R}^2)$.

Example 2. We now study $K_1 = \{aR_\theta : a > 0,\ 0 \leq \theta < 2\pi\}$. One can easily form P by separately discretizing the a's and θ's and then find an appropriate "fundamental domain" F. However, it is important to note that P can be of the form $\{h^n : n \in \mathbb{Z}\}$ for a single fixed h. Let

$$h = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} = \sqrt{2} \begin{pmatrix} \cos\frac{\pi}{4} & \sin\frac{\pi}{4} \\ -\sin\frac{\pi}{4} & \cos\frac{\pi}{4} \end{pmatrix}$$

and let $P = \{h^n : n \in \mathbb{Z}\}$. The action on $\widehat{\mathbb{R}^2}$ of h is a simultaneous rotation and dilation. It is easily verified that P is a separated set, with respect to $U = \widehat{\mathbb{R}^2} \backslash \{(0,0)\}$, which is the open orbit for K_1. Let F be the trapezoid with vertices at $(\frac{1}{16}, 0), (1,0), (\frac{1}{16}, \frac{1}{16})$, and $(1,1)$. Then $U = \bigcup_{n \in \mathbb{Z}} Fh^n$. That is, (P, F) is a frame generator.

Remark. It is known that one can do much better in this situation (see [13], for example). If we let $\Gamma = \mathbb{Z}^2 \subseteq \mathbb{R}^2$, combining translations by elements of Γ and dilations by powers of h admits a multiresolution analysis and leads to an orthonormal basis in $L^2(\mathbb{R}^2)$, not just a frame as is given by Theorem 3.

Example 3. We select one representative example from the family $\{H_c : c \neq 0\}$ of groups from Example 3 in §3. If $c = 1$,

$$H_1 = \left\{ \begin{pmatrix} a & 0 \\ b & a \end{pmatrix} : a, b \in \mathbb{R},\ a > 0 \right\}.$$

There are two open H_1-orbits in $\widehat{\mathbb{R}^2}$. We will consider one of them

$$U_+ = \{(\gamma_1, \gamma_2) : \gamma_2 > 0\}$$

for illustration. Let

$$P = \left\{ \begin{pmatrix} 2^n & 0 \\ m2^n & 2^n \end{pmatrix} : m, n \in \mathbb{Z} \right\}$$

and let F denote the trapezoid with vertices at $(0,1), (1,1), (0,2)$ and $(2,2)$. It is a pleasant exercise to show that the open upper half-plane is tiled by the translates of F under P. That is, $U_+ = \bigcup_{m,n \in \mathbb{Z}} F\left(\begin{smallmatrix} 2^n & 0 \\ m2^n & 2^n \end{smallmatrix}\right)$. It is also easily checked that P is a separated set relative to U_+. Therefore, (P, F) is a frame generator.

6. Concluding remarks. Our purpose was to show the wide availability of square-integrable representations on $\mathbb{R}^n$. The group that is being represented is of the form $\mathbb{R}^n \rtimes H$ with the elements of $\mathbb{R}^n$ acting on $\mathbb{R}^n$ as translations and the elements of H acting as generalized dilations. There is a great deal of liberty in selecting H. Among the n-dimensional closed subgroups of $GL_n(\mathbb{R})$, many (perhaps most) will act in such a manner that open free H-orbits exist in $\widehat{\mathbb{R}^n}$. We have illustrated this point with two-dimensional examples.

Once H has been selected so that there exists an open free H-orbit U in $\widehat{\mathbb{R}^n}$, then a representation ρ_U is naturally defined on the Hardy space $\mathcal{H}^2_U$. We showed that this ρ_U is square integrable and gave an elementary proof of the Duflo–Moore relations for such a representation. Moreover, for ρ_U, the admissibility conditions are particularily easy to check and the "formal dimension" operator has an explicit description.

We made the first steps towards constructing good frames in $\mathcal{H}_U^2$ in §4. There is clearly much remaining to be done. Especially, in cases where one knows that a particular dilation group H is well suited to an application at hand, one needs to refine the analysis to obtain tight frames or even an orthonormal basis.

Acknowledgments. We would like to acknowledge useful conversations with James Brooke, Patrick Browne, Qingde Yang, and Eberhard Kaniuth while we were developing these ideas.

REFERENCES

[1] S. T. ALI, J. P. ANTOINE, AND J. P. GAZEAU, *Square-integrability of group representations on homogeneous spaces* I: *Reproducing triples and frames* and II: *Generalized square-integrability and equivalent families of coherent states*, Ann. Inst. H. Poincaré, 55 (1991), pp. 829–890.

[2] J. P. ANTOINE, P. CARRETTE, R. MURENZI, AND B. PIETTE, *Image analysis with two-dimensional continuous wavelet transform*, Signal Processing, 31 (1993), pp. 241–272.

[3] L. BAGGETT AND K. F. TAYLOR, *Groups with completely reducible regular representation*, Proc. Amer. Math. Soc., 72 (1978), pp. 593–600.

[4] ———, *A sufficient condition for the complete reducibility of the regular representation*, J. Funct. Anal., 34 (1979), pp. 250–265.

[5] A. L. CAREY, *Square integrable representations of non-unimodular groups*, Bull. Austral. Math. Soc., 15 (1976), pp. 1–12.

[6] I. DAUBEHIES, *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[7] S. DE BIÈVRE, *Coherent states over symplectic homogeneous spaces*, J. Math. Phys., 30 (1989), pp. 1401–1407.

[8] J. DIXMIER, *C^*-Algebras*, North–Holland, Amsterdam, 1977.

[9] M. DUFLO AND C. C. MOORE, *On the regular representation of a nonunimodular locally compact group*, J. Funct. Anal., 21 (1976), pp. 209–243.

[10] H. FEICHTINGER AND K. GROCHENIG, *Banach space related to integrable group representations and their atomic decompositions* I, J. Funct. Anal., 86 (1989), pp. 307–340.

[11] J. M. G. FELL AND R. S. DORAN, *Representations of *-Algebras, Locally Compact Groups, and Banach *-Algebraic Bundles*, Vol. I and II, Academic Press, Boston, 1988.

[12] J. GLIMM, *Locally compact transformation groups*, Trans. Amer. Math. Soc., 101 (1961), pp. 124–138.

[13] K. GROCHENIG AND W. R. MADYCH, *Multiresolution analysis, Haar bases and self-similar tilings of* $\mathbb{R}^n$, IEEE Trans. Inform. Theory, 38 (1992), pp. 556–568.

[14] A. GROSSMANN AND J. MORLET, *Decomposition of Hardy functions into square integrable wavelets of constant shape*, SIAM J. Math. Anal., 15 (1984), pp. 723–736.

[15] A. GROSSMANN, J. MORLET, AND T. PAUL, *Transforms associated to square integrable group representations* I: *General results*, J. Math. Phys., 27 (1985), pp. 2473–2479.

[16] C. E. HEIL AND D. F. WALNUT, *Continuous and discrete wavelet transforms*, SIAM Review, 31 (1989), pp. 628–666.

[17] J. PHILLIPS, *A note on square-integrable representations*, J. Funct. Anal., 20 (1975), pp. 83–92.

[18] K. F. TAYLOR, *Geometry of the Fourier algebras and locally compact groups with atomic unitary representations*, Math. Ann., 262 (1983), pp. 183–190.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 609–619, May 1996

POLYCONVEX FUNCTIONALS FOR NEARLY CONFORMAL DEFORMATIONS*

TADEUSZ IWANIEC† AND ADAM LUTOBORSKI†

Abstract. Variational integrals whose absolute minima are conformal deformations are studied. Polyconvexity and mean coercivity of these functionals are proved in even dimensions. Existence of nearly conformal deformations is established.

Key words. polyconvex functions, mean coercive functionals, nearly conformal deformations

AMS subject classifications. 49A, 30C60, 73C60

1. Introduction. In this article, we investigate a class of variational integrals with nonconvex and noncoercive integrands. Such functionals arise naturally in quasiconformal analysis and nonlinear elasticity. They are referred to as energy functionals for a deformation $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ and take the form

$$\mathcal{E}[f] = \int_\Omega E(\nabla f)dx. \tag{1.1}$$

In elasticity, the integrand $E : \mathbb{R}^{n\times n} \to \mathbb{R}_+$ is called the stored-energy function and it encodes the mechanical properties of the material. The minimum points $A \in \mathbb{R}^{n\times n}$ of E are called potential wells. Obviously, the derivative $E' : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ vanishes on the potential wells and E posesses stress-free states. Our basic assumption is that E vanishes exactly on the matrices of linear conformal mappings, that is, on the set $\mathcal{C}^+(n) = \{\lambda A\,; A \in SO(n),\ \lambda \geq 0\}$. Therefore, $\mathcal{E}[f]$ measures (in an average sense) how far is f from a conformal deformation. If f is conformal, then $\mathcal{E}[f] = 0$ and we say that f is an absolute minimizer of $\mathcal{E}$. The minimizers (with prescribed boundary values) of our energy functional will be referred to as nearly conformal mappings.

The simplest example of such a functional is the complex p-harmonic intergral

$$\mathcal{E}[f] = \int_\Omega \left|\frac{\partial f}{\partial \bar{z}}\right|^p, \qquad 1 < p < \infty,$$

for functions $f : \Omega \subset \mathbb{C} \to \mathbb{C}$ of Sobolev class $W^{1,p}(\Omega)$. Its minima, called the complex p-harmonic functions, are found by solving the nonhomogeneous Cauchy–Riemann equation

$$\frac{\partial f}{\partial \bar{z}} = |h(z)|^{q-2}\overline{h(z)},$$

where h is a holomorphic function and q is the Hölder conjugate to p.

Define a nonlinear operator $H : \mathbb{R}^{n\times n} \to \mathbb{R}_+$ acting on gradient matrices as

$$H(\nabla f) = |\nabla f|^n - n^{\frac{n}{2}} \det \nabla f. \tag{1.2}$$

Its differential $H' : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ induces another operator

$$H'(\nabla f) = n|\nabla f|^{n-2}\nabla f - n^{\frac{n}{2}}\ \mathrm{adj}\nabla f. \tag{1.3}$$

*Received by the editors March 18, 1994; accepted for publication (in revised form) September 26, 1994.

†Department of Mathematics, Syracuse University, 215 Carnegie Building, Syracuse, NY 13244-1150. The research of the first author was supported in part by National Science Foundation grant DMS-9401104.

Both H and H' vanish on conformal matrices. H is nonnegative due to Hadamard's inequality. When $n = 2$, we find that $H(\nabla f) = |\partial f/\partial \bar{z}|^2$ and $H'(\nabla f) = 2\partial f/\partial \bar{z}$. Therefore, a natural generalization of the complex p-harmonic integral to higher dimensions might be the functional

$$\mathcal{I}_p[f] = \int_\Omega (|\nabla f|^n - n^{\frac{n}{2}} \det \nabla f)^{p/2}, \quad 1 < p < \infty, \tag{1.4}$$

defined for all mappings $f : \Omega \to \mathbb{R}^n$ in the Sobolev class $W^{1,\frac{np}{2}}(\Omega, \mathbb{R}^n)$.

For $n > 2$, the function $H : \mathbb{R}^{n\times n} \to \mathbb{R}_+$ is not convex. Another serious obstacle in proving the existence of minima for $\mathcal{I}_p$ is the lack of coercivity of H. In [IL], using polyconvexity and *mean coercivity*, we succeeded in showing that for $p \geq 2$ the functional $\mathcal{I}_p$ attains its minimum in the Sobolev class $W^{1,\frac{np}{2}}(\Omega, \mathbb{R}^n)$ with prescribed boundary values of $g \in W^{1,\frac{np}{2}}(\Omega, \mathbb{R}^n)$. The case $1 < p < 2$ is still open. For $p \geq 2$, Liouville's theorem asserts that the absolute minimizers of $\mathcal{I}_p$ in the class $W^{1,\frac{np}{2}}(\Omega, \mathbb{R}^n)$ must be Möbius transformations (or constant mappings); see [G], [R1], [BI], and [IM].

Let $f = (f^1, \ldots, f^n) : \Omega \to \mathbb{R}^n$, $n = 2l$, be a mapping of the Sobolev class $W^{1,l}(\Omega, \mathbb{R}^n)$. Set $N = \binom{n}{l}$ and denote by ∇f^{l*l} the $N \times N$ matrix of all $l \times l$ minors of ∇f. More precisely, the entries of ∇f^{l*l} are indexed by ordered l-tuples $I = (i_1, \ldots, i_l)$, $1 \leq i_1 < \cdots < i_l \leq n$, $J = (j_1, \cdots, j_l)$, $1 \leq j_1 < \cdots < j_l \leq n$, and

$$(\nabla f^{l*l})_I^J = \begin{vmatrix} \frac{\partial f^{j_1}}{\partial x_{i_1}} & \cdots & \frac{\partial f^{j_1}}{\partial x_{i_l}} \\ \vdots & & \vdots \\ \frac{\partial f^{j_l}}{\partial x_{i_1}} & \cdots & \frac{\partial f^{j_l}}{\partial x_{i_l}} \end{vmatrix}.$$

These determinants are integrable functions on Ω. For $p \geq 1$, we introduce the class $\mathcal{L}^{p,l}(\Omega, \mathbb{R}^n)$ of mappings $f \in W^{1,l}(\Omega, \mathbb{R}^n)$ such that $\nabla f^{l*l} \in L^p(\Omega, \mathbb{R}^{N\times N})$.

In Lemma 2.1 of §2 we show that the expression $H_l(A) = |A^{l*l}|^2 - N \det A$ for $A \in \mathbb{R}^{n\times n}$ is nonnegative and vanishes only when A is either a conformal matrix or of rank less than l. Thus an alternative to the p-harmonic integral in the complex plane is the functional

$$\mathcal{F}_p[f] = \int_\Omega (|\nabla f^{l*l}|^2 - N \det \nabla f)^{\frac{p}{2}} \tag{1.5}$$

for mappings $f \in \mathcal{L}^{p,l}(\Omega, \mathbb{R}^n)$. $\mathcal{F}_p$ is polyconvex for all $1 \leq p < \infty$; see Lemma 2.3. $\mathcal{F}_p$ is possibly mean coercive, as it is if $p \geq 2 - \epsilon$ for some $\epsilon = \epsilon(n) > 0$; see [IL].

Another functional of interest to us is

$$\mathcal{W}_\alpha[f] = \int_\Omega |\nabla f^{l*l}(x)|^{2\alpha} [|\nabla f^{l*l}(x)|^2 - N \det \nabla f(x)]\, dx \tag{1.6}$$

for mappings $f \in \mathcal{L}^{2\alpha+2,l}(\Omega, \mathbb{R}^n)$. The case $l = 1$ reduces to the functional offered by Alibert and Dacorogna [AD] in two dimensions,

$$\mathcal{W}_\alpha[f] = \int_\Omega |\nabla f|^{2\alpha} (|\nabla f|^2 - 2 \det \nabla f) = \int (|f_z|^2 + |f_{\bar{z}}|^2)^\alpha |f_{\bar{z}}|^2 \tag{1.7}$$

Our main result addresses the polyconvexity of the functional $\mathcal{W}_\alpha$.

THEOREM 1.1. *Let l be a positive integer, $n = 2l$, $0 \leq \alpha \leq 1$, and $A, B \in \mathbb{R}^{n\times n}$. Then*

$$W_\alpha(A) - W_\alpha(B) \geq \langle \mathcal{M}, A^{l*l} - B^{l*l} \rangle + \lambda(\det A - \det B), \tag{1.8}$$

where $\mathcal{M} = \mathcal{M}(B^{l*l}) \in \mathbb{R}^{N\times N}$ *and* $\lambda = \lambda(B^{l*l}) \in \mathbb{R}$ *are given explicitly by the formulas*

$$\mathcal{M}(X) = 4|X|^{2\alpha-2}(|X|^2 + 2\alpha|X^-|^2)X^+, \tag{1.9}$$

$$\lambda(X) = -4|X|^{2\alpha-2}(|X|^2 + \alpha|X^-|^2) \tag{1.10}$$

for $X \in \mathbb{R}^{N\times N}$ *(see* §2 *for notation used here).*

This result is new even in two dimensions. Only the cases of $\alpha = 1$ (see [AD], [IL]) and the case $\alpha = 0$, which is obvious, have been known. Our next result deals with the mean coercivity of $\mathcal{W}_\alpha$.

THEOREM 1.2. *Suppose* $f, g \in \mathcal{L}^{2\alpha+2,l}(\Omega, \mathbb{R}^n)$ *are mappings such that* $f - g \in \overset{\circ}{W}{}^{1,l}(\Omega, \mathbb{R}^n)$. *Then*

$$\delta \int_\Omega |\nabla f^{l*l}|^{2\alpha+2} \le \mathcal{W}_\alpha[f] + \int_\Omega |\nabla^{l*l} g|^{2\alpha+2}, \tag{1.11}$$

where we point out that $\delta = \delta(l, \alpha) > 0$ *is independent of* $f, g,$ *and* Ω.

In particular, we see that a minimizer f of $\mathcal{W}_\alpha$ subject to the Dirichlet condition $f - g \in \overset{\circ}{W}{}^{1,l}(\Omega, \mathbb{R}^n)$ automatically belongs to $\mathcal{L}^{2\alpha+2,l}(\Omega, \mathbb{R}^n)$.

Theorems 1.1 and 1.2 allow us to apply the direct method in solving the following minimization problem.

THEOREM 1.3. *Let* $0 \le \alpha \le 1$, $\epsilon > 0$, *and* $g \in \mathcal{L}^{2\alpha+2,l}(\Omega, \mathbb{R}^n)$ *be given. Then the functional*

$$\mathcal{F}[f] = \epsilon\Big(\int_\Omega |\nabla f|^l\Big)^{2\alpha+2} + \mathcal{W}_\alpha[f] \tag{1.12}$$

attains its minimum in the class of mappings $f \in W^{1,l}(\Omega, \mathbb{R}^n)$ *such that* $f - g \in \overset{\circ}{W}{}^{1,l}(\Omega, \mathbb{R}^n)$.

Although we have shown that $\mathcal{W}_\alpha$ is coercive in the mean for all $\alpha \ge 0$, it ceases to be polyconvex for large α. We address this problem in two dimensions in §4. Therein we examine the range of parameter α for which W_α is rank-one convex.

2. Preliminaries from multilinear algebra. We begin by introducing some matrix notations. Let l be a positive integer and $n = 2l$, $N = \binom{n}{l}$. We denote by $\mathbb{R}^{n\times n}$ the space of $n \times n$ matrices $A, B, C, \dots$ and by $\mathbb{R}^{N\times N}$ the space of $N \times N$ matrices $X, Y, Z, \dots$. Spaces of matrices are equipped with the inner product $\langle X, Y\rangle = \operatorname{trace}(X^T Y)$ for $X, Y \in \mathbb{R}^{N\times N}$ and an associated norm $|X|^2 = \langle X, X\rangle$. We also define the matrix signum function

$$\operatorname{sgn}(X) = \begin{cases} \frac{X}{|X|} & \text{if } X \ne 0, \\ 0 & \text{if } X = 0. \end{cases}$$

For $A \in \mathbb{R}^{n\times n}$, we define $A^{l*l} \in \mathbb{R}^{N\times N}$

$$(A^{l*l})^I_J = \begin{vmatrix} a^{i_1}_{j_1} & \cdots & a^{i_1}_{j_l} \\ \vdots & & \vdots \\ a^{i_l}_{j_1} & \cdots & a^{i_l}_{j_l} \end{vmatrix},$$

where $I = (i_1, \dots, i_l)$ and $J = (j_1, \dots, j_l)$ are l-tuples such that $1 \le i_1 < \cdots < i_l \le n$ and $1 \le j_1 < \cdots < j_l \le n$.

It is standard in the theory of determinants that for $A, B \in \mathbb{R}^{n\times n}$, we have $(A^T)^{l*l} = (A^{l*l})^T$ and $(AB)^{l*l} = B^{l*l}A^{l*l}$. Hence if U is orthogonal, then U^{l*l} is also orthogonal. We will need the following version of the Hadamard inequality, which follows from [IL, Lem. 2.1].

LEMMA 2.1. *Let $n = 2l$, $N = \binom{n}{l}$, and $A \in \mathbb{R}^{n\times n}$. Then*

$$H_l(A) = |A^{l*l}|^2 - N \det A \geq 0 \tag{2.1}$$

and equality occurs iff one of the following two conditions holds:

(i) *$A^{l*l} = 0$, that is, rank$(A) < l$;*

(ii) *$A^T A = \lambda I$, where $\lambda > 0$, that is, A is a similarity matrix.*

For $X = (X^I_J) \in \mathbb{R}^{N\times N}$, we define the complementary matrix $\tilde{X} \in \mathbb{R}^{N\times N}$ by $\tilde{X}^I_J = (-1)^{|I|+|J|}X^{I'}_{J'}$ where I' and J' denote the l-tuples complementary to I and J. Clearly, the transformation $\tilde{}: \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}$ is an idempotent self-adjoint operator, which is an isometry with respect to the inner product in $\mathbb{R}^{N\times N}$,

$$\tilde{\tilde{X}} = X, \tag{2.2}$$

$$\langle X, \tilde{Y}\rangle = \langle \tilde{X}, Y\rangle, \tag{2.3}$$

$$\langle \tilde{X}, \tilde{Y}\rangle = \langle X, Y\rangle \tag{2.4}$$

for all $X, Y \in \mathbb{R}^{N\times N}$.

Indeed, (2.2) follows directly from the definition, whereas (2.4) is implied by (2.2) and (2.3).

$$\begin{aligned}\langle X, \tilde{Y}\rangle &= \sum_{I,J}(-1)^{|I|+|J|}X^I_J Y^{I'}_{J'} = \sum_{I',J'}(-1)^{|I'|+|J'|}X^{I'}_{J'}Y^I_J\\ &= \langle \tilde{X}, Y\rangle.\end{aligned}$$

For $X \in \mathbb{R}^{N\times N}$, we define the matrices X^+ and X^-, called the conformal and anticonformal parts of X, by

$$X^+ = \frac{1}{2}(X + \tilde{X}), \qquad X^- = \frac{1}{2}(X - \tilde{X}), \tag{2.5}$$

and hence

$$X = X^+ + X^-. \tag{2.6}$$

For any $X, Y \in \mathbb{R}^{N\times N}$,

$$\langle X^+, Y^-\rangle = 0. \tag{2.7}$$

Indeed, by (2.3) and (2.4),

$$4\langle X^+, Y^-\rangle = \langle X, Y\rangle - \langle X, \tilde{Y}\rangle + \langle \tilde{X}, Y\rangle - \langle \tilde{X}, \tilde{Y}\rangle = 0.$$

In other words, the decomposition of X into conformal and anticonformal parts defines an orthogonal decomposition of $\mathbb{R}^{N\times N}$. Hence, if $X = X^+ + X^-$, then

$$|X|^2 = |X^+|^2 + |X^-|^2. \tag{2.8}$$

LEMMA 2.2. *For any $X, Y \in \mathbb{R}^{N\times N}$,*

$$|X^+| - |Y^+| \geq \langle sgn\, Y^+, X - Y\rangle. \tag{2.9}$$

Proof. From the Cauchy–Schwarz inequality, we see that $|X^+| \geq \langle \text{sgn } Y^+, X^+\rangle$ and $|Y^+| = \langle \text{sgn } Y^+, Y^+\rangle$. Hence (2.7) yields

$$|X^+| - |Y^+| \geq \langle \text{sgn } Y^+, X^+ - Y^+\rangle = \langle \text{sgn } Y^+, X - Y\rangle.$$

LEMMA 2.3. *Let $n = 2l$, $N = \binom{n}{l}$. For $A \in \mathbb{R}^{n\times n}$, let $X = A^{l*l}$. Then*

$$|X^+|^2 - |X^-|^2 = N \det A. \tag{2.10}$$

In particular, the form $[H_l(A)]^{p/2} = 2^{p/2}|X^-|^p$ is convex with respect to the $l \times l$ minors of A for all $p \geq 1$.

Proof. Using the Laplace expansion formula [DS, p. 46], in the penultimate step, we get

$$\begin{aligned}|X^+|^2 - |X^-|^2 = \langle X, \tilde{X}\rangle &= \sum_{I,J}(-1)^{|I|+|J|}X_J^I X_{J'}^{I'} \\ &= \sum_I \det A = N \det A.\end{aligned}$$

Remark 1. For $A \in \mathbb{R}^{n\times n}$, and $X = A^{l*l}$ we have that $X^T\tilde{X} = \det A\, I$

3. Polyconvex conformal stored energies in even dimensions. Polyconvex integrands have been introduced by J. M. Ball [B] as a generalization of null Lagrangians. For a thorough discussion of the subject, we refer the reader to [D] and [IL].

DEFINITION 3.1. *A function $W : \mathbb{R}^{n\times n} \to \mathbb{R}$ is said to be polyconvex if for every $B \in \mathbb{R}^{n\times n}$ there exists $\mathcal{B} \in \mathbb{R}^{2^n\times 2^n}$ such that*

$$W(A) - W(B) \geq \langle \mathcal{B}, T(A) - T(B)\rangle$$

for any $A \in \mathbb{R}^{n\times n}$, where $T(A)$ denotes the $2^n \times 2^n$ matrix of all minors of A of all orders.

The main result of this section is the following.

THEOREM 3.1. *Let $0 \leq \alpha \leq 1$, $n = 2l$, $N = \binom{n}{l}$, where l is a positive integer. Then the matrix function $W_\alpha : \mathbb{R}^{n\times n} \to \mathbb{R}$ given by*

$$W_\alpha(A) = |A^{l*l}|^{2\alpha}\Big(|A^{l*l}|^2 - N \det A\Big) \tag{3.1}$$

is polyconvex. More precisely,

$$W_\alpha(A) - W_\alpha(B) \geq \langle \mathcal{M}, A^{l*l} - B^{l*l}\rangle + \lambda(\det A - \det B), \tag{3.2}$$

where $\mathcal{M} = \mathcal{M}(B^{l*l})$, $\lambda = \lambda(B^{l*l})$, and

$$\mathcal{M}(X) = 4|X|^{2\alpha-2}(|X|^2 + 2\alpha|X^-|^2)X^+, \tag{3.3}$$

$$\lambda(X) = -4|X|^{2\alpha-2}(|X|^2 + \alpha|X^-|^2) \tag{3.4}$$

for $X \in \mathbb{R}^{N\times N}$.

Our proof is preceeded by two lemmas.

LEMMA 3.2. *Let S be an open subset of $\mathbb{R}^n$ and let $h \in C^2(\overline{S})$ be a nonnegative function vanishing on the boundary of S whose Hessian matrix is positive in S. Then*

$$h(a) - h(b) \geq \langle \nabla h(b), a - b \rangle \tag{3.5}$$

for all $a, b \in \overline{S}$.

Proof. Inequality (3.5) is well known in the theory of convex functions if the segment $[a, b]$ is contained in $\overline{S}$. To generalize it, we denote by $[x, b] \subset [a, b]$ the largest subsegment contained in $\overline{S}$. If $x \neq a$, then x belongs to the boundary of S. Accordingly,

$$h(x) - h(b) \geq \langle \nabla h(b),\ x - b \rangle.$$

Since $h(x) = 0$ and $a - b = \lambda(x - b)$ for some $\lambda > 1$, we obtain

$$\begin{aligned}\langle \nabla h(b)\ ,\ a - b \rangle &= \lambda \langle \nabla h(b),\ x - b \rangle \\ &\leq \lambda h(x) - \lambda h(b) \leq h(a) - h(b).\end{aligned}$$

LEMMA 3.3. *For $0 \leq \alpha \leq 1$, consider the function*

$$h(t, d) = (t^2 - 2d)^\alpha (t^2 - 4d) \tag{3.6}$$

in $S = \{(t, d) \in \mathbb{R}^2 : t^2 - 4d \geq 0\}$.

Then its Hessian matrix

$$H = \begin{pmatrix} h_{dd} & h_{dt} \\ h_{td} & h_{tt} \end{pmatrix}$$

is positive definite.

Proof. The zero set and the support S of the function h are depicted in Fig. 1. An elementary calculation shows that

$$h_{dd} = 2^{4-\alpha} \alpha [2t^2 - 4d]^{\alpha - 2} [2t^2 + (1 + \alpha)(t^2 - 4d)] \geq 0 \tag{3.7}$$

and

(3.8)
$$\det H = 4\alpha [2t^2 - 4d]^{2\alpha - 3} [(\alpha + 1) 3t^2 (t^2 - 4d) + (\alpha + 1)(2\alpha + 1)(t^2 - 4d)^2 - 2(\alpha - 1) t^4] \\ \geq 0$$

as desired.

Proof of Theorem 3.1. From Lemmas 3.2 and 3.3, we infer that

$$h(t_A, d_A) - h(t_B, d_B) \geq \tau(t_B, d_B)(t_A - t_B) + \delta(t_B, d_B)(d_A - d_B) \tag{3.9}$$

for all pairs (t_A, d_A) and (t_B, d_b) from S, where τ and δ are the partials of h with respect to t and d. More explicitly,

$$\tau(t, d) = 2t(t^2 - 2d)^{\alpha - 1} \Big[\alpha(t^2 - 4d) + t^2 - 2d\Big], \tag{3.10}$$

$$\delta(t, d) = -2(t^2 - 2d)^{\alpha - 1} \Big[\alpha(t^2 - 4d) + 2(t^2 - 2d)\Big] \tag{3.11}$$

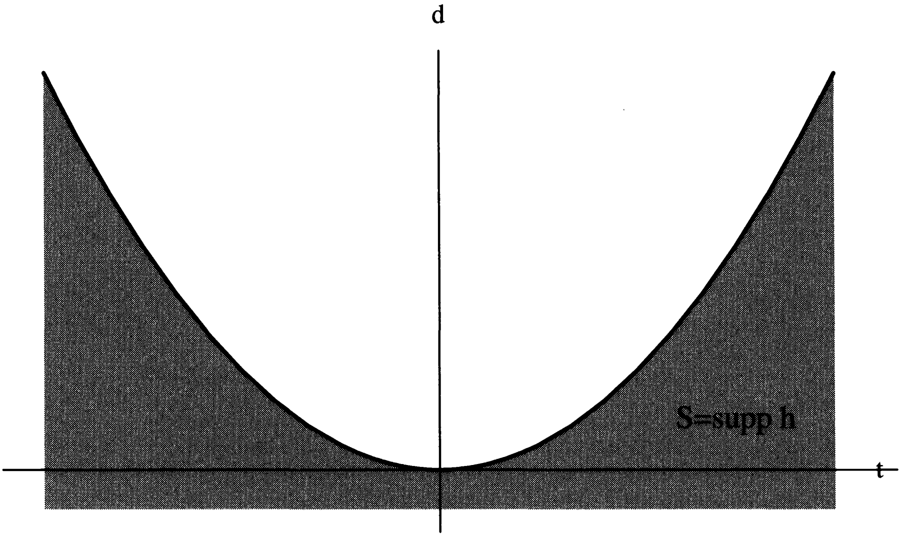


FIG. 1.

for all (t_A, d_A), $(t_B, d_B) \in S$. Next, we define two real functions $t, d : \mathbb{R}^{N\times N} \to \mathbb{R}$,

$$t(X) = \sqrt{2}|X^+| \tag{3.12}$$

$$d(X) = \frac{1}{2}(|X^+|^2 - |X^-|^2) \tag{3.13}$$

for $X \in \mathbb{R}^{N\times N}$.

To shorten notations, we write $X = A^{l*l}$ and $Y = B^{l*l}$ for $A, B \in \mathbb{R}^{n\times n}$.

According to (2.8),

$$(t^2 - 2d)(X) = |X|^2, \tag{3.14}$$

$$(t^2 - 4d)(X) = 2|X^-|^2. \tag{3.15}$$

Hence for $0 \leq \alpha \leq 1$ in the definition of the function h introduced in Lemma 3.3, we obtain that

$$\begin{aligned} W_\alpha(A) &= 2|X|^{2\alpha}|X^-|^2 \\ &= (t^2(X) - 2d(X))^\alpha (t^2(X) - 4d(X)) \\ &= h(t(X), d(X)) \end{aligned} \tag{3.16}$$

Due to Lemma 3.3, for any $A, B \in \mathbb{R}^{n\times n}$, we have

$$W_\alpha(A) - W_\alpha(B) \geq \tau_B(t(X) - t(Y)) + \delta_B(d(X) - d(Y)), \tag{3.17}$$

where

$$\tau_B = 2\sqrt{2}|Y|^{2(\alpha-1)}|Y^+|(2\alpha|Y^-|^2 + |Y|^2), \tag{3.18}$$

$$\delta_B = -4|Y|^{2(\alpha-1)}(\alpha|Y^-|^2 + |Y|^2). \tag{3.19}$$

We note that $\tau_B \geq 0$, because $\alpha \geq 0$ for all $B \in \mathbb{R}^{n\times n}$.

We may now use Lemma 2.2. Accordingly,

$$\begin{aligned} t(X) - t(Y) &= \sqrt{2}(|X^+| - |Y^+|) \\ &\geq \sqrt{2}\langle \operatorname{sgn} Y^+, X - Y\rangle. \end{aligned}$$

Finally, we conclude with the desired estimate

$$W_\alpha(A) - W_\alpha(B) \geq \langle \mathcal{M}, A^{l*l} - B^{l*l}\rangle + \lambda(\det A - \det B), \tag{3.20}$$

where

$$\mathcal{M} = 4|B^{l*l}|^{2(\alpha-1)}\Big[2\alpha|(B^{l*l})^-|^2 + |B^{l*l}|^2\Big](B^{l*l})^+, \tag{3.21}$$

$$\lambda = -4|B^{l*l}|^{2(\alpha-1)}\Big[\alpha|(B^{l*l})^-|^2 + |B^{l*l}|^2\Big]. \tag{3.22}$$

4. Mean coercivity. In this section we prove Theorem 1.2.

Proof of Theorem 1.2. To each ordered l-tuple $I = (i_1, \ldots, i_l)$, $1 \leq i_1 < \cdots < i_l \leq 1$, we assign two l-forms on Ω, namely,

$$\varphi_I = df^{i_1} \wedge \cdots \wedge df^{i_l}, \qquad \psi_I = dg^{i_1} \wedge \cdots \wedge dg^{i_l}. \tag{4.1}$$

The coefficients of these forms are the $l \times l$ minors of ∇f and ∇g, respectively. Of course, $(\sum |\varphi_I|^2)^{\frac{1}{2}} = |\nabla f^{l*l}| \in L^{2\alpha+2}(\Omega)$ and $(\sum |\psi_I|^2)^{\frac{1}{2}} = |\nabla g^{l*l}| \in L^{2\alpha+2}(\Omega)$. The hypothesis that $f, g \in W^{1,l}(\Omega, \mathbb{R}^n)$ are essential to ensure that the forms φ_I and ψ_I are closed in the distributional sense. Notice also that

$$\varphi_I - \psi_I = d\sum_{k=1}^{l}(-1)^k(f^{i_k} - g^{i_k})dg^{i_1} \wedge \cdots \wedge dg^{i_k - 1} \wedge df^{i_k + 1} \wedge \cdots \wedge df^{i_l}. \tag{4.2}$$

Since $f - g \in \overset{\circ}{W}{}^{1,l}(\Omega, \mathbb{R}^n)$, it follows that the zero extension of $\varphi_I - \psi_I$,

$$\omega_I = \begin{cases} \varphi_I - \psi_I & \text{in } \Omega, \\ 0 & \text{in } \mathbb{R}^n - \Omega, \end{cases}$$

remains exact in the entire space $\mathbb{R}^n$. For each ordered l-tuple I, we consider its complementary l-tuple I' ordered in such a way that $\operatorname{sgn}(I, I') = 1$. Therefore,

$$\varphi_I \wedge \varphi_{I'} = df^{i_1} \wedge \cdots \wedge df^{i_l} \wedge df^{i'_1} \wedge \cdots \wedge df^{i'_l} = \det \nabla f. \tag{4.3}$$

As before, the form $\omega_{I'} \in \mathcal{L}^{2\alpha+2}(\mathbb{R}^n)$ is also exact. We now use Proposition 9.1 in [IL] with $n = 2l$. Accordingly,

$$(1-\delta)\int_\Omega (|\omega_I|^2 + |\omega_{I'}|^2)^{\alpha+1} \leq \int_\Omega (|\omega_I|^2 + |\omega_{I'}|^2)^\alpha[|\omega_I|^2 + |\omega_{I'}|^2 - 2\omega_I \wedge \omega_{I'}], \tag{4.4}$$

where $\delta = \delta(l, \alpha) \in [0, 1)$.

Recalling that $\omega_I = \varphi_I - \psi_I$ and $\omega_{I'} = \varphi_{I'} - \psi_{I'}$, with the aid of Young's inequality, we routinely arrive at the estimate

$$\begin{aligned} \epsilon\int_\Omega (|\varphi_I|^2 + |\varphi_{I'}|^2)^{\alpha+1} &\leq \int_\Omega (|\varphi_I|^2 + |\varphi_{I'}|^2)^\alpha[|\varphi_I|^2 + |\varphi_{I'}|^2 - 2\varphi_I \wedge \varphi_{I'}] \\ &\quad + C_\epsilon \int_\Omega (|\psi_I|^2 + |\psi_{I'}|^2)^{\alpha+1} \end{aligned} \tag{4.5}$$

for some $\epsilon = \epsilon(l, \alpha) > 0$.

On the other hand, for each l-tuple I, we have

$$0 \leq |\varphi_I|^2 + |\varphi_{I'}|^2 - 2\varphi_I \wedge \varphi_{I'} \leq \sum_I (|\varphi_I|^2 + |\varphi_{I'}|^2 - 2\varphi_I \wedge \varphi_{I'}) \tag{4.6}$$
$$= 2\Big[|\nabla f^{l*l}|^2 - N \det \nabla f\Big].$$

Hence, summing (4.5) with respect to I, we conclude with the desired inequality

$$\delta \textstyle\int_\Omega |\nabla f^{l*l}|^{2\alpha+2} \leq \displaystyle\int_\Omega |\nabla f^{l*l}|^{2\alpha}\Big[|\nabla f^{l*l}|^2 - N \det \nabla f\Big] + \int_\Omega |\nabla g^{l*l}|^{2\alpha+2}$$

for some $\delta = \delta(l,\alpha) > 0$.

5. Existence of nearly conformal deformations. Having at our disposal the polyconvexity and mean coercivity of the integrand W_α, we may now implement the direct method of the calculus of variations.

Proof of Theorem 1.3. Let $\{f_j\}_{j\geq 1}$ be a minimizing sequence in $\mathcal{L}^{2\alpha+2,l}(\Omega, \mathbb{R}^n)$ for the functional $\mathcal{F}$ such that $f_j - g \in \overset{\circ}{W}{}^{1,l}(\Omega, \mathbb{R}^n)$.

Mean coercivity yields

$$\epsilon\Big(\int_\Omega |\nabla f_j|^l\Big)^{2\alpha+2} + \delta \int_\Omega |\nabla f_j^{l*l}|^{2\alpha+2} \leq \mathcal{F}[f_j] + \int_\Omega |\nabla g^{l*l}|^{2\alpha+2}. \tag{5.1}$$

This estimate, together with the boundary condition $f_j - g \in \overset{\circ}{W}{}^{1,l}(\Omega, \mathbb{R}^n)$, implies that the minimizing sequence is bounded in $W^{1,l}(\Omega, \mathbb{R}^n)$. Therefore, there is no loss of generality in assuming that $\{f_j\}_{j\geq 1}$ converges weakly to a mapping f. Of course, f satisfies the Dirichlet boundary condition $f - g \in \overset{\circ}{W}{}^{1,l}(\Omega, \mathbb{R}^n)$. It remains to be shown that f minimizes $\mathcal{F}$.

Clearly,

$$\Big(\int_\Omega |\nabla f|^l\Big)^{2\alpha+2} \leq \liminf_{j\to\infty}\Big(\int_\Omega |\nabla f_j|^l\Big)^{2\alpha+1}. \tag{5.2}$$

Using weak continuity of the minors (see [R2], [B]), we observe that $\nabla f_j^{l*l} \to \nabla f^{l*l}$ in the sense of Schwarz distributions. On the other hand, by (5.1), the sequence ∇f_j^{l*l} is bounded in $L^{2\alpha+2}(\Omega, \mathbb{R}^{N\times N})$. These facts ensure that ∇f_j^{l*l} is actually weakly convergent in $L^{2\alpha+2}(\Omega, \mathbb{R}^{N\times N})$ and hence $f \in \mathcal{L}^{2\alpha+2,l}(\Omega, \mathbb{R}^n)$.

Also, by Proposition 5.2 of [IL], we see that $\det \nabla f_j \to \det \nabla f$ in the sense of distributions. By Hadamard's inequality (2.1), we see that $\det \nabla f_j$ is a bounded sequence in $L^{1+\alpha}(\Omega)$, where we note that $1+\alpha > 1$. These arguments show that $\det \nabla f_j \to \det \nabla f$ weakly in $L^{1+\alpha}(\Omega)$.

The final step requires the polyconvexity of W_α established in Theorem 3.1.

$$\int_\Omega W_\alpha(\nabla f_j) - \int_\Omega W_\alpha(\nabla f) \geq \int_\Omega \langle \mathcal{M}, \nabla f_j^{l*l} - \nabla f^{l*l}\rangle \tag{5.3}$$
$$+ \int_\Omega \lambda(\det \nabla f_j - \det \nabla f).$$

From (3.21), we see at once that $\mathcal{M} = \mathcal{M}(\nabla f^{l*l})$ belongs to $L^{\frac{2\alpha+2}{2\alpha+1}}(\Omega, \mathbb{R}^{N\times N})$, which is the dual of $L^{2\alpha+2}(\Omega, \mathbb{R}^{N\times N})$. Similarly, by (3.22), $\lambda = \lambda(\nabla f^{l*l})$ belongs to $L^{(\alpha+1)/\alpha}(\Omega)$, which is dual to $L^{\alpha+1}(\Omega)$. Consequently, the right-hand side of (5.3) converges to zero, which implies the inequality

$$\int_\Omega W_\alpha(\nabla f) \geq \liminf_{j\to\infty} \int_\Omega W_\alpha(\nabla f_j). \tag{5.4}$$

Combining (5.4) with (5.2) yields

$$\mathcal{F}[f] \leq \liminf_{j\to\infty} \mathcal{F}[f_j],$$

completing the proof of Theorem 1.3.

6. Rank-one convexity. Considerable progress has been made in the study of nonconvex variational problems in two dimensions. Specific classes of integrands have been analyzed in great detail [AD], [A], [DDGR], [DK], [DM], [RS], [Š2].

These efforts resulted in showing that the sets of convex, polyconvex, quasiconvex, and rank-one-convex matrix functions form an increasing sequence. All but the last inclusion have been shown to be strict in two dimensions. In higher dimensions, the fact that rank-one convexity does not imply quasi convexity has been established by V. Šverák in [Š1].

In the context of the quoted work, it is appropriate to summarize the results of §3 in two dimensions. When $n = 2l = N = \binom{n}{l} = 2$, Theorem 3.1 becomes the following.

PROPOSITION 6.1. *The function $W_\alpha : \mathbb{R}^{2\times 2} \to \mathbb{R}$ given by*

$$W_\alpha(A) = |A|^{2\alpha}(|A|^2 - 2\det A) \tag{6.1}$$

is polyconvex for all $0 \leq \alpha \leq 1$.

Polyconvexity in the case $\alpha = 1$ has been proved in [AD] and [IL], and the case $\alpha = 0$ is obvious. The precise values of α for which W_α is polyconvex, quasi-convex, and rank-one convex remain unknown. A new step in this direction is the following.

PROPOSITION 6.2. *If $-\frac{1}{8} \leq \alpha \leq \frac{3+\sqrt{13}}{4}$, then the function W_α is rank-one convex.*

Proof. Polyconvexity implies rank-one convexity. Hence it follows from Proposition 6.1 that W_α is rank-one convex for $0 \leq \alpha \leq 1$.

In [DDGR, Prop. 1.1], Dacorogna, Douchet, Gangbo and Rappaz found the necessary and sufficient conditions for rank-one convexity in two dimensions. Accordingly, W_α is rank-one convex iff

$$\Phi_\alpha(u,v,y) = 2[(\alpha+1)\alpha - 2\alpha(\alpha-1)y]u^2 - 4\alpha uv + \alpha + 1 - 2\alpha y \geq 0 \tag{6.2}$$

for all $y, u, v \in \mathbb{R}$ satisfying $u^2 + v^2 \leq 1$, $(u+v)^2 - 1 \leq 2y \leq 1 - (u-v)^2$.

Part 1: *Proof of* (6.2) *for* $1 < \alpha \leq \frac{3+\sqrt{13}}{4}$. Since $\Phi_\alpha(u,v,y)$ is linear in y, it is enough to show that (6.2) holds for the maximal value of y, namely, $y^+ = \frac{1}{2}(1-(u-v)^2)$. Therefore, (6.2) is reduced to proving that for all $u^2+v^2 \leq 1$,

$$\frac{1}{\alpha}\Phi_\alpha(u,v,y^+) = \frac{1}{\alpha} + (u-v)(5u-v) + 2(\alpha-1)u^2(u-v)^2 \geq 0. \tag{6.3}$$

Using our assumption on α and dropping a nonnegative term, we get

$$\frac{1}{\alpha}\Phi_\alpha(u,v,y^+) \geq \sqrt{13} - 3 + (u-v)(5u-v) \geq (\sqrt{13}-3)(u^2+v^2) + (u-v)(5u-v)$$
$$= \left(u\sqrt{\sqrt{13}+2} - v\sqrt{\sqrt{13}-2}\right)^2 \geq 0.$$

Part 2: *Proof of* (6.2) *for* $-\frac{1}{8} \leq \alpha < 0$. It suffices to prove (6.2) for $|2y| \leq 1$, hence

$$\begin{aligned}\Phi_\alpha(u,v,y) &\geq 1 - |\alpha|(4 + 2|\alpha - 1| + 2|\alpha + 1|) \\ &= 1 + 8\alpha \geq 0\end{aligned}$$

as desired.

PROPOSITION 6.3. *For* $\alpha = 2$ *and* $\alpha < -\frac{1}{3}$, *the integrand* W_α *is not rank-one convex.*

Indeed, when $\alpha = 2$ in Part 1, then $\Phi_2(\frac{\sqrt{5}}{5}, \frac{2\sqrt{5}}{5}, \frac{2}{5}) = -\frac{1}{25} < 0$. In Part 2, if $u = -v = \frac{\sqrt{2}}{2}$ and $2y = -1$ and α is rank-one convex, then we infer that $\alpha > \frac{\sqrt{2}-2}{2} > -\frac{1}{3}$.

Acknowledgment. Both authors thank Vladimir Šverák for discussions at Cortona and at the Institute for Advanced Study, respectively.

REFERENCES

[AD] J.-J. ALIBERT AND B. DACOROGNA, *An example of a quasiconvex function that is not polyconvex in two dimensions*, Arch. Rational Mech. Anal., 117 (1992), pp. 155–166.

[A] G. AUBERT, *On a counterexample of a rank* 1 *convex function which is not polyconvex in the case* $N = 2$, Proc. Roy. Soc. Edinburgh Sect A, 106 (1987), pp. 237–240.

[B] J. M. BALL, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 63 (1977), pp. 337–403.

[BI] B. BOJARSKI AND T. IWANIEC, *Another approach to Liouville theorem*, Math. Nachr., 107 (1982), pp. 253–262.

[D] B. DACOROGNA, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, New York, 1989.

[DDGR] B. DACOROGNA, J. DOUCHET, W. GANGBO, AND J. RAPPAZ, *Some examples of rank one convex functions in dimension two*, Proc. Roy. Soc. Edinburgh Sect. A, 114 (1990), pp. 135–150.

[DK] B. DACOROGNA AND H. KOSHIGOE, *On the different notions of convexity for rotationally invariant functions*, Ann. Fac. Sci. Toulouse Math., II (1993), pp. 163–184.

[DM] B. DACOROGNA AND P. MARCELLINI, *A counterexample in the vectorial calculus of variations*, in Material Instabilities in Continuum Mechanics, J. M. Ball, ed., Oxford University Press, Oxford, 1988, pp. 77–83.

[DS] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part* 1, Interscience, New York, 1967.

[IL] T. IWANIEC AND A. LUTOBORSKI, *Integral estimates for null Lagrangians*, Arch. Rational Mech. Anal., 125 (1993), pp. 25–79.

[IM] T. IWANIEC AND G. MARTIN, *Quasiregular mappings in even dimensions*, Acta Math., 170 (1993), pp. 29–81.

[G] F. W. GEHRING, *Rings and quasiconformal mappings in space*, Trans. Amer. Math. Soc., 103 (1962), pp. 353–393.

[R1] Y. G. RESHETNYAK, *Liouville's conformal mapping theorem under minimal regularity assumption*, Siberian Math. J., 8 (1967), pp. 835–840.

[R2] ———, *Stability theorems for mappings with bounded distortion*, Siberian Math. J., 8 (1968), pp. 499–512.

[RS] P. ROSAKIS AND H. C. SIMPSON, *On the relation between polyconvexity and rank-one convexity in nonlinear elasticity*, J. Elasticity, 37 (1995), pp. 113–137.

[Š1] V. ŠVERÁK, *Rank-one convexity does not imply quasiconvexity*, Proc. Roy. Soc. Edinburgh Sect. A, 120 (1992), pp. 185–189.

[Š2] ———, *Quasiconvex functions with subquadratic growth*, Proc. Roy. Soc. London, 433 (1991), pp. 723–725.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 620–637, May 1996

THE GRADIENT THEORY OF THE PHASE TRANSITIONS IN CAHN–HILLIARD FLUIDS WITH DIRICHLET BOUNDARY CONDITIONS*

KAZUHIRO ISHIGE†

Abstract. We are interested in the asymptotic behavior of minimizers (as $\epsilon \to 0$) of the variational problems under the Dirichlet condition

$$\inf\left\{ \int_\Omega \left[\epsilon|\nabla u|^2 + \frac{1}{\epsilon}W(x,u)\right] dx \;\middle|\; u \in W^{1,2}(\Omega : \mathbf{R}^n),\, u = g \quad \text{on } \partial\Omega \right\},$$

where $W(x,\cdot)$ is a nonnegative function with only two zeros α and β. Here α and β are independent of the space variable x. In this paper, we will show that the limit of a sequence of minimizers $\{u_\epsilon\}_{\epsilon>0}$ (as $\epsilon \to 0$) is a solution of another variational problem without boundary condition. However the limit variational problem has a boundary integral corresponding to transition layers near $\partial\Omega$. Our analysis relies mainly on the theory of gamma convergence. In order to overcome the difficulty of inhomogeneity of the boundary condition, we approximate $g(x)$ by suitable piecewise smooth functions near the boundary $\partial\Omega$.

Key words. phase transition, Cahn–Hilliard fluids, gamma convergence, singular perturbation

AMS subject classifications. 49J45, 49Q20

1. Introduction.

In this paper, we will investigate the asymptotic behavior of minimizers $\{u_\epsilon\}_{\epsilon>0}$ (as $\epsilon \to 0$) of the following variational problem:

$$(P_\epsilon) \qquad \inf\left\{ \int_\Omega \left[\epsilon|\nabla u|^2 + \frac{1}{\epsilon}\, W(x,u)\right] dx \;\middle|\; u \in W^{1,2}(\Omega : \mathbf{R}^n),\, u = g \text{ on } \partial\Omega \right\},$$

where Ω is a bounded domain in $\mathbf{R}^N$ with C^2 smooth boundary $\partial\Omega$ and g is a Lipschitz continuous function from $\partial\Omega$ into $\mathbf{R}^n$. Here $W(x,\cdot)$ is a nonnegative continuous function with only two zeros $\alpha, \beta \in \mathbf{R}^n$, and α, β are independent of the space variable x. This type of problem is related to the study of the phase transitions of Cahn–Hilliard fluids. See [13], [18], and [20].

In [12], R. V. Kohn and P. Sternberg conjectured that minimizers of the variational problem

$$(SP_\epsilon) \qquad \inf\left\{ \int_\Omega \left[\epsilon|\nabla u|^2 + \frac{1}{\epsilon}(u^2-1)^2\right] dx \;\middle|\; u \in W^{1,2}(\Omega),\, u|_{\partial\Omega} = g \right\},$$

which is a special case of (P_ϵ), converge to a solution of

$$\inf\left\{ \frac{8}{3}P_\Omega\{u=1\} + 2\int_{\partial\Omega} |d(u)-d(g)|d\mathcal{H}_{N-1} \;\middle|\; u \in BV(\Omega),\, |u| = 1 \text{ a.e.} \right\},$$

where $d(t) = \int_{-1}^t |s^2-1|ds$ and $\mathcal{H}_{N-1}$ is the $N-1$-dimensional Hausdorff measure. This conjecture was affirmatively solved by N. C. Owen, J. Rubinstein, and P. Sternberg in [18].

In this paper, we will study the asymptotic behavior of minimizers of (P_ϵ) and extend the result of [18] to the vector case ($n \geq 2$). Our proof for the vectorial case is more complicated than that for the scalar case ($n = 1$).

*Received by the editors March 30, 1994; accepted for publications (in revised form) August 24, 1994.

†Graduate School of Information Sciences, Tohoku University, Aoba-ku, Sendai 980-77, Japan (ishige@math.is.tohoku.ac.jp).

Our approach is based on the theory of gamma convergence, in which the asymptotic behavior of minimizers $\{u_\epsilon\}_{\epsilon>0}$ of the variational problem (P_ϵ) is characterized by a solution of another variational problem without the Dirichlet conditions.

Recently, using the theory of gamma convergence, several authors have studied the asymptotic behavior of the minimizers of the problem

$$(E_\epsilon) \qquad \inf\left\{ \int_\Omega \left[\epsilon|\nabla u|^2 + \frac{1}{\epsilon} W(u) \right] dx \;\middle|\; u \in W^{1,2}(\Omega : \mathbf{R}^n), \int_\Omega u(x)dx = m \right\},$$

where m is a constant vector in $\mathbf{R}^n$. For the scalar case (i.e., $n = 1$), see [1], [6], [10], [12]–[17], and [19]. For the vector case (i.e., $n \geq 2$), see [3], [4], [9], and [11]. Our results on the problem (P_ϵ) depend mainly on the study of the asymptotic-behavior of minimizers of (E_ϵ). However, there are several different aspects between the asymptotic-behavior of minimizers of (P_ϵ) and those of (E_ϵ). In fact, minimizers of (E_ϵ) generate only an interior layer, while minimizers of (P_ϵ) generate both an interior and a boundary layer as $\epsilon \to 0$.

On the other hand, we can easily see that minimizers of (SP_ϵ) satisfy the equation

$$(CP_\epsilon) \qquad \begin{cases} \epsilon^2 \Delta u - u(u-1)(u+1) = 0 & \text{in } \Omega, \\ u(x) = g(x) & \text{on } \partial\Omega. \end{cases}$$

In this case, there are several results for the solutions of (CP_ϵ) obtained by using the method of matched expansion. Our results also seem to be closely related to [5] and [8].

We give the precise hypotheses on the functions $W(x,u)$ and $g(x)$. Let $W(x,u) : \overline{\Omega} \times \mathbf{R}^n \to \mathbf{R}$ be a continuous nonnegative function, and for any $x \in \overline{\Omega}$, $W(x,u) = 0$ if and only if $u = \alpha$ or β. Here we denote two constant vectors independent of x by α and β . We assume that there exist two constants K_1 and K_2 such that

$$(1.1) \qquad \sup_{u \in \partial[K_1,K_2]^n} W(x,u) \leq W(x,v) \qquad \text{for all } x \in \overline{\Omega},\ v \notin [K_1,K_2]^n$$

and

$$(1.2) \qquad g(x) \in [K_1,K_2]^n \qquad \text{for all } x \in \partial\Omega.$$

Moreover, we assume that for any $\epsilon > 0$, there exists a positive constant δ such that

$$(1.3) \qquad |W^{1/2}(x,u) - W^{1/2}(y,u)| \leq \epsilon W^{1/2}(x,u)$$

for all $x, y \in \overline{\Omega}$ with $|x-y| \leq \delta$ and for all $u \in \mathbf{R}^n$.

In order to state our main theorem, we will introduce a Riemannian metric d on $\mathbf{R}^n$, which is a modification of the one introduced in [3] and [9]. For $x \in \overline{\Omega}$ and $a, b \in \mathbf{R}^n$, let $d(x,a,b)$ be the metric defined by

$$(1.4) \qquad d(x,a,b) = \inf\left\{ \int_0^1 W^{1/2}(x,\gamma(t))|\,\dot\gamma(t)\,|\,dt \;\middle|\; \gamma \in C^1([0,1] : \mathbf{R}^n), \gamma(0) = a,\ \gamma(1) = b \right\}.$$

For example, in the case of $W(x,u) = (u^2-1)^2$ and $n = 1$, we have

$$d(x,-1,b) = \int_{-1}^{b} |s^2 - 1| ds \qquad \text{for} \quad b \geq -1.$$

We now state our main theorem in this paper.

THEOREM 1. *Let $g : \partial\Omega \to \mathbf{R}^n$ be a Lipschitz-continuous function which satisfies* (1.2). *Suppose that the function W satisfies* (1.1) *and* (1.3). *For $\epsilon > 0$, let u_ϵ be a solution of the variational problem*

$$(P_\epsilon)\qquad \inf\left\{ \int_\Omega \left[\epsilon|\nabla u|^2 + \frac{1}{\epsilon} W(x,u)\right] dx \;\middle|\; u \in W^{1,2}(\Omega : \mathbf{R}^n),\ u|_{\partial\Omega}(x) = g(x)\right\}.$$

If there exist a positive sequence $\{\epsilon_i\}_{i=1}^\infty$ and a function $u_0(x) \in L^1(\Omega : \mathbf{R}^n)$ such that

$$(1.5)\qquad \lim_{i\to\infty} \epsilon_i = 0 \quad \text{and} \quad \lim_{i\to\infty} u_{\epsilon_i} = u_0 \quad \text{in } L^1(\Omega : \mathbf{R}^n),$$

then

$$W(x, u_0(x)) = 0 \quad \text{for almost all } x \in \Omega,$$

that is, $u_0(x) = \alpha$ or β for almost all $x \in \Omega$. Moreover, the set $E_0 = \{x \in \Omega \,|\, u_0(x) = \alpha\}$ is a solution of the following variational problem:

$$(P_0)\qquad \inf\left\{ \int_{\Omega\cap\partial^* E} d(x,\alpha,\beta)d\mathcal{H}_{N-1} + \int_{\partial\Omega} d(x, v|_{\partial\Omega}(x), g(x))d\mathcal{H}_{N-1} \;\middle|\; E \subset \Omega,\ P_\Omega(E) < \infty,\ v = \alpha\chi_E + \beta\chi_{\Omega\setminus E}\right\},$$

where $P_\Omega(E)$ is the perimeter of E in Ω and $v|_{\partial\Omega}$ is the trace of v on $\partial\Omega$. Furthermore, we have

$$\lim_{i\to\infty}\int_\Omega \left[\epsilon_i|\nabla u_{\epsilon_i}|^2 + \frac{1}{\epsilon_i} W(x,u_{\epsilon_i})\right] dx = 2\int_{\Omega\cap\partial^* E_0} d(x,\alpha,\beta)d\mathcal{H}_{N-1}$$
$$+ 2\int_{\partial\Omega\cap\partial^* E_0} d(x,\alpha,g(x))d\mathcal{H}_{N-1} + 2\int_{\partial\Omega\setminus\partial^* E_0} d(x,\beta,g(x))d\mathcal{H}_{N-1}.$$

Here $\partial^* E_0$ is the reduced boundary of E_0.

Remark 1. It is not restrictive to assume that there exists a subsequence $\{u_{\epsilon_i}\}_{i=1}^\infty$ satisfying (1.5). In fact, the following is proved in [9] and [11]: if there exist constants C and R such that

$$(1.6)\qquad \inf_{x\in\overline{\Omega}} W(x,u) \geq C|u| \qquad \text{for} \quad |u| \geq R,$$

then there exists a subsequence $\{u_{\epsilon_i}\}_{i=1}^\infty$ satisfying (1.5).

Remark 2. Consider two continuous functions $W(u)$ and $h(x)$, where $W(u)$ satisfies condition (1.1) and $h(x)$ is positive function in $\overline{\Omega}$. If the function $W(x,u)$ has a form of $h(x)W(u)$, then $W(x,u)$ satisfies conditions (1.1) and (1.3).

It is worth noting that the proof of the vector case has completely different difficulties relative to that of the scalar case ($n = 1$). For the vector case, it is important to select a minimizing sequence $\{\gamma_k(x,t)\}_{k=1}^\infty$ achieving the value of $d(x,\alpha,g(x))$. One of the difficulties in the treatment of the vector case is that $\gamma_k(x,t)$ is not necessary continuous in the space variable x. Hence it seems difficult to apply the method of [18] directly. In order to overcome this difficulty, we approximate $W(\cdot,u)$ and $g(\cdot)$ by suitable piecewise smooth functions near the transition layer and the boundary $\partial\Omega$. For further details, see Step 2 in §4.

The plan of this paper is as follows: In §2, we present some preliminary results and state Propositions A and B, which are crucial in our analysis. In §3, adopting the method of [4] to our problem, we will prove Proposition A. In §4, we will construct approximate functions $\{w_\epsilon\}_{\epsilon>0}$ to the minimizers $\{u_\epsilon\}_{\epsilon>0}$ and prove Proposition B. In §5, we give some direct applications of the results obtained in the previous sections.

2. Preliminary results and main propositions. We first introduce the notation used in this paper. Let Ω be an open bounded set in $\mathbf{R}^N$ with C^2 smooth boundary $\partial\Omega$. For any $\nu \in \mathbf{S}^{N-1}$, we denote the open unit cube centered at the origin with two of its surfaces normal to ν by Q_ν, i.e., if $\{\nu_1, \dots, \nu_{N-1}, \nu\}$ is a orthonormal system of $\mathbf{R}^N$, then

$$Q_\nu \equiv \left\{ x \in \mathbf{R}^N \,\middle|\, |x \cdot \nu_i| < \frac{1}{2}, |x \cdot \nu| < \frac{1}{2}, i = 1, \dots, N-1 \right\}.$$

We denote the Banach space of functions of bounded variation by $BV(\Omega)$. We denote the Lebesgue N-dimensional measure by $\mathcal{L}_N$ and set $|E| = \mathcal{L}_N(E)$ for any measurable set E of $\mathbf{R}^N$. Furthermore, we denote the $N-1$-dimensional Hausdorff measure by $\mathcal{H}_{N-1}$.

In this section, we recall some properties of functions in $BV(\Omega)$ and state two propositions which are the hearts of the matter in our analysis.

For $u \in L^1(\Omega)$, we say $u \in BV(\Omega)$ if and only if there are Radon measures $\mu_1, \mu_2, \dots, \mu_N$ defined in Ω such that for $i = 1, 2, \dots, N$, the total variation of $\mu_i < \infty$ and

$$\int_\Omega u \frac{\partial}{\partial x_i} \varphi dx = -\int_\Omega \varphi d\mu_i \qquad \text{for all} \quad \varphi \in C_0^\infty(\Omega).$$

In what follows, we write $\nabla u = (\mu_1, \mu_2, \dots, \mu_N)$. For any positive continuous function h defined on $\overline{\Omega}$, we set

$$\int_\Omega h(x)|\nabla u| \equiv \sup\left\{ \int_\Omega u \operatorname{div} g dx \,\middle|\, g \in C_0^1(\Omega : \mathbf{R}^N),\ |g| \le h \right\}.$$

If $u_\epsilon \to u$ in $L^1(\Omega)$, then we have

$$\liminf_{\epsilon \to 0_+} \int_\Omega h(x)|\nabla u_\epsilon| \ge \int_\Omega h(x)|\nabla u|. \tag{2.1}$$

On the other hand, for $u \in BV(\Omega)$ and any continuous function $\varphi(x, y) \in C(\Omega \times \mathbf{R})$, the co-area formula is read as

$$\int_\Omega \varphi(x, u(x))|\nabla u| = \int_{-\infty}^{+\infty} dt \int_{\{u(x)=t\}} \varphi(x, t) d\mathcal{H}_{N-1}(x). \tag{2.2}$$

Furthermore, even if $\partial\Omega$ is a Lipschitz-continuous boundary, we can define the trace of u on $\partial\Omega$ for $u \in BV(\Omega)$.

For any measurable subset E of $\mathbf{R}^N$, let χ_E be the characteristic function of E. If $\chi_E \in BV(\Omega)$, we say that E has finite perimeter in Ω. Then we set

$$P_\Omega(E) = \int_\Omega |\nabla \chi_E|, \qquad \nu_r = -\int_{B(x,r)} \nabla \chi_E \Big/ \int_{B(x,r)} |\nabla \chi_E|. \tag{2.3}$$

We define the reduced boundary of E as $\partial^* E$. We say $x \in \partial^* E$ if and only if

$$\int_{B(x,r)} |\nabla \chi_E| > 0 \qquad \text{for all} \quad r > 0$$

and the limit $\nu(x) = \lim_{r \to 0} \nu_r(x)$ exists with $|\nu(x)| = 1$. We have

$$\int_\Omega h(x)|\nabla \chi_E| = \int_{\partial^* E} h(x) d\mathcal{H}_{N-1}. \tag{2.4}$$

For further details on functions of bounded variation, see [7] and [21].

For $u \in L^1(\Omega : \mathbf{R}^n)$ and $\epsilon > 0$, we define $F_\epsilon(u)$, $F_0(u)$ by

$$F_\epsilon(u) = \begin{cases} \int_\Omega \left[\epsilon|\nabla u|^2 + \frac{1}{\epsilon}W(x,u)\right] dx & \text{if } u \in W^{1,2}(\Omega : \mathbf{R}^n) \text{ and } u = g \text{ on } \partial\Omega, \\ +\infty & \text{otherwise,} \end{cases}$$

$$F_0(u) = \begin{cases} 2\int_\Omega d(x,\alpha,\beta)|\nabla \chi_{\{u(x)=\alpha\}}| + 2\int_{\partial\Omega} d(x, u|_{\partial\Omega}(x), g(x)) d\mathcal{H}_{N-1} \\ \qquad \text{if } u \in BV(\Omega : \mathbf{R}^n) \text{ and } W(x,u(x)) = 0 \text{ for almost all } x \in \Omega, \\ +\infty \qquad \text{otherwise.} \end{cases}$$

In order to prove our main theorem, we need the following two propositions, which are crucial in our analysis.

PROPOSITION A. *Suppose that $\{v_\epsilon\}_{\epsilon>0}$ is a sequence in $L^1(\Omega : \mathbf{R}^n)$ which converges in $L^1(\Omega : \mathbf{R}^n)$ as $\epsilon \to 0_+$ to a function v_0. If*

$$\liminf_{\epsilon \to 0_+} F_\epsilon(v_\epsilon) < +\infty, \tag{2.5}$$

then v_0 is a function in $BV(\Omega : \mathbf{R}^n)$ such that $W(x, v_0(x)) = 0$ a.e. $x \in \Omega$, $E = \{x \in \Omega \,|\, v_0(x) = \alpha\}$ has finite perimeter, and

$$F_0(v_0) \le \liminf_{\epsilon \to 0_+} F_\epsilon(v_\epsilon). \tag{2.6}$$

PROPOSITION B. *Suppose that $w_0 \in L^1(\Omega : \mathbf{R}^n)$ can be written as $w_0 = \alpha\chi_E + \beta\chi_{\Omega \setminus E}$, where E is a measurable subset of Ω with finite perimeter. Then there exists a sequence $\{w_\epsilon\}_{\epsilon>0}$ in $W^{1,2}(\Omega : \mathbf{R}^n)$ which converges in $L^1(\Omega : \mathbf{R}^n)$ as $\epsilon \to 0_+$ to w_0 and such that*

$$\limsup_{\epsilon \to 0_+} F_\epsilon(w_\epsilon) \le F_0(w_0). \tag{2.7}$$

Using Propositions A and B, we can prove Theorem 1 in the same manner as in [13]. For completeness, we will give the proof of Theorem 1.

Proof of Theorem 1. Let $\{u_{\epsilon_i}\}_{i=0}^\infty$ be a sequence in $W^{1,2}(\Omega : \mathbf{R}^n)$ with $\lim_{i \to \infty} u_{\epsilon_i} = u_0$ in $L^1(\Omega : \mathbf{R}^n)$, where each u_{ϵ_i} is a solution of (P_{ϵ_i}). From Proposition B, there exists a sequence $\{v_\epsilon\}_{\epsilon>0}$ such that $\lim_{\epsilon \to 0} v_\epsilon = \alpha$ in $L^1(\Omega : \mathbf{R}^n)$ and $\limsup_{\epsilon \to 0} F_\epsilon(v_\epsilon) \le F_0(\alpha) < \infty$. Then we have $\limsup_{i \to \infty} F_{\epsilon_i}(u_{\epsilon_i}) < \infty$, and from Proposition A, we see that

$$F_0(u_0) \le \liminf_{i \to \infty} F_{\epsilon_i}(u_{\epsilon_i}), \tag{2.8}$$

$W(x, u_0(x)) = 0$ a.e. $x \in \Omega$, and $E_0 = \{x \in \Omega \,|\, u_0(x) = \alpha\}$ has finite perimeter. Assume that u_0 is not a solution of (P_0), i.e., there exists $\tilde{u}_0 \in BV(\Omega : \mathbf{R}^n)$ such that

$$F_0(\tilde{u}_0) < F_0(u_0).$$

From Proposition B, there exists a sequence $\{\tilde{u}_\epsilon\}_{\epsilon>0}$ in $W^{1,2}(\Omega : \mathbf{R}^n)$ such that

$$\limsup_{\epsilon\to 0} F_\epsilon(\tilde{u}_\epsilon) \le F_0(\tilde{u}_0) < F_0(u_0) \le \liminf_{i\to\infty} F_{\epsilon_i}(u_{\epsilon_i}). \tag{2.9}$$

Then (2.9) contradicts to the minimality of u_{ϵ_i}. Therefore, u_0 is a solution of (P_0). Furthermore, from Proposition B and the minimality of u_{ϵ_i}, we have

$$\lim_{i\to\infty} F_{\epsilon_i}(u_{\epsilon_i}) = F_0(u_0).$$

From the definition of the functional F_0, we obtain

$$\begin{aligned}\lim_{i\to\infty} F_{\epsilon_i}(u_{\epsilon_i}) = 2\int_{\Omega\cap\partial^* E_0} d(x,\alpha,\beta)d\mathcal{H}_{N-1} &+ 2\int_{\partial\Omega\cap\partial^* E_0} d(x,\alpha,g(x))d\mathcal{H}_{N-1} \\ &+ 2\int_{\partial\Omega\setminus\partial^* E_0} d(x,\beta,g(x))d\mathcal{H}_{N-1},\end{aligned}$$

and the proof of Theorem 1 is completed. □

It remains to prove Proposition A and B, which we will undertake in §§3 and 4, respectively.

3. Proof of Proposition A. From (2.5), it follows that there exists a positive sequence $\{\epsilon_i\}_{i=1}^\infty$ such that $\lim_{i\to\infty}\epsilon_i = 0$ and $\lim_{i\to\infty} F_{\epsilon_i}(v_{\epsilon_i}) = \liminf_{\epsilon\to 0_+} F_\epsilon(v_\epsilon) < \infty$. Furthermore, by taking a subsequence if necessary, we can assume that $\{v_{\epsilon_i}\}_{i=1}^\infty$ converges to v_0 a.e. $x \in \Omega$ as $i \to +\infty$. Then by Fatou's lemma, we have

$$\begin{aligned}0 \le \int_\Omega W(x, v_0(x))\,dx &\le \liminf_{i\to\infty} \int_\Omega W(x, v_{\epsilon_i}(x))\,dx \\ &\le \liminf_{i\to\infty} \int_\Omega [{\epsilon_i}^2|\nabla v_{\epsilon_i}| + W(x, v_{\epsilon_i}(x))]\,dx \\ &\le \liminf_{i\to\infty} \epsilon_i F_{\epsilon_i}(v_{\epsilon_i}) = 0.\end{aligned} \tag{3.1}$$

Therefore, we obtain $W(x, v_0(x)) = 0$ for a.e. x in Ω.

In order to prove the remaining part of Proposition A, we can assume without loss of generality that $v_{\epsilon_i}(x) \in [K_1, K_2]^n$ for all $x \in \Omega$. In fact, let $v_{\epsilon_i}^\#$ be a function which is obtained by truncation of each scalar component of v_{ϵ_i} by K_1 and K_2 given in (1.1), i.e., for $v_{\epsilon_i} = (v_{\epsilon_i}^1, v_{\epsilon_i}^2, \dots, v_{\epsilon_i}^n)$,

$$v_{\epsilon_i}^{\#^j} = \max\{K_1, \min\{v_{\epsilon_i}^j, K_2\}\}, \quad j = 1, 2, \dots, n.$$

Then from (1.1) and (1.2), we have the following relations:

$$\lim_{i\to+\infty} v_{\epsilon_i}^\# = v_0 \quad \text{in} \quad L^1(\Omega : \mathbf{R}^n), \qquad v_{\epsilon_i}^\# = g \quad \text{on} \quad \partial\Omega,$$

$$\int_\Omega \left[\epsilon_i|\nabla v_{\epsilon_i}^\#|^2 + \frac{1}{\epsilon_i} W(x, v_{\epsilon_i}^\#)\right] dx \le \int_\Omega \left[\epsilon_i|\nabla v_{\epsilon_i}|^2 + \frac{1}{\epsilon_i} W(x, v_{\epsilon_i})\right] dx.$$

For the purpose of proving that $v_0 \in BV(\Omega : \mathbf{R}^n)$, we will introduce a function $d_\infty : \mathbf{R}^{2n} \to \mathbf{R}^+$ as follows:

$$(3.2)\qquad d_\infty(a,b) = \inf\left\{ \int_0^1 W_\infty^{1/2}(\gamma(s))\,|\,\dot\gamma(s)\,|\,ds \;\middle|\; \gamma \in C^1([0,1] : \mathbf{R}^n),\ \gamma(0) = a, \gamma(1) = b \right\}$$

for all $a, b \in \mathbf{R}^n$, where $W_\infty(\cdot) = \inf_{x\in\overline{\Omega}} W(x,\cdot)$. Here we set $\psi_\infty(\xi) = d_\infty(a,\xi)$ for $\xi \in \mathbf{R}^n$ and for fixed $a \in \mathbf{R}^n$, and we can easily see that $\psi_\infty(\xi)$ is continuous. Then from Proposition 2–1 in [3], we have $\psi_\infty(v_{\epsilon_i}) \in W^{1,1}(\Omega)$ and

$$\begin{aligned}\liminf_{i\to\infty} \int_\Omega |\nabla(\psi_\infty(v_{\epsilon_i}))|dx &\le \liminf_{i\to\infty} \int_\Omega W_\infty^{1/2}(v_{\epsilon_i})|\nabla v_{\epsilon_i}|\,dx \\ &\le \lim_{i\to\infty} \frac{1}{2} \int_\Omega [\epsilon_i|\nabla v_{\epsilon_i}|^2 + \frac{1}{\epsilon_i} W_\infty(v_{\epsilon_i})]\,dx \\ &\le \lim_{i\to\infty} \frac{1}{2} F_{\epsilon_i}(v_{\epsilon_i}) < \infty.\end{aligned}$$

On the other hand, from the equiboundedness of $\{v_{\epsilon_i}\}_{i=1}^\infty$ in $L^\infty(\Omega : \mathbf{R}^n)$, we can see that $\psi_\infty(v_{\epsilon_i}) \to \psi_\infty(v_0)$ in $L^1(\Omega)$. Thus we have from (2.1)

$$\int_\Omega |\nabla\psi_\infty(v_0)| \le \liminf_{i\to\infty} \int_\Omega |\nabla(\psi_\infty(v_{\epsilon_i}))|dx < \infty,$$

and so $\psi_\infty(v_0) \in BV(\Omega)$. Therefore, by (3.1), we have $\psi_\infty(v_0) \in \{\psi_\infty(\alpha), \psi_\infty(\beta)\}$ for almost all x in Ω, and we deduce that $v_0 \in BV(\Omega : \mathbf{R}^n)$.

Next we will prove inequality (2.6). In order to give the estimates of $|\nabla v_{\epsilon_i}|\cdot W^{1/2}(x, v_{\epsilon_i})$ on the boundary $\partial\Omega$ of Ω, we need to extend the domain Ω.

Let Ω' be any bounded open set in $\mathbf{R}^N$ with $\overline{\Omega} \subset \Omega'$. From the regularity of $\partial\Omega$, without loss of generality, we can extend the function $W(x,v)$ to the domain $\Omega' \times \mathbf{R}^n$. Then from the uniform continuity of $W(x,v)$ on $\overline{\Omega'} \times [K_1, K_2]^n$, for any $\epsilon > 0$, there exists a constant $\delta > 0$ such that

$$(3.3)\qquad |W^{1/2}(x,v) - W^{1/2}(y,v)| \le \epsilon[1 + W^{1/2}(x,v)]$$

for all $x, y \in \Omega'$ with $|x-y| \le \delta$ and for all $v \in [K_1, K_2]^n$. Furthermore, from the regularity of $\partial\Omega$ and g, there exists a function G in $W^{1,\infty}(\Omega' : \mathbf{R}^n)$ whose trace on $\partial\Omega$ equals to g. Here for any $v \in L^1(\Omega : \mathbf{R}^n)$, we set

$$(3.4)\qquad \tilde v = \begin{cases} v & \text{if } x \in \Omega, \\ G & \text{if } x \in \Omega' \setminus \overline{\Omega}. \end{cases}$$

In particular, $\tilde v_{\epsilon_i} \in W^{1,2}(\Omega' : \mathbf{R}^n)$ and we have

$$\sup_i \int_{\Omega'} |\nabla \tilde v_{\epsilon_i}| W^{1/2}(x, \tilde v_{\epsilon_i})dx < \infty.$$

Therefore, from the weak compactness for measures (see [7, Thm. 2, p. 55]) and the definition of $\tilde v_{\epsilon_i}$, there exists a nonnegative Radon measure μ on $\overline{\Omega}$ and a subsequence $\{\epsilon_{i'}\}_{i'=1}^\infty$ of $\{\epsilon_i\}_{i=1}^\infty$ such that for any $\varphi \in C_0^\infty(\Omega')$,

$$(3.5)\qquad \int_{\Omega'} \varphi|\nabla\tilde v_{\epsilon_{i'}}|W^{1/2}(x, \tilde v_{\epsilon_{i'}})\,dx \to \int_{\overline{\Omega}} \varphi d\mu + \int_{\Omega'\setminus\overline{\Omega}} \varphi|\nabla G|W^{1/2}(x,G)\,dx$$

as $i' \to \infty$. In what follows, we omit the prime of $\epsilon_{i'}$ for simplicity.

From the Lebesgue decomposition theorem (see [7, Thm. 3, p. 42]), we can write μ as a sum of the singular nonnegative measures $\mu = \mu_1 \mathcal{H}_{N-1} \lfloor ((\Omega \cap \partial^* E) \cup \partial\Omega) + \mu_2$. Then we can prove the following lemma. The proof of the following lemma is given by modifying the arguments of [4].

LEMMA 3.1. *The singular measure* μ_1 *has the following properties:*

$$\mu_1(x) \geq d(x, \alpha, \beta) \qquad \mathcal{H}_{N-1} - \text{ a.e. } x \text{ in } \Omega \cap \partial^* E; \tag{3.6}$$

$$\mu_1(x) \geq d(x, \alpha, g(x)) \qquad \mathcal{H}_{N-1} - \text{ a.e. } x \text{ in } \partial\Omega \cap \partial^* E; \tag{3.7}$$

$$\mu_1(x) \geq d(x, \beta, g(x)) \qquad \mathcal{H}_{N-1} - \text{ a.e. } x \text{ in } \partial\Omega \setminus \partial^* E. \tag{3.8}$$

Proof. We begin by proving the inequality (3.6). Let x_0 be any point in $\Omega \cap \partial^* E$. From the Radon–Nikodym theorem, we have $\mu_1(x) < \infty$ for $\mathcal{H}_{N-1}-$ a.e. x in $(\Omega \cap \partial^* E) \cup \partial\Omega$, and so we can assume $\mu_1(x_0) < \infty$. Then we have

$$+\infty > \mu_1(x_0) = \limsup_{\delta \to 0} \mu(x_0 + \delta B_1)/\mathcal{H}_{N-1}(\{x_0 + \delta B_1\} \cap \{\Omega \cap \partial^* E\}) \tag{3.9}$$

and for any $\varphi \in C_0^\infty(B_1)$ with $0 \leq \varphi \leq 1$, we get

$$\begin{aligned} \mu(x_0 + \delta B_1) &\geq \int_{x_0 + \delta B_1} \varphi\left(\frac{y - x_0}{\delta}\right) d\mu(y) \\ &\geq \lim_{i \to 0} \int_{x_0 + \delta B_1} \varphi\left(\frac{y - x_0}{\delta}\right) |\nabla v_{\epsilon_i}| W^{1/2}(y, v_{\epsilon_i}(y)) dy \\ &\geq \liminf_{i \to 0} I_1^i + \liminf_{i \to 0} I_2^i, \end{aligned} \tag{3.10}$$

where

$$I_1^i = \int_{x_0 + \delta B_1} \varphi\left(\frac{y - x_0}{\delta}\right) |\nabla v_{\epsilon_i}| W^{1/2}(x_0, v_{\epsilon_i}(y)) dy$$

and

$$I_2^i = \int_{x_0 + \delta B_1} \varphi\left(\frac{y - x_0}{\delta}\right) |\nabla v_{\epsilon_i}| [W^{1/2}(y, v_{\epsilon_i}(y)) - W^{1/2}(x_0, v_{\epsilon_i}(y))] dy.$$

From Proposition 2–1 in [3], we have

$$\begin{aligned} &\int_{x_0 + \delta B_1} \varphi\left(\frac{y - x_0}{\delta}\right) |\nabla v_{\epsilon_i}|\, W^{1/2}(x_0, v_{\epsilon_i}(y)) dy \\ &\qquad \geq \int_{x_0 + \delta B_1} \varphi\left(\frac{y - x_0}{\delta}\right) |\nabla_y d(x_0, \alpha, v_{\epsilon_i}(y))|, \end{aligned}$$

and so from (2.1), we obtain

$$\begin{aligned} \liminf_{i \to \infty} I_1^i &\geq \int_{x_0 + \delta B_1} \varphi\left(\frac{y - x_0}{\delta}\right) |\nabla_y d(x_0, \alpha, v_0(y))| \\ &\geq \int_{(x_0 + \delta B_1) \cap \partial^* E} \varphi\left(\frac{y - x_0}{\delta}\right) d(x_0, \alpha, \beta) d\mathcal{H}_{N-1}(y). \end{aligned} \tag{3.11}$$

On the other hand, from (1.3), there exists a nonnegative function $\epsilon(\delta)$ with $\lim_{\delta\to 0}\epsilon(\delta) = 0$ such that

$$(3.12)\qquad |I_2^i| \leq \epsilon(\delta)\int_{x_0+\delta B_1}\varphi\Big(\frac{y-x_0}{\delta}\Big)|\nabla v_{\epsilon_i}|W^{1/2}(y,v_\epsilon(y))dy \leq \epsilon(\delta)\mu(x_0+\delta B_1).$$

In what follows, we denote $\epsilon(\delta)$ by $0_\delta(1)$. Therefore, from (3.10)–(3.12), we have

$$\mu(x_0+\delta B_1) \geq d(x_0,\alpha,\beta)\mathcal{H}_{N-1}(\{x_0+\delta B_1\}\cap\partial^* E) + 0_\delta(1)\mu(x_0+\delta B_1),$$

and so from (3.9), we obtain

$$\mu_1(x_0) \geq d(x_0,\alpha,\beta)$$

and we conclude inequality (3.6).

Next we will prove the equality (3.7). Let x_0 be a point in $\partial^* E\cap\partial\Omega$. Without loss of generality, we can assume $\mu_1(x_0) < +\infty$, and we have

$$\begin{aligned}(3.13)\qquad \mu(x_0+\delta B_1) &\geq \lim_{i\to\infty}\int_{x_0+\delta B_1}\varphi\Big(\frac{x_0-y}{\delta}\Big)|\nabla\tilde{v}_{\epsilon_i}|W^{1/2}(y,\tilde{v}_{\epsilon_i})dy \\ &\quad -\int_{(x_0+\delta B_1)\cap(\Omega'\setminus\overline{\Omega})}\varphi\Big(\frac{x_0-y}{\delta}\Big)|\nabla G|W^{1/2}(x,G)dx \\ &= \lim_{i\to\infty}J_1^i + \lim_{i\to\infty}J_2^i + 0(\delta^N),\end{aligned}$$

where

$$J_1^i \equiv \int_{x_0+\delta B_1}\varphi\Big(\frac{x_0-y}{\delta}\Big)|\nabla\tilde{v}_{\epsilon_i}|W^{1/2}(x_0,\tilde{v}_{\epsilon_i})dy$$

and

$$J_2^i \equiv \int_{x_0+\delta B_1}\varphi\Big(\frac{x_0-y}{\delta}\Big)|\nabla\tilde{v}_{\epsilon_i}|[W^{1/2}(y,\tilde{v}_{\epsilon_i}) - W^{1/2}(x_0,\tilde{v}_{\epsilon_i})]dy.$$

By means of an argument similar to the one we used to prove (3.6), we have

$$\begin{aligned}(3.14)\qquad \lim_{i\to\infty}J_1^i &\geq \lim_{i\to\infty}\int_{x_0+\delta B_1}\varphi\Big(\frac{x_0-y}{\delta}\Big)|\nabla d(x_0,\alpha,\tilde{v}_{\epsilon_i})| \\ &\geq \int_{x_0+\delta B_1}\varphi\Big(\frac{x_0-y}{\delta}\Big)|\nabla d(x_0,\alpha,\tilde{v}_0)| \\ &\geq \int_{\{x_0+\delta B_1\}\cap(\partial\Omega\cap\partial^* E)}\varphi\Big(\frac{x_0-y}{\delta}\Big)d(x_0,\alpha,g(y))d\mathcal{H}_{N-1}(y)\end{aligned}$$

and from (1.3) and (3.3),

$$\begin{aligned}(3.15)\qquad |J_2^i| &\leq 0_\delta(1)\int_{x_0+\delta B_1}\varphi\Big(\frac{x_0-y}{\delta}\Big)|\nabla\tilde{v}_{\epsilon_i}|W^{1/2}(y,\tilde{v}_{\epsilon_i})dy \\ &\qquad + 0_\delta(1)\int_{\{x_0+\delta B_1\}\cap(\Omega'\setminus\Omega)}\varphi\Big(\frac{x_0-y}{\delta}\Big)dy \\ &\leq 0_\delta(1)[\mu(x_0+\delta B_1) + \delta^N].\end{aligned}$$

Therefore, from [21, Lem. 5.5.4, p. 236] and (3.13)–(3.15), we obtain

$$\mu_1(x_0) \geq d(x_0, \alpha, g(x_0)),$$

and so we prove inequality (3.7). Furthermore, we can prove (3.8) in just the same way as (3.9).

For any $\varphi \in C_0^\infty(\Omega')$, $0 \leq \varphi \leq 1$, we have from (3.4)

$$\begin{aligned}(3.16)\qquad \lim_{i\to\infty} \int_\Omega \left[\epsilon_i |\nabla v_{\epsilon_i}|^2 + \frac{1}{\epsilon_i} W(x, v_{\epsilon_i})\right] dx &\geq 2 \lim_{i\to\infty} \int_\Omega |\nabla v_{\epsilon_i}| W^{1/2}(x, v_{\epsilon_i}) dx \\ &\geq 2 \lim_{i\to\infty} \int_{\Omega'} \varphi |\nabla \tilde{v}_{\epsilon_i}| W^{1/2}(x, \tilde{v}_{\epsilon_i}) dx - 2 \int_{\Omega'\setminus\overline{\Omega}} |\nabla G| W^{1/2}(x, G) dx \\ &\geq 2 \int_{\overline{\Omega}} \varphi d\mu - 2 \int_{\Omega'\setminus\overline{\Omega}} (1-\varphi) |\nabla G| W^{1/2}(x, G) dx.\end{aligned}$$

Consider an increasing sequence φ_j, $j = 1, 2, \ldots$, such that $\lim_{j\to\infty} \varphi_j(x) = 1$ for all $x \in \Omega'$, and taking the limit of (3.16) as $j \to \infty$, we have

$$\liminf_{i\to\infty} F_{\epsilon_i}(v_{\epsilon_i}) \geq 2 \int_{\overline{\Omega}} d\mu.$$

Therefore, from Lemma 3.1, we obtain

$$\begin{aligned}\liminf_{i\to\infty} F_{\epsilon_i}(v_{\epsilon_i}) \geq 2 &\int_{\Omega\cap\partial^* E} d(x, \alpha, \beta) d\mathcal{H}_{N-1} \\ &+ 2 \int_{\partial\Omega\cap\partial^* E} d(x, \alpha, g(x)) d\mathcal{H}_{N-1} + 2 \int_{\partial\Omega\setminus\partial^* E} d(x, \beta, g(x)) d\mathcal{H}_{N-1},\end{aligned}$$

and the proof of Proposition A is completed. □

4. Proof of Proposition B. First we prove Proposition B for the special case where $w_0 \equiv \alpha$ in Ω. We divide the proof into three steps. The arguments used to prove the second and third steps are completely different from and more complicated than the ones used in the scalar case. In the second step, we construct the essential parts of w_ϵ, whose energy tends to $F_0(w_0)$ as $\epsilon \to 0$. In the third step, we complete the construction of w_ϵ, matching the trace on $\partial\Omega$ to coincide with the given function g on $\partial\Omega$.

Before starting a proof of Proposition B, we present the following two lemmas. The first lemma is obtained easily by means of the inverse-mapping theorem.

LEMMA 4.1. *Let Ω be a bounded domain with C^2-smooth boundary $\partial\Omega$. For $x \in \partial\Omega$, let $\nu(x)$ be a inner normal vector to $\partial\Omega$ at x. Define a mapping $\pi : \partial\Omega \times [0, \infty) \to \mathbf{R}^N$ by*

$$(4.1)\qquad \pi(x, t) = \pi_t(x) = x + t\nu(x).$$

Then there exists a constant s_0 such that $\pi(\partial\Omega \times (0, s_0])$ is contained in Ω and the C^1-smooth inverse mapping π^{-1} of π exists on $\pi(\partial\Omega \times [0, s_0])$.

LEMMA 4.2. (*See* [13] *and* [20].) *Let Ω be an open bounded subset of $\mathbf{R}^N$ with Lipschitz-continuous boundary. Let A be an open subset of $\mathbf{R}^N$ with C^2 and with a compact, nonempty boundary such that $\mathcal{H}_{N-1}(\partial A \cap \partial\Omega) = 0$. Define a distance function to ∂A, $d_{\partial A} : \Omega \to \mathbf{R}$, by*

$$d_{\partial A}(x) = \text{ dist } (x, A).$$

Then, for some $s_1 > 0$, $d_{\partial A}$ *is a* C^2 *function in* $\{0 < d_{\partial A}(x) < s_1\}$ *with*

$$|\nabla d_{\partial A}| = 1. \tag{4.2}$$

Furthermore, $\lim_{s\to 0}\mathcal{H}_{N-1}(\{d_{\partial A}(x) = s\}) = \mathcal{H}_{N-1}(\partial A \cap \Omega)$ *and*

$$|\{x \mid |d_{\partial A}(x)| < s\}| = O(s). \tag{4.3}$$

Here $d_{\partial\Omega}(x)$ denotes the distance function $\operatorname{dist}(x,\partial\Omega)$. From Lemma 4.2, we see that $d_{\partial\Omega}$ is a C^2 function in a neighborhood of $\partial\Omega$. We set $s^* = \min\{s_0, s_1\}$. Furthermore, for $x \in \partial\Omega$, $\eta > 0$, and sufficiently small δ with $0 < \delta < s^*$, we set $\partial\Omega_\eta(x) = \partial\Omega \cap (x + \eta Q_{\nu(x)})$ and $\Omega^\delta_\eta(x) = \cup_{\delta<t<s^*}\pi_t(\partial\Omega_\eta(x))$.

Now we begin the proof of Proposition B in the special case where $w_0 \equiv \alpha$.

Step 1. Let x_0 be any point in $\partial\Omega$. In this step, for any sufficiently small $\eta > 0$, we will construct a family $\{w^\delta_\epsilon\}_{\epsilon,\delta>0} \subset W^{1,2}(\Omega^\delta_\eta(x_0) : \mathbf{R}^n)$ such that

$$\limsup_{\epsilon,\delta\to 0} \int_{\Omega^\delta_\eta} \left[\epsilon|\nabla w^\delta_\epsilon|^2 + \frac{1}{\epsilon}W(x_0, w^\delta_\epsilon)\right] dx \le 2d(x_0, \alpha, g(x_0))\mathcal{H}_{N-1}(\partial\Omega_\eta(x_0)). \tag{4.4}$$

In this step, for simplicity, we set $\Omega^\delta_\eta = \Omega^\delta_\eta(x_0)$.

In order to construct $\{w^\delta_\epsilon\}_{\epsilon,\delta>0}$, we fix $\epsilon, \delta > 0$, and we consider the following ordinary differential equation:

$$\begin{cases} \dfrac{d}{dt} y_\epsilon(t) = \dfrac{[\epsilon^{1/2} + W(x_0, \gamma(y_\epsilon(t)))]^{1/2}}{\epsilon|\,\dot\gamma(y_\epsilon(t))\,|}, \\ y_\epsilon(\delta) = 0. \end{cases} \tag{4.5}$$

Here we denote $d\gamma(t)/dt$ by $\dot\gamma$, and we assume that $\gamma \in C^1([0,1] : [K_1, K_2]^n)$, $\gamma(0) = \alpha$, $\gamma(1) = g(x_0)$. We set

$$\psi_\epsilon(t) = \int_0^t \frac{\epsilon|\dot\gamma(t)|}{[\epsilon^{1/2} + W(x_0, \gamma(t))]^{1/2}} dt$$

for $t \in (0,1)$ and set $\tau_\epsilon = \psi_\epsilon(1)$. Then $\psi_\epsilon(t)$ is a monotone increasing function and

$$\tau_\epsilon = \psi_\epsilon(1) \le \epsilon^{3/4} \cdot \text{ length of } \gamma. \tag{4.6}$$

We set $\tilde y_\epsilon(t) = \psi_\epsilon^{-1}(t - \delta)$, and we can see that $\tilde y_\epsilon(t)$ satisfies (4.5) in $[\delta, \delta + \tau_\epsilon]$; we define $y_\epsilon(t)$ by

$$y_\epsilon(t) \equiv \max\{0, \min\{1, \tilde y_\epsilon(t)\}\} = \begin{cases} 0, & t \le \delta, \\ \tilde y_\epsilon(t), & \delta \le t \le \delta + \tau_\epsilon, \\ 1, & t \ge \delta + \tau_\epsilon. \end{cases} \tag{4.7}$$

We partition Ω^δ_η into three subdomains $\Omega^\delta_{\eta,i}$, $i = 1, 2, 3$ as follows:

$$\begin{aligned} \Omega^\delta_{\eta,1} &\equiv \{x \in \Omega^\delta_\eta : d_{\partial\Omega}(x) < \delta + \tau_\epsilon,\ d_S(x) \le \eta\tau_\epsilon\}, \\ \Omega^\delta_{\eta,2} &\equiv \{x \in \Omega^\delta_\eta : d_{\partial\Omega}(x) < \delta + \tau_\epsilon,\ d_S(x) \ge \eta\tau_\epsilon\}, \\ \Omega^\delta_{\eta,3} &\equiv \{x \in \Omega^\delta_\eta : d_{\partial\Omega}(x) \ge \delta + \tau_\epsilon\}, \end{aligned} \tag{4.8}$$

where $d_S(x)$ is a distance function to $\cup_{\delta<t<s^*}\pi_t[\partial\Omega\cap(x_0+\eta\partial Q_{\nu(x_0)})]$. Here we define $w_\epsilon(x)$ on $\cup_{i=2,3}\Omega^\delta_{\eta,i}$ as follows:

$$w_\epsilon(x) = \begin{cases} \gamma(y_\epsilon(d_{\partial\Omega}(x))), & \text{if } x \in \Omega^\delta_{\eta,2}, \\ \alpha, & \text{if } x \in \Omega^\delta_{\eta,3}. \end{cases} \tag{4.9}$$

We extend w_ϵ to $\Omega^\delta_{\eta,1}$ in such a way that for any $x \in \Omega^\delta_\eta$ with $d_S(x) = 0$ or $d_{\partial\Omega}(x) = \delta + \tau_\epsilon$, $w_\epsilon(x) = \alpha$ and

$$|\nabla w_\epsilon| \leq 2/(K_2 - K_1)\eta\tau_\epsilon + C/\epsilon \leq C(\eta\tau_\epsilon)^{-1} + C\epsilon^{-1}.$$

For sufficiently small $\epsilon > 0$, we have the length of $\gamma < \epsilon^{-1/8}$ and $\tau_\epsilon \leq \epsilon^{5/8}$. Therefore, we obtain

$$\begin{aligned} \int_{\Omega^\delta_{\eta,1}} \left[\epsilon|\nabla w_\epsilon|^2 + \frac{1}{\epsilon}W(x_0, w_\epsilon)\right] dx &\leq C[\epsilon/\eta^2\tau_\epsilon^2 + 1/\epsilon]\tau_\epsilon^N \mathcal{H}_{N-1}(\partial\Omega_\eta) \\ &\leq C(\epsilon/\eta^2 + \epsilon^{1/4})\tau_\epsilon^{N-2}\mathcal{H}_{N-1}(\partial\Omega_\eta). \end{aligned} \tag{4.10}$$

We note that constants C are independent of ϵ and η. On the other hand, for sufficiently small $\delta > 0$ and $\epsilon > 0$, we have $\delta + \tau_\epsilon < s^* \equiv \min\{s_0, s_1\}$ and we obtain from Lemma 4.2 and (4.9)

$$\begin{aligned} &\int_{\cup_{i=2,3}\Omega^\delta_{\eta,i}} \left[\epsilon|\nabla w_\epsilon|^2 + \frac{1}{\epsilon}W(x_0, w_\epsilon)\right] dx \\ &\qquad \leq \int_{\Omega^\delta_{\eta,2}} \frac{2}{\epsilon}[\epsilon^{1/2} + W(x_0, \gamma(y_\epsilon(d_{\partial\Omega}(x))))]|\nabla d_{\partial\Omega}(x)|dx, \end{aligned}$$

and from (2.2), we get

$$\begin{aligned} &\int_{\cup_{i=2,3}\Omega^\delta_{\eta,i}} \left[\epsilon|\nabla w_\epsilon|^2 + \frac{1}{\epsilon}W(x_0, w_\epsilon)\right] dx \\ &\qquad \leq 2\int_\delta^{\tau_\epsilon+\delta} dt \int_{\Omega^\delta_\eta\cap\{d_{\partial\Omega}(x)=t\}} \epsilon^{-1}[\epsilon^{1/2} + W(x_0, \gamma(y_\epsilon(t)))]d\mathcal{H}_{N-1} \\ &\qquad \leq 2\kappa^\delta_\epsilon \int_\delta^{\tau_\epsilon+\delta} \epsilon^{-1}(\epsilon^{1/2} + W(x_0, \gamma(y_\epsilon(t))))dt, \end{aligned}$$

where $\kappa^\delta_\epsilon = \sup_{\delta\leq d_S(x)\leq\delta+\epsilon}(\Omega^\delta_\eta \cap \pi_t(\partial\Omega))$. Then from (4.5), we obtain

$$\int_{\cup_{i=1,2}\Omega^\delta_{\eta,i}} \left[\epsilon|\nabla w_\epsilon|^2 + \frac{1}{\epsilon}W(x_0, w_\epsilon)\right] dx \leq 2\kappa^\delta_\epsilon \int_0^1 [\epsilon^{1/2} + W(x_0, \gamma(t))]^{1/2}|\dot\gamma(t)|dt. \tag{4.11}$$

From the regularity of $\partial\Omega$ and the definition of $\Omega^0_\eta(x_0)$, there exists a constant η_0 independent of x_0 (dependent only on $\partial\Omega$) such that for any $0 < \eta < \eta_0$, we have $\mathcal{H}_{N-1}(\partial\Omega^0_\eta(x_0)\cap\partial\Omega) = 0$. So from Lemma 4.2, we have $\lim_{\epsilon,\delta\to 0}\kappa^\delta_\epsilon = \mathcal{H}_{N-1}(\partial\Omega_\eta(x_0))$ for any $\eta \in (0, \eta_0)$. Here we set $w^{\delta,\gamma}_\epsilon = w_\epsilon$. Therefore, from (4.10) and (4.11), for any $\eta \in (0, \eta_0)$ we obtain

$$\begin{aligned} &\int_{\Omega^\delta_\eta(x_0)} \left[\epsilon|\nabla w^{\delta,\gamma}_\epsilon|^2 + \frac{1}{\epsilon}W(x_0, w^{\delta,\gamma}_\epsilon)\right] dx \\ &\qquad \leq 2\mathcal{H}_{N-1}(\partial\Omega_\eta)\int_0^1 W^{1/2}(x_0, \gamma(t))|\,\dot\gamma(t)\,|dt \\ &\qquad\qquad + \mathcal{H}_{N-1}(\partial\Omega_\eta)[0(\epsilon/\eta^2) + 0(\epsilon^{1/4}) + 0_{\sqrt{\epsilon^2+\delta^2}}(1)]. \end{aligned} \tag{4.12}$$

Since for any $\epsilon > 0$ there exists a sequence of C^1-curves $\{\gamma_i\}_{i=1}^{\infty}$ such that the length of $\gamma_i \leq \epsilon^{-1/8}$ and

$$\lim_{i\to\infty}\int_0^1 W^{1/2}(x_0,\gamma_i(t))|\,\dot{\gamma}_i(t)\,|\,dt = d(x_0,\alpha,g(x_0)),$$

by the diagonal argument and (4.12), we can construct a sequence $\{w_\epsilon^\delta\}_{\epsilon,\delta>0}$ satisfying (4.4). Therefore, the aim of the first step is completed. □

Step 2. Let Ω_δ be the domain $\{x \in \Omega : \delta < d_{\partial\Omega}(x) < s^*\} = \cup_{\delta<t<s^*}\pi_t(\partial\Omega)$. In this step, we construct a sequence $\{w_\epsilon^\delta\}_{\epsilon,\delta>0}$ in $W^{1,2}(\Omega_\delta, \mathbf{R}^n)$ such that

$$\limsup_{\delta,\epsilon\to\infty}\int_{\Omega_\delta}\left[\epsilon|\nabla w_\epsilon^\delta|^2 + \frac{1}{\epsilon}W(x,w_\epsilon^\delta)\right]dx \leq 2\int_{\partial\Omega} d(x,a,g(x))d\mathcal{H}_{N-1}. \tag{4.13}$$

In order to construct this sequence $\{w_\epsilon^\delta\}_{\epsilon,\delta>0}$, we partition $\partial\Omega$ into subdomains. In view of the regularity of $\partial\Omega$, for sufficiently small $\eta > 0$, there exist p points $\{x_i\}_{i=1}^p \subset \partial\Omega$ and a subset ω_η of $\partial\Omega$ such that

$$\partial\Omega\setminus\cup_{1\leq i\leq p}\partial\Omega_\eta(x_i)\subset\omega_\eta,\quad \partial\Omega_\eta(x_i)\cap\partial\Omega_\eta(x_j)=\emptyset,\quad i\neq j,\ i,j=1,2,\ldots,p \tag{4.14}$$

and $\lim_{\eta\to 0}\mathcal{H}_{N-1}(\omega_\eta) = 0$. Here we note that p depends on η and $\lim_{\eta\to 0}p(\eta) = \infty$.

Fix $\eta, \delta, \epsilon > 0$. For any $i \in \{1, 2, \ldots, p\}$, from (4.12) we can construct functions $w_\epsilon^{i,\delta,\eta} \in W^{1,2}(\Omega_\eta^\delta(x_i))$ such that

$$\begin{aligned}\int_{\Omega_\eta^\delta(x_i)}&\left[\epsilon|\nabla w_\epsilon^i|^2 + \frac{1}{\epsilon}W(x_i,w_\epsilon^i)\right]dx \\ &\leq 2\mathcal{H}_{N-1}(\partial\Omega_\eta(x_i))d(x_i,\alpha,g(x_i)) \\ &\qquad + \mathcal{H}_{N-1}(\partial\Omega_\eta(x_i))[0(\epsilon/\eta^2) + 0(\epsilon^{1/4}) + 0_{\sqrt{\epsilon^2+\delta^2}}(1)].\end{aligned} \tag{4.15}$$

Then we define $w_\epsilon^{\delta,\eta} \in W^{1,2}(\Omega_\delta : \mathbf{R}^n)$ as follows:

$$w_\epsilon^{\delta,\eta} = \begin{cases} w_\epsilon^{i,\delta,\eta} & \text{if } x \in \Omega_\eta^\delta(x_i), \\ \alpha & \text{otherwise.}\end{cases}$$

By the argument of Step 1, we can easily see that $w_\epsilon^{\delta,\eta} \in W^{1,2}(\Omega_\delta : \mathbf{R}^n)$. Then we have

$$\begin{aligned}\int_{\Omega_\delta}&\left[\epsilon|\nabla w_\epsilon^{\delta,\eta}|^2 + \frac{1}{\epsilon}W(x,w_\epsilon^{\delta,\eta})\right]dx \\ &= \sum_{i=1}^p\int_{\Omega_\eta^\delta(x_i)}\left[\epsilon|\nabla w_\epsilon^{i,\delta,\eta}|^2 + \frac{1}{\epsilon}W(x,w_\epsilon^{i,\delta,\eta})\right]dx.\end{aligned} \tag{4.16}$$

On the other hand, we have

$$\begin{aligned}&\int_{\Omega_\eta^\delta(x_i)}\left[\epsilon|\nabla w_\epsilon^i|^2 + \frac{1}{\epsilon}W(x,w_\epsilon^i)\right]dx \\ &= \int_{\Omega_\eta^\delta(x_i)}\left[\epsilon|\nabla w_\epsilon^i|^2 + \frac{1}{\epsilon}W(x_i,w_\epsilon^i)\right]dx + \int_{\Omega_\eta^\delta(x_i)}\frac{1}{\epsilon}\left[W(x,w_\epsilon^i) - W(x_i,w_\epsilon^i)\right]dx \\ &\equiv I_1^i + I_2^i\end{aligned}$$

(for simplicity we omit the index δ,η of $w_\epsilon^{i,\delta,\eta}$). From (4.15), we obtain

$$\sum_{i=1}^{p(\eta)} I_1^i \leq 2\sum_{i=1}^{p(\eta)}[d(x_i,\alpha,g(x_i))\mathcal{H}_{N-1}(\partial\Omega_\eta(x_i))] \\ + 0(\epsilon/\eta^2) + 0(\epsilon^{1/4}) + 0_{\sqrt{\epsilon^2+\delta^2}}(1), \tag{4.17}$$

and from (1.3) and (4.15),

$$\sum_{i=1}^{p(\eta)} |I_2^i| \leq \sum_{i=1}^{p(\eta)} \int_{\Omega_\eta^\delta(x_i)} 0_{|x-x_i|}(1)\frac{1}{\epsilon}W(x_i,w_\epsilon^i)dx \leq 0_\eta(1)\sum_{i=1}^{p(\eta)} I_1^i.$$

We set $\eta^2 = \epsilon^{3/4}$. This inequality, together with (4.16) and (4.17), yields

$$\limsup_{\delta,\epsilon\to 0} \int_{\Omega_\delta} \left[\epsilon|\nabla w_\epsilon^{\delta,\eta(\epsilon)}|^2 + \frac{1}{\epsilon}W(x,w_\epsilon^{\delta,\eta(\epsilon)})\right] dx \\ \leq \limsup_{\epsilon\to 0} 2\sum_{i=1}^{p(\eta)} d(x_i,\alpha,g(x_i))\mathcal{H}_{N-1}(\partial\Omega_\eta(x_i)). \tag{4.18}$$

On the other hand, for any $x \in \partial\Omega_\eta(x_i)$, we have

$$|d(x_i,\alpha,g(x_i)) - d(x,\alpha,g(x))| \\ \leq |d(x,\alpha,g(x)) - d(x_i,\alpha,g(x))| + |d(x_i,\alpha,g(x)) - d(x_i,\alpha,g(x_i))| \leq 0_\eta(1).$$

Thus we obtain

$$\sum_{i=1}^{p(\eta)} d(x_j,\alpha,g(x_j))\mathcal{H}_{N-1}(\partial\Omega_\eta(x_i)) \leq \int_{\cup_{1\leq j\leq p}\partial\Omega_\eta(x_i)} d(x,\alpha,g(x))d\mathcal{H}_{N-1} + 0_\eta(1) \\ \leq \int_{\partial\Omega} d(x,\alpha,g(x))d\mathcal{H}_{N-1} + 0_\eta(1).$$

Recalling (4.18), we can see that the sequence $\{w_\epsilon^{\delta,\eta(\epsilon)}\}_{\epsilon,\delta>0}$ satisfies (4.13). Hence we set $w_\epsilon^\delta = w_\epsilon^{\delta,\eta(\epsilon)}$, and the claim of Step 2 is proved. □

Step 3. In this step, we complete the proof of Proposition B for the special case $w_0 \equiv \alpha$. For any $\delta,\epsilon > 0$, we define w_ϵ^δ as follows:

$$w_\epsilon^\delta = \begin{cases} \alpha & \text{if } x \in \Omega\setminus\Omega_0, \\ {w^*}_\epsilon^\delta & \text{if } x \in \Omega_\delta, \end{cases}$$

where $\Omega_0 = \cup_{0<t<s^*}\pi_t(\partial\Omega)$ and ${w^*}_\epsilon^\delta$ is a function constructed in Step 2. In $\Omega^\delta \equiv \Omega_0\setminus\Omega_\delta$, we construct w_ϵ^δ using a convex combination between $g(x)$ and ${w^*}_\epsilon^\delta(\pi_\delta(x))$, i.e., for $x \in \Omega_0\setminus\Omega_\delta$,

$$w_\epsilon^\delta(x) = \frac{d_{\partial\Omega}(x)}{\delta}{w^*}_\epsilon^\delta|_{(\partial\Omega)_\delta}(\pi_\delta\circ\pi_{d_{\partial\Omega}(x)}^{-1}(x)) + \left(1-\frac{d_{\partial\Omega}(x)}{\delta}\right)g(\pi_{d_{\partial\Omega}(x)}^{-1}(x)). \tag{4.19}$$

Here $\pi_\delta(x)$ and $\pi_{d_{\partial\Omega}}(x)$ are the functions appearing in Lemma 4.1. Then we can easily see that $w_\epsilon^\delta \in W^{1,2}(\Omega)$ and $w_\epsilon^\delta(x) = g(x)$ for all $x \in \partial\Omega$.

In order to estimate the gradient of w_ϵ^δ, we fix ϵ, δ, $\{\Omega_\eta^\delta(x_i)\}_{i=1}^p$, and ω_η. We set

$$\begin{aligned}\Omega_1^\delta &= \{x \in \Omega^\delta : \pi_\delta \circ \pi_{d_{\partial\Omega}(x)}^{-1}(x) \in \cup_{1\le i\le p}\partial(\Omega_{\eta,1}^\delta(x_i))\},\\ \Omega_2^\delta &= \{x \in \Omega^\delta : \pi_\delta \circ \pi_{d_{\partial\Omega}(x)}^{-1}(x) \in \cup_{1\le i\le p}\partial(\Omega_{\eta,2}^\delta(x_i))\},\\ \omega_\eta^\delta &= \cup_{0<t<\delta}\pi_t(\omega_\eta)\end{aligned} \tag{4.20}$$

and have $\Omega^\delta = \Omega_1^\delta \cup \Omega_2^\delta \cup \omega_\eta^\delta$. Here $\Omega_{\eta,i}^\delta(x)$, $i = 1, 2$ is the domain appearing in Step 1. Furthermore, for simplicity, we set

$$\begin{aligned}\hat{g}(x) &= g(\pi_{d_{\partial\Omega}(x)}^{-1}(x)),\\ \hat{w}_\epsilon^\delta(x) &= {w^*}_\epsilon^\delta|_{(\partial\Omega)_\delta}(\pi_\delta \circ \pi_{d_{\partial\Omega}(x)}^{-1}(x)), \qquad x \in \Omega^\delta.\end{aligned}$$

Then from Lemma 4.1, we can see that there exists a constant C such that $|\nabla\hat{g}(x)| \le C$ for almost all $x \in \Omega^\delta$.

Now in the domains $\Omega_{\eta,1}^\delta$, $\Omega_{\eta,2}^\delta$, and Ω^δ, we will estimate the gradient of w_ϵ^δ and obtain the inequality (2.7). If $x \in \omega_\eta^\delta$, then from the construction of w_ϵ in Step 2, we see $v_\epsilon^{\delta,\eta} \equiv \alpha$ in a neighborhood of x, and so for almost all $x \in \omega_\eta^\delta$, we have

$$|\nabla w_\epsilon^{\delta,\eta}| \le \frac{C}{\delta} + C.$$

So we obtain

$$\int_{\omega^\delta}\left[\epsilon|\nabla w_\epsilon^{\delta,\eta}|^2 + \frac{1}{\epsilon}W(x, w_\epsilon^{\delta,\eta})\right]dx \le C\left(\frac{\epsilon}{\delta^2} + \epsilon + \frac{1}{\epsilon}\right)\delta\mathcal{H}_{N-1}(\omega). \tag{4.21}$$

For almost all $x \in \Omega_{\eta,1}^\delta(x_i)$, we have

$$\begin{aligned}|\nabla w_\epsilon^{\delta,\eta}| \le \frac{|\nabla d_{\partial\Omega}(x)|}{\delta}\hat{w}_\epsilon^{\delta,\eta}(x) &+ \frac{d_{\partial\Omega}(x)}{\delta}|\nabla\hat{w}_\epsilon^{\delta,\eta}(x)|\\ &+ \frac{|\nabla d_{\partial\Omega}(x)|}{\delta}\hat{g}(x) + \left(1 - \frac{d_{\partial\Omega}(x)}{\delta}\right)|\nabla\hat{g}(x)|.\end{aligned}$$

Here from the argument in Step 1, there exists a constant C_2 such that $|\nabla v_\epsilon^{\delta,\eta}(x)| \le C/(\epsilon^{5/8}\eta)$ for all $x \in \Omega_{\eta,1}^\delta$. Moreover, we have

$$|\Omega_{\eta,1}^\delta| \le C\delta(\epsilon^{5/8}\eta^{N-1})(\mathcal{H}_{N-1}(\partial\Omega)/\eta^{N-1}) \le C\delta\epsilon^{5/8}.$$

So we obtain

$$\begin{aligned}\int_{\Omega_1^\delta}\left[\epsilon|\nabla w_\epsilon^{\delta,\eta}|^2 + \frac{1}{\epsilon}W(x, w_\epsilon^{\delta,\eta})\right]dx &\le C\left[\epsilon\left(\frac{1}{\delta} + \frac{1}{\epsilon^{5/8}\eta} + 1\right)^2 + \frac{1}{\epsilon}\right]\delta\epsilon^{5/8}\\ &\le C\left(\frac{\epsilon}{\delta} + \frac{\delta}{\eta^2\epsilon^{1/4}} + \frac{\delta}{\epsilon}\right)\epsilon^{5/8}.\end{aligned} \tag{4.22}$$

For any $x \in \Omega_{\eta,2}^\delta(x_i)$, from Step 1 we see $w_\epsilon^*(x) \equiv g(x_i)$ in a neighborhood of x. Then from the Lipschitz continuity of $g(x)$ on $\partial\Omega$ and (4.19), we have

$$\begin{aligned}|\nabla w_\epsilon^{\delta,\eta}| &\le \frac{|\nabla d_{\partial\Omega}(x)|}{\delta}|g(x_i) - \hat{g}(x)| + \left(1 - \frac{d_{\partial\Omega}(x)}{\delta}\right)|\nabla\hat{g}|\\ &\le \frac{C}{\delta}|g(x_i) - \hat{g}(x)| + C \le C\frac{\eta}{\delta} + C.\end{aligned}$$

Therefore, we obtain

$$(4.23)\qquad \int_{\Omega_2^\delta}\left[\epsilon|\nabla w_\epsilon^{\delta,\eta}|^2+\frac{1}{\epsilon}W(x,w_\epsilon^{\delta,\eta})\right]dx\le C\left(\epsilon\left(\frac{\eta}{\delta}\right)^2+\epsilon+\frac{1}{\epsilon}\right)\delta\mathcal{H}_{N-1}(\partial\Omega).$$

Let $\sigma(\cdot)$ be a positive function with $\sigma(0)=0$ such that $\lim_{\epsilon\to 0}\mathcal{H}_{N-1}(\omega_{\eta(\epsilon)})/\sigma(\epsilon)=0$ and $\lim_{\epsilon\to 0}\epsilon^{5/8}/\sigma(\epsilon)=0$. Set $\delta_\epsilon=\epsilon\sigma(\epsilon)$ and define $w_\epsilon=w_\epsilon^{\delta_\epsilon}$. Then from (4.21)–(4.23), we obtain

$$(4.24)\qquad \lim_{\epsilon\to 0}\int_{\Omega^{\epsilon\sigma(\epsilon)}}\left[\epsilon|\nabla w_\epsilon|^2+\frac{1}{\epsilon}W(x,w_\epsilon)\right]dx=0.$$

Therefore, from (4.13) and (4.24), we obtain

$$\limsup_{\epsilon\to 0}\int_{\Omega}\left[\epsilon|\nabla w_\epsilon|^2+\frac{1}{\epsilon}W(x,w)\right]dx\le 2\int_{\partial\Omega}d(x,\alpha,g(x))d\mathcal{H}_{N-1}.$$

Hence the proof of Proposition B for the special case $w_0=\alpha$ is completed. □

Next, using the arguments of Steps 1–3, we will prove Proposition B. First we recall an approximation theorem for sets of finite perimeter by sets with smooth boundary. See [13].

LEMMA 4.3. *Let Ω be an open bounded subset of $\mathbf{R}^N$ with Lipschitz-continuous boundary. Let $A\subset\Omega$ be a set of finite perimeter in Ω with $0<|A|<|\Omega|$. Then there exists a sequence of open sets $\{A_k\}$ satisfying the following conditions:*

(i) $\partial A_k\cap\Omega\in C^2$;

(ii) $|(A_k\cap\Omega)\Delta A|\to 0$ *as* $k\to\infty$;

(iii) $P_\Omega(A_k)\to P_\Omega(A)$ *as* $k\to\infty$;

(iv) $\mathcal{H}_{N-1}(\partial A_k\cap\partial\Omega)=0$;

(v) $|A_k\cap\Omega|=|A|$ *for all sufficiently large* k.

We assume that the measurable set E has a C^2-smooth boundary in Ω; otherwise, the proof of Proposition B follows from Lemma 4.3 and the diagonal argument.

Using the same argument as in [14], we separate the domain Ω into six domains as follows:

$$\Omega_\alpha=\{x\in\Omega: d_{\partial\Omega}(x)\ge\tau_\epsilon+\delta_\epsilon,\ d_{\partial E}(x)<0\},$$
$$\Omega_\beta=\{x\in\Omega: d_{\partial\Omega}(x)\ge\tau_\epsilon+\delta_\epsilon,\ d_{\partial E}(x)>\tau_\epsilon+\delta'_\epsilon\},$$
$$\Omega_{\alpha,\beta}=\{x\in\Omega: d_{\partial\Omega}(x)\ge\tau_\epsilon+\delta_\epsilon,\ 0\le d_{\partial E}(x)\le\tau_\epsilon+\delta'_\epsilon\},$$
$$\Omega_{\alpha,g}=\{x\in\Omega: d_{\partial\Omega}(x)\le\tau_\epsilon+\delta_\epsilon,\ d_{\partial E}(x)<0\},$$
$$\Omega_{\beta,g}=\{x\in\Omega: d_{\partial\Omega}(x)\le\tau_\epsilon+\delta_\epsilon,\ d_{\partial E}(x)>\tau_\epsilon+\delta'_\epsilon\},$$
$$\Omega_r=\{x\in\Omega: d_{\partial\Omega}(x)\le\tau_\epsilon+\delta_\epsilon,\ 0\le d_{\partial E}(x)\le\tau_\epsilon+\delta'_\epsilon\},$$

where $d_{\partial E}(x)$ is a distance function to ∂E appearing in Lemma 4.2. We note that δ_ϵ and δ'_ϵ depend on $\partial\Omega$ and ∂E respectively. Here we set $w_\epsilon=\alpha$ in Ω_α and $w_\epsilon=\beta$ in Ω_β. In the domains $\Omega_{\alpha,\beta}$, $\Omega_{\alpha,g}$, and $\Omega_{\beta,g}$, by using the argument of Steps 1–3, we construct w_ϵ such that

$$(4.25)\qquad \limsup_{\epsilon\to 0}\int_{\Omega_{\alpha,\beta}}\left[\epsilon|\nabla w_\epsilon|^2+\frac{1}{\epsilon}W(x,w_\epsilon)\right]dx\le 2\int_{\partial E\cap\Omega}d(x,\alpha,\beta)d\mathcal{H}_{N-1},$$

$$(4.26)\qquad \limsup_{\epsilon\to 0}\int_{\Omega_{\alpha,g}}\left[\epsilon|\nabla w_\epsilon|^2+\frac{1}{\epsilon}W(x,w_\epsilon)\right]dx\le 2\int_{\partial E\cap\partial\Omega}d(x,\alpha,g(x))d\mathcal{H}_{N-1},$$

and

$$\limsup_{\epsilon\to 0}\int_{\Omega_{\beta,g}}\left[\epsilon|\nabla w_\epsilon|^2+\frac{1}{\epsilon}W(x,w_\epsilon)\right]dx\le 2\int_{\partial\Omega\setminus\partial E}d(x,\beta,g(x))d\mathcal{H}_{N-1}. \tag{4.27}$$

Furthermore, we separate the domain Ω_r into two domains Ω_r^1 and Ω_r^2 as follows:

$$\Omega_r^1=\Omega_r\cap\{x\in\Omega:d_{\partial\Omega}(x)\ge\delta_\epsilon,\ d_{\partial E}(x)\ge{\delta'}_\epsilon\},$$
$$\Omega_r^2=\Omega_r\cap\{x\in\Omega:d_{\partial\Omega}(x)\le\delta_\epsilon\ \text{or}\ d_{\partial E}(x)\le{\delta'}_\epsilon\}.$$

By means of the same argument used to obtain ω^δ in Step 3, we can construct w_ϵ such that $w_\epsilon(x)=\beta$ for $x\in\Omega_r^1$ and

$$|\nabla w_\epsilon|\le C/\delta_\epsilon+C/{\delta'}_\epsilon+C\qquad\text{for almost all } x\in\Omega_r^2,$$

and we have, as $|\Omega_r^2|=0(\tau_\epsilon\delta_\epsilon+\tau_\epsilon{\delta'}_\epsilon+\delta_\epsilon{\delta'}_\epsilon)$,

$$\lim_{\epsilon\to 0}\int_{\Omega_r}\left[\epsilon|\nabla w_\epsilon|^2+\frac{1}{\epsilon}W(x,w_\epsilon)\right]dx=\lim_{\epsilon\to 0}\left[\frac{C\epsilon}{\delta_\epsilon{\delta'}_\epsilon}+C\epsilon+\frac{C}{\epsilon}\right]|\Omega_r^2|=0. \tag{4.28}$$

Here we use the fact $\lim_{\epsilon\to 0}\epsilon^{1+5/8}/\delta_\epsilon=0$ and $\lim_{\epsilon\to 0}\epsilon^{1+5/8}/{\delta'}_\epsilon=0$.

Therefore, we have from (4.25)–(4.28) that

$$\limsup_{\epsilon\to 0}\int_\Omega\left[\epsilon|\nabla w_\epsilon|^2+\frac{1}{\epsilon}W(x,w_\epsilon)\right]dx\le 2\int_{\partial E\cap\Omega}d(x,\alpha,g(x))d\mathcal{H}_{N-1}$$
$$+2\int_{\partial\Omega\cap\partial E}d(x,\alpha,\beta)d\mathcal{H}_{N-1}+2\int_{\partial\Omega\setminus\partial E}d(x,\beta,g(x))d\mathcal{H}_{N-1}.$$

Furthermore, by the construction of w_ϵ, we can easily see that $w_\epsilon(x)=g(x)$ for all $x\in\partial\Omega$ and $\lim_{\epsilon\to 0}w_\epsilon=w_0$ in $L^1(\Omega:\mathbf{R}^n)$. Hence the proof of Proposition B is complete. □

5. Corollaries of Theorem 1. In this section, we will give some corollaries obtained from Theorem 1. From the definition of gamma convergence and by Propositions A and B, the result below follows.

THEOREM 2. *Let F_ϵ and F_0 be the functionals from $L^1(\Omega:\mathbf{R}^n)$ into $[0,\infty]$, which are given in §2. Then*

$$\Gamma-\lim_{\epsilon\to 0_+}F_\epsilon=F_0\quad\textit{in } L^1(\Omega:\mathbf{R}^n)\ \textit{topology}.$$

We recall the definition of gamma convergence (see [2]).

DEFINITION. *Given (X,τ), a topological space, and $F_n,F_\infty:X\to\overline{\mathbf{R}}$, a family of real (extended) valued functions, the sequence $\{F_n\}_{n=1}^\infty$ is said to* gamma converge *to F_∞ at $x\in X$ if the following two conditions hold:*

(i) *for every convergent sequence $x_n\to x$ in (X,τ),*

$$F_\infty(x)\le\liminf_{n\to\infty}F_\epsilon(x_n);$$

(ii) *there exists a convergent sequence $x_n\to x$ in (X,τ) such that*

$$\limsup_{n\to\infty}F_n(x_n)\le F_\infty(x).$$

When this property holds for every $x \in X$, the sequence $\{F_n\}_{n=1}^{\infty}$ is said to gamma converge to F_{∞} and $F_{\infty} = \Gamma(\tau) - \lim_{n\to\infty} F_n$.

On the other hand, in Theorem 1, the sequence of minimizers $\{u_\epsilon\}_{\epsilon>0}$ does not always generate interior layers. For example, if we consider problem (SP_ϵ) with $g \equiv 0$, we have $E_0 = \Omega$ or $\emptyset$. In addition, considering the family of *local* minimizers, from Theorem 2 and the results of [12], we obtain the following theorem.

THEOREM 3. *Let $u_0 \in L^1(\Omega : \mathbf{R}^n)$ be a isolated L^1-local minimizer of F_0, that is,*

$$\text{there exists a positive constant } \delta \text{ such that } F_0(u_0) < F_0(v)$$
$$\text{whenever} \quad u_0 \neq v \quad \text{and} \quad \|u_0 - v\|_{L^1(\Omega:\mathbf{R}^n)} \leq \delta.$$

Then there exist a constant $\epsilon_0 > 0$ and a sequence $\{u_\epsilon\}_{\epsilon<\epsilon_0}$ such that u_ϵ is a local minimizer of F_ϵ and $u_\epsilon \to u_0$ in $L^1(\Omega : \mathbf{R}^n)$ as $\epsilon \to 0$.

REFERENCES

[1] L. AMBROSIO, *Metric space valued functions of bounded variation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 17 (1990), pp. 439–473.

[2] H. ATTOUCH, *Variational Convergence for Functions and Operators*, Pitman, Boston, 1984.

[3] S. BALDO, *Minimal interface criterion for phase transitions in mixtures of Cahn-Hilliard fluid*, Ann. Inst. H. Poincaré Anal. Non Lin eaire, 7 (1990), pp. 67–90.

[4] A. C. BARROSO AND I. FONSECA, *Anisotropic singular perturbation: The vectorial case*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 527–571.

[5] M. S. BERGER AND L. E. FRAENKEL, *On the asymptotic solution of a nonlinear Dirichlet problem*, Math. Mech., 19 (1970), pp. 553–585.

[6] G. BOUCHITTE, *Singular perturbation of variational problems arising from a two-phase transition model*, Appl. Math. Optim., 21 (1990), pp. 289–314.

[7] L. C. EVANS AND R. F. GARIEPY, *Measure Theory and Finite Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[8] P. C. FIFE AND W. M. GREENLEE, *Interior transition layers for elliptic boundary value problems with a small parameter*, Russian Math. Surveys, 29 (1974), pp. 103–131.

[9] I. FONSECA AND L. TARTAR, *The gradient theory of phase transitions for systems with two potential wells*, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 89–102.

[10] M. E. GURTIN AND H. MATANO, *On the structure of equilibrium phase transitions within the gradient theory of fluids*, Quart. Appl. Math., 46 (1988), pp. 301–317.

[11] K. ISHIGE, *Singular perturbations of variational problems of vector valued functions*, Nonlinear Anal., 23 (1994), pp. 1453–1466.

[12] R. V. KOHN AND P. STERNBERG, *Local minimizer and singular perturbations*, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 69–84.

[13] L. MODICA, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal., 98 (1987), pp. 123–142.

[14] ———, *Gradient theory of phase transitions with boundary contact energy*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 5 (1987), pp. 453–486.

[15] L. MODICA AND S. MORTOLA, *Il limite nella Γ-convergenza di una famiglia di funzionali ellittichi*, Boll. Un. Math. Ital., A14 (1977), pp. 526–529.

[16] ———, *Un esenpio di Γ-convergenza*, Boll. Un. Mat. Ital. B, 14 (1977), pp. 285–299.

[17] N. C. OWEN, *Nonconvex variational problems with general singular perturbations*, Trans. Amer. Math. Soc., 310 (1988), pp. 393–404.

[18] N. C. OWEN, J. RUBINSTEIN, AND P. STERNBERG, *Minimizers and gradient flows for singularly perturbed bi-stable potentials with a Dirichlet condition*, Proc. Roy. Soc. London Ser. A, 429 (1990), pp. 505–532.

[19] N. C. OWEN AND P. STERNBERG, *Nonconvex problems with anisotropic perturbations*, Nonlinear Anal., 16 (1991), pp. 705–720.

[20] P. STERNBERG, *The effect of a singular perturbation on nonconvex variational problems*, Arch. Rational Mech. Anal., 101 (1988), pp. 209–260.

[21] W. P. ZIEMER, *Weakly Differentiable Functions*, Springer-Verlag, Berlin, 1989.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 638–660, May 1996

SOME A PRIORI ESTIMATES FOR A SINGULAR EVOLUTION EQUATION ARISING IN THIN-FILM DYNAMICS*

STEPHEN H. DAVIS†, EMMANUELE DIBENEDETTO‡, AND DAVID J. DILLER‡

Abstract. This paper considers the singular evolution equation $u_t - \Delta \ln u = 0$, particularly the corresponding Cauchy problem. This equation arises in the study of thin film dynamics and as the formal limit as $m \to 0$ of the porous-medium equation. Through the use of local sup-estimates, similar to those for the porous-medium equation (see E. DiBenedetto and Y. L. Kwong, Intrinsic Harnack estimates and extinction profile for certain parabolic equations, *Trans. Amer. Math. Soc.*, 330 (1992), pp. 783–811), and a global Harnack-type inequality, a critical decay rate for solutions in three or more space dimensions as $|x| \to \infty$ is found. In particular, if the initial data decays faster than this critical rate, there is no solution to the Cauchy problem.

Key words. thin-film dynamics, ill-posed problem, porous-medium equation, global Harnack inequality, finite extinction time

AMS subject classifications. 35B45, 35K65, 35Q35

1. Introduction.

We establish a spectrum of qualitative and quantitative properties of nonnegative solutions to the singular evolution equation

$$u_t - \Delta \ln u = 0 \qquad \text{in } \mathbb{R}^N \times \mathbb{R}^+. \tag{1}$$

Singular equations of this type have been studied, primarily in one space dimension, in, for example, [5] and [11]. This equation is singular since its modulus of ellipticity, u^{-1}, blows up at points where $u = 0$. At such points, ln u is not defined. We regard ln u as an element of $L^p_{\text{loc}}(\mathbb{R}^N \times \mathbb{R}^+)$, for some $p \geq 1$, so that (1) can be interpreted in the sense of distributions. Precisely, we say that a measurable function $u : \mathbb{R}^N \times \mathbb{R}^+ \to \mathbb{R}^+$ is a weak solution to (1) if

$$\begin{cases} u, \ln u \in L^1_{\text{loc}}(\mathbb{R}^N \times \mathbb{R}^+) \text{ and, for all } \phi \in C_0^\infty(\mathbb{R}^N \times \mathbb{R}^+), \\ \displaystyle\iint_{\mathbb{R}\times\mathbb{R}^+} (u\phi_t + \ln u \Delta\phi)dx\,dt = 0. \end{cases} \tag{2}$$

The equation has theoretical significance as it arises in the two-dimensional Ricci flow (see [10], [16]) and a physical significance in connection with the dynamics of thin liquid films (see [3], [4], [15]). In one space dimension, it has also been proposed as a model for the limiting density distribution in the kinetics of two gases moving against each other and obeying the Boltzmann equation (see [6], [12]). Since our interest is mainly in the physical aspects of the partial differential equation (PDE), we will describe below some currently proposed models of film rupture and their connection with (1). The equation can also be viewed, at least formally, as a limiting case of the porous medium equation,

$$v_\tau - \Delta v^m = 0, \qquad v \geq 0, \qquad m > 0, \tag{3}$$

*Received by the editors May 9, 1994; accepted for publication (in revised form) October 4, 1994.

†Department of Engineering Sciences and Applied Mathematics, The Technological Institute, Northwestern University, Evanston, IL 60208.

‡Department of Mathematics, Lunt Hall, Northwestern University, Evanston, IL 60208. The research of the second author was partially funded by NATO grant 5-205 (CRG.920/26) 458/92/JARC-501.

when $m \to 0$. Indeed, the formal change of variables,

$$\tau = mt, \qquad u(x,t) = v(x,mt),$$

transforms (3) into

$$u_t - \Delta \frac{u^m - 1}{m} = 0,$$

from which (1) follows, formally, by letting $m \to 0$. This formal connection prompts some of our analysis below. We will show that some peculiar properties of the porous-medium equation, such as sup-bounds, semiconvexity estimates, and extinction in finite time, continue to hold for solutions of (1), whereas others, such as behavior at infinity, do not. A dramatic difference is that while the Cauchy problem in $\mathbb{R}^N \times \mathbb{R}^+$ for the porous-medium equation (3) can always be solved globally in time for every initial datum

$$u_0 \geq 0 \quad u_0 \in C_0^\infty(\mathbb{R}^N),$$

the corresponding Cauchy problem for (1) cannot be solved for this datum in space dimensions $N \geq 3$. Another difference is the behavior of the solutions as $|x| \to \infty$. For any $m > 0$, however small, solutions of (3) have unrestricted behavior as $|x| \to \infty$ (see [7]), whereas the behavior of solutions of (1) as $|x| \to \infty$ is rather rigid. One of the results of this note is that if $N \geq 3$, any solution of (1) in $\mathbb{R}^N \times (0,T)$ as $|x| \to \infty$ decays no faster than $|x|^{-2}$.

In this note, we assume that u is a classical solution of (1) and derive some a priori estimates on its behavior. The point here is to find estimates and properties that are *independent* of any sort of lower bound on u. Thus, by a limiting process, these are indeed a priori estimates and properties of the weak solutions in the sense of (2).

1.1. Main results: Upper bounds and behavior at infinity. Nonnegative solutions of (1) are locally bounded in $\mathbb{R}^N \times \mathbb{R}^+$ if

$$u \in L^r_{\text{loc}}(\mathbb{R}^N \times \mathbb{R}^+), \qquad r > \max\left\{1, \frac{N}{2}\right\}. \tag{4}$$

We show by a counterexample that the indicated order of integrability is optimal for a sup-bound to hold (see Remark 2.6). One of the interests of this result is that if $u_0 \in L^r_{\text{loc}}(\mathbb{R}^n)$, then (4) holds, and then the local boundedness of u can be established. This supplies necessary conditions for the forward Cauchy problem to have a locally bounded solution.

For the porous-medium equation (3), local boundedness is guaranteed if (see [7])

$$v \in L^r_{\text{loc}}(\mathbb{R}^N \times \mathbb{R}^+), \qquad N(m-1) + 2r > 0. \tag{5}$$

Thus the integrability condition (4) can be regarded as the limiting case of the integrability condition prescribed by (5).

These estimates are *global* in $\mathbb{R}^N$ if, for some $t_0 > 0$, the function $x \to u(x,t_0)$, roughly speaking, does not decay slower than $|x|^{-2}$ as $x \to \infty$. This is meant in the sense that

$$\sup_{\rho > 1} \fint_{B_\rho} [|x|^2 u(x,t_0)]^r \, dx < \infty, \quad \text{for some } r > \max\left\{1, \frac{N}{2}\right\}. \tag{6}$$

Here, for a region Ω of finite Lebesgue measure and for $f \in L^1(\Omega)$,

$$⨍_\Omega f\,dx = \frac{1}{\text{meas}\ \Omega}\int_\Omega f\,dx$$

denotes the integral average of f over Ω. If (6) holds, the solution $u(\cdot, t)$ decays, for $t > t_0$, pointwise no slower than $|x|^{-2}$ as $x \to \infty$. Such an upper bound, in the growth of u, is optimal as shown by the lower bounds described below.

1.2. Main results: Lower bounds and behavior at infinity. We first establish that a solution u for which $\ln u \in L^\infty_{\text{loc}}(\mathbb{R}^+; L^1_{\text{loc}}(\mathbb{R}^N))$ is bounded away from 0, locally in $\mathbb{R}^N \times \mathbb{R}^+$. Here, we say that $f \in L^\infty_{\text{loc}}(\mathbb{R}^+; L^1_{\text{loc}}(\mathbb{R}^N))$ if, for each $\rho > 0$ and $0 < t_0 < t_1 < \infty$,

$$\sup_{t_1 \le t \le t_2} \int_{B_\rho} |f(x,t)|\,dx \le C(\rho, t_0, t_1).$$

This is the content of §3.1. Then, using the estimates leading to this fact, we prove in §3.3–3.5 that if $N \ge 3$, solutions of (1) decrease, as $|x| \to \infty$, no faster than $|x|^{-2}$. Thus the rate $|x|^{-2}$ of increase/decrease of $u(\cdot, t)$ as $x \to \infty$ is optimal.

1.3. Ill-posed problems. A key fact is that the results indicated in the previous sections hold for solutions of (1) originating from either initial or final data. Thus, if $N \ge 3$, *neither* the forward nor the backward Cauchy problem is well posed if the data decay faster than $|x|^{-2}$ as $|x| \to \infty$. Thus, in particular, if $N \ge 3$ and if the initial datum is compactly supported, neither the forward nor the backward Cauchy problem associated with (1) is well posed.

1.4. Extinction in finite time. We can show that, if $u_0 \in L^{N/2}(\mathbb{R}^N)$ and $N \ge 3$, then any classical solution with this initial data must become extinct after some finite time T. Furthermore, T is bounded by a constant depending only on N and the $L^{N/2}(\mathbb{R}^N)$ norm of u_0. We also indicate through a family of examples why a result of this type is impossible for an initial datum $u_0 \in L^r(\mathbb{R}^N) r > \frac{N}{2}$. This result is merely a sufficient condition to guarantee that a solution has a finite extinction time. The explicit solutions given by Example 1.5 below show that this condition is not necessary for finite extinction to occur. This result can be viewed as the limiting case of the corresponding result for the porous-medium equation. For the latter, finite extinction occurs when the initial data is in $L^r(\mathbb{R}^N)$, where $r = \frac{N(1-m)}{2}$, as long as $r > 1$ (see [2]).

1.5. Background on thin-film dynamics. When a uniform viscous-liquid film lies on a rigid plate, small disturbances to the planar interface can grow, leading to the rupture of the film locally and laying bare the substrate. The instability is driven by the presence in ultrathin films ($100\overset{\circ}{A}$–$1000\overset{\circ}{A}$) of van der Waals attractions that drive thin regions to zero thickness. The van der Waals attractions can be modeled as a potential, ϕ, of an extra body force, $\phi \sim u^3$, where u is the local film thickness. Williams and Davis [15] derive an evolution equation governing the dynamics, viz.

$$u_t \pm \Delta \ln u + \text{ div } (u^3 \nabla(\Delta u)) = 0 \quad \text{in } \mathbb{R}^2 \times \mathbb{R}^+, \tag{7}$$

with the upper sign corresponding to the unstable case above. When the lower sign holds, the van der Waals forces are repulsive and the uniform film is stable. The last term represents the stabilizing effects of surface tension on the liquid–gas interface.

Williams and Davis [15] numerically integrated the unstable version of equation (7) in $\mathbb{R} \times \mathbb{R}^+$ and found that in a finite time, $t = t_R$, there is a value of x for which $u = 0$, i.e., the film ruptures. Burelbach et al. [4] did more extensive numerical computations and found that, at a rupture point, as $t \to t_R^-$, $u \sim t_R - t$, suggesting that locally $u \sim (t_R - t)f(x)$, which corresponds to a solution of the unstable version of (7) with the surface tension neglected. Thus the equations, locally, should be

$$u_t \pm \Delta \ln u = 0 \quad \text{in } \mathbb{R}^2 \times \mathbb{R}^+. \tag{8}$$

As mentioned in §1.3, most of our results hold for both the stable and unstable versions of (7).

1.6. Some explicit solutions. As a way of checking our analysis, we list here some known explicit solutions of (1) in $\mathbb{R}^N \times \mathbb{R}^+$. Some are obtained from the explicit Barenblatt–Pattle (see [13]) solutions by formally letting $m \to 0$, and others are new.

Example 1.1. For $N = 1$ and $\lambda > 0$,

$$u(x,t) = \frac{\lambda^2 (T-t)_+}{1+\cosh(\lambda x)}.$$

Example 1.2. For $N = 1$ and $\lambda > 0$,

$$u(x,t) = \frac{2t}{\lambda t^2 + |x|^2}.$$

Example 1.3. For $N = 2$ and $T > 0$,

$$u(x,t) = \frac{8\lambda (T-t)_+}{(\lambda + |x|^2)^2}.$$

Example 1.4. For $N = 2$, $\beta > 0$, and $\alpha > 0$,

$$u(x,t) = \frac{\beta}{e^{\frac{4\alpha t}{\beta}} + \alpha |x|^2}.$$

Example 1.5. For $N \geq 3$, $\lambda \geq 0$, and $T > 0$,

$$u(x,t) = \frac{2(N-2)(T-t)_+^{\frac{N}{N-2}}}{\lambda + (T-t)_+^{\frac{2}{N-2}} |x|^2}.$$

2. Some sup-bounds. For what follows, we assume that we have a classical solution to the Cauchy problem

$$\begin{cases} u_t - \Delta \ln u = 0 & \text{in } \mathbb{R}^N \times \mathbb{R}^+, \\ u(x,0) = u_0(x), & u_0 \in C^\infty(\mathbb{R}^N). \end{cases} \tag{9}$$

In addition, we assume that $u_0 > 0$ to make sense out of (9). Our estimates in this section are entirely local and hold for classical solutions of (9) with no specification of the behavior of $u(x,t)$ as $|x| \to \infty$. These a priori estimates can be used to actually construct weak solutions of (9) and to infer the asymptotic behavior of $u(x,t)$ as $|x| \to \infty$.

For $\rho > 0$ and $x_0 \in \mathbb{R}^N$, we let $B_\rho(x_0) = \{|x - x_0| < \rho\}$ denote the ball of radius ρ about x_0. If $x_0 = 0$, then we write $B_\rho(0) = B_\rho$. We let r be a real number satisfying

$$r > \max\left\{1, \frac{N}{2}\right\} \tag{10}$$

and set

$$\kappa_r = 2r - N. \tag{11}$$

For $N \geq 2$, the requirement (10) is equivalent to $\kappa_r > 0$. In the estimates to follow, we will denote by γ and b generic positive constants depending only on N and r.

2.1. Local sup-bounds.

PROPOSITION 2.1. *Let u be a classical solution to* (9) *and let r satisfy* (10). *Then there exists a constant γ depending only on N and r such that*

$$\|u(\cdot, t)\|_{\infty, B_\rho(x_0)} \leq \gamma \left\{ (t-\tau)^{-\frac{N}{\kappa_r}} \sup_{\tau < s < t} \|u(\cdot, t)\|_{r, B_{2\rho}(x_0)}^{\frac{2r}{\kappa_r}} + \frac{t-\tau}{\rho^2} \right\}, \tag{12}$$

for all $x_0 \in \mathbb{R}^N$ and for all $0 \leq \tau < t$.

Proof. Without loss of generality, we may assume that $x_0 = 0$ and $\tau = 0$. Let $0 < \sigma < 1$ and consider the sequences of radii and time levels

$$\rho_n = \rho(1 + \sigma 2^{-n}) \quad \text{and} \quad t_n = \frac{t}{2}(1 - 2^{-n-1}).$$

Then $\rho_0 = \rho(1+\sigma)$ and $t_0 = t/4$. Also, as $n \to \infty$, ρ_n decreases to ρ, and t_n increases to $\frac{t}{2}$. So, with $Q_n = B_{\rho_n} \times (t_n, t)$, we have the following relations:

$$B_{\rho(1+\sigma)} \times \left(\frac{t}{4}, t\right) = Q_0 \supseteq Q_1 \supseteq \cdots \subseteq Q_{n-1} \supseteq Q_n \supseteq \cdots \supseteq Q_\infty = B_\rho \times \left(\frac{t}{2}, t\right).$$

Let $\zeta_n \in C_0^\infty(Q_n)$ satisfy

$$\begin{cases} 0 \leq \zeta_n \leq 1 & \text{in } Q_n, \\ \zeta_n = 1 & \text{in } Q_{n+1}, \\ |D\zeta_n|^2, |\Delta\zeta_n| \leq \dfrac{\gamma 4^n}{(\sigma\rho)^2} & \text{in } Q_n, \\ 0 \leq \dfrac{\partial \zeta_n}{\partial t} \leq \dfrac{\gamma 2^n}{t} & \text{in } Q_n. \end{cases}$$

Multiply the first term of (9) by the test function $(u - k_{n+1})_+^{r-1}\zeta_n^2$ and integrate over Q_n^s. Here $k_n = k(1 - 2^{-n})$, where $k > 0$ will be chosen later, and $Q_n^s = B_{\rho_n} \times (t_n, s)$, where $t_n \leq s \leq t$. Note that $k_0 = 0$ and, as $n \to \infty$, k_n increases to k. Set

$$M = \sup_{Q_0} u.$$

Then

$$\begin{aligned}
0 &= \iint_{Q_n^s} (u_t - \Delta \ln u)(u - k_{n+1})_+^{r-1}\zeta_n^2\, dx\, d\tau \\
&= \frac{1}{r}\iint_{Q_n^s} \frac{\partial (u - k_{n+1})_+^r}{\partial t}\zeta_n^2\, dx\, d\tau \\
&\quad + \iint_{Q_n^s} \frac{Du}{u}[D(u - k_{n+1})_+^{r-1}\zeta_n^2 + 2(u - k_{n+1})_+^{r-1}\zeta_n D\zeta_n]\, dx\, d\tau \\
&= \frac{1}{r}\int_{B_{\rho_n}} (u - k_{n+1})_+^r(x, s)\zeta_n^2(x, s)\, dx - \frac{2}{r}\iint_{Q_n^s} (u - k_{n+1})_+^r\zeta_n\zeta_{n,t}\, dx\, d\tau \\
&\quad + \frac{4(r-1)}{r^2}\iint_{Q_n^s} \frac{1}{u}|D(u - k_{n+1})_+^{\frac{r}{2}}|^2\zeta_n^2\, dx\, d\tau \\
&\quad + 2\iint_{Q_n^s} \frac{Du}{u}(u - k_{n+1})_+^{r-1}\zeta_n D\zeta_n\, dx\, d\tau \\
&= \frac{1}{r}T_1 - \frac{2}{r}T_2 + \frac{4(r-1)}{r^2}T_3 + 2T_4.
\end{aligned} \tag{13}$$

$$|T_2| \le \gamma\frac{2^n}{t}\iint_{Q_n^s} (u - k_{n+1})_+^r\, dx\, d\tau. \tag{14}$$

$$\begin{aligned}
\iint_{Q_n^s} |D[(u - k_{n+1})_+^{\frac{r}{2}}\zeta_n]|^2\, dx\, d\tau \le \iint_{Q_n^s} &|D(u - k_{n+1})_+^{\frac{r}{2}}|^2\zeta_n^2\, dx\, d\tau \\
&+ \frac{\gamma 4^n}{\sigma^2\rho^2}\iint_{Q_n^s} (u - k_{n+1})_+^r\, dx\, d\tau.
\end{aligned}$$

Thus

$$\begin{aligned}
T_3 &\ge \frac{1}{M}\iint_{Q_n^s} |D(u - k_{n+1})_+^{\frac{r}{2}}|^2\zeta_n^2 dx d\tau \\
&\ge \frac{1}{M}\iint_{Q_n^s} |D[(u - k_{n+1})_+^{\frac{r}{2}}\zeta_n|^2 dx d\tau - \frac{\gamma 4^n}{M\sigma^2\rho^2}\iint_{Q_n^s} (u - k_{n+1})_+^r dx d\tau.
\end{aligned} \tag{15}$$

Set

$$f_n(\delta) = \int_{k_{n+1}}^{\delta} \frac{(s - k_{n+1})_+^{r-1}}{s} ds.$$

Then

$$|f_n(\delta)| \le \frac{(\delta - k_{n+1})_+^r}{k_{n+1}} \le \frac{2(\delta - k_{n+1})_+^r}{k}.$$

Thus

$$\begin{aligned}
T_4 &= \iint_{Q_n^s} Df_n(u)\zeta_n D\zeta_n dx d\tau \\
&= -\iint_{Q_n^s} f_n(u)(|D\zeta_n|^2 + \zeta_n\Delta\zeta_n) dx d\tau.
\end{aligned}$$

$$
\begin{aligned}
|T_4| &\le \gamma \frac{4^n}{\sigma^2\rho^2} \iint_{Q_n^s} |f_n(u)| dx d\tau \\
&\le \gamma \frac{4^n}{k\sigma^2\rho^2} \iint_{Q_n^s} (u-k_{n+1})_+^r dx d\tau.
\end{aligned} \tag{16}
$$

Then, by combining (13)–(16), taking the supremum for $s \in (t_n, t)$, and assuming that $\frac{t}{\rho^2} \le k \le M$, we find

$$
\begin{aligned}
\sup_{t_n \le s \le t} \int_{B_{\rho n}} (u-k_{n+1})_+^r(x,s)\zeta_n^2(x,s)dx &+ \frac{1}{M} \iint_{Q_n} \left|D[(u-k_{n+1})_+^{\frac{r}{2}}\zeta_n]\right|^2 dx d\tau \\
&\le \gamma \frac{4^n}{\sigma^2 t} \iint_{Q_n} (u-k_{n+1})_+^r \, dx d\tau \\
&\le \gamma \frac{4^n}{\sigma^2 t} Y_n,
\end{aligned} \tag{17}
$$

where

$$
Y_n = \iint_{Q_n} (u-k_n)_+^r dx d\tau.
$$

Since $(u-k_{n+1})_+^r \zeta_n^2$ vanishes on the lateral boundary of Q_n, we can apply the space–time version of the Sobolev Embedding Theorem (see [7, Chap. 1, §3]), that is,

$$
\begin{aligned}
\iint_{Q_n} [(u-k_{n+1})_+^{\frac{r}{2}}\zeta_n]^{2\frac{N+2}{N}} dx d\tau &\le \gamma \left(\iint_{Q_n} |D[(u-k_{n+1})_+^{\frac{r}{2}}\zeta_n]|^2 dx d\tau \right) \\
&\times \left(\sup_{t_n \le s \le t} \int_{B_{\rho n}} [(u-k_{n+1})_+^{\frac{r}{2}}\zeta_n]^2(x,s)dx \right)^{\frac{2}{N}}.
\end{aligned} \tag{18}
$$

By Hölder's inequality,

$$
\begin{aligned}
Y_{n+1} &= \iint_{Q_{n+1}} (u-k_{n+1})_+^r \, dx d\tau \\
&\le \iint_{Q_n} (u-k_{n+1})_+^r \zeta_n^2 dx d\tau \\
&\le \left(\iint_{Q_n} [(u-k_{n+1})_+^r \zeta_n^2]^{\frac{N+2}{N}} dx d\tau \right)^{\frac{N}{N+2}} |A_n|^{\frac{2}{N+2}},
\end{aligned}
$$

where $A_n = \{(x,t) \in Q_n | u(x,t) > k_{n+1}\}$. Then, by applying (17) and (18),

$$
Y_{n+1} \le \gamma \frac{4^n}{\sigma^2 t} M^{\frac{N}{N+2}} Y_n |A_n|^{\frac{2}{N+2}}. \tag{19}
$$

To estimate $|A_n|$,

$$\begin{aligned} Y_n &= \iint_{Q_n} (u-k_n)_+^r dx d\tau \\ &\geq \iint_{Q_n} (u-k_n)_+^r \chi_{\{u>k_{n+1}\}} dx d\tau \\ &\geq (k_{n+1}-k_n)^r |A_n| \\ &= \frac{k^r}{2^{(n+1)r}} |A_n|. \end{aligned}$$

Combining this with (19) yields

$$Y_{n+1} \leq \gamma \frac{b^n M^{\frac{N}{N+2}}}{\sigma^2 t k^{\frac{2r}{N+2}}} Y_n^{1+\frac{2}{N+2}},$$

where $b > 1$ is some number depending only on N and r. Thus (see [7, Chap. 1, Lem. 4.1])

$$\lim_{n\to\infty} Y_n = 0 \quad \text{if } k > \gamma \frac{M^{\frac{N}{2r}} Y_0^{\frac{1}{r}}}{(\sigma^2 t)^{\frac{N+2}{2r}}}.$$

Thus, with

$$M = \sup_{B_{\rho(1+\sigma)}\times[\frac{t}{4},t]} u,$$

$$\sup_{B_\rho\times(\frac{t}{2},t)} u \leq \frac{\gamma M^{\frac{N}{2r}}}{(\sigma^2 t)^{\frac{N+2}{2r}}} \left(\int_0^t \int_{B_{\rho(1+\sigma)}} u^r dx d\tau \right)^{\frac{1}{r}} + \frac{t}{\rho^2} \tag{20}$$

for any $\rho > 0$ and $0 < \sigma \leq 1$. Note that the $\frac{t}{\rho^2}$ term is included to satisfy the assumption that $k \geq \frac{t}{\rho^2}$.

Fix $R > 0$ and $T > 0$. Consider the sequences

$$\rho_n = R\sum_{i=0}^{n} 2^{-i} \quad \text{and} \quad t_n = \frac{T}{2^{n+1}}.$$

Thus $B_{\rho_n} \times (t_n, T) \subseteq B_{\rho_{n+1}} \times (t_{n+1}, T)$. Define σ_n so that $\rho_{n+1} = \rho_n(1+\sigma_n)$. Then it follows that $\sigma_n \geq 2^{-n-2}$. By applying (20) with $t = t_n$, $\rho = \rho_n$, and $\sigma = \sigma_n$,

$$\begin{aligned} \sup_{B_{\rho n}\times(t_n,t_{n-1})} u &\leq \gamma \frac{M_{n+1}^{\frac{N}{2r}}}{(\sigma_n^2 t_{n-1})^{\frac{N+2}{2r}}} \left(\int_0^{t_{n-1}} \int_{B_{\rho_{n+1}}} u^r dx d\tau \right)^{\frac{1}{r}} + \frac{t_{n-1}}{\rho^2} \\ &\leq \gamma \frac{M_{n+1}^{\frac{N}{2r}} b^n}{T^{\frac{N+2}{2r}}} \left(\int_0^T \int_{B_{2R}} u^r dx d\tau \right)^{\frac{1}{r}} + \frac{T}{\rho^2}, \end{aligned} \tag{21}$$

where

$$M_n = \sup_{B_{\rho n}\times(t_n,T)} u.$$

Thus, from (21),

$$M_n \leq \gamma M_{n+1}^{\frac{N}{2r}} b^n \left(T^{-\frac{N+2}{2}} \int_0^T \int_{B_{2R}} u^t dx d\tau \right)^{\frac{1}{r}} + \frac{T}{R^2}, \tag{22}$$

where γ and b are constants depending only on N and r. Fix $\eta \in (0,1)$ and apply Young's inequality to (22) to arrive at

$$M_n \leq \eta M_{n+1} + \gamma b^n \left(T^{-\frac{N+2}{2}} \int_0^T \int_{B_{2R}} u^r dx d\tau \right)^{\frac{2}{2r-N}} + \frac{T}{R^2}. \tag{23}$$

Then inductively, with $F = \left(T^{-\frac{N+2}{2}} \int_0^T \int_{B_{2R}} u^r dx d\tau \right)^{\frac{2}{2r-N}}$,

$$M_0 \leq \eta^k M_k + \gamma F \sum_{j=0}^{k-1} b^j \eta^j + \frac{T}{R^2} \sum_{j=0}^{k-1} \eta^j.$$

By selecting $\eta < \frac{1}{b}$ and letting $k \to \infty$,

$$\begin{aligned} \sup_{B_R \times (\frac{T}{2}, T)} u &\leq \gamma \left[T^{-\frac{N+2}{2r-N}} \left(\int_0^T \int_{B_{2R}} u^r dx d\tau \right)^{\frac{2}{2r-N}} + \frac{T}{R^2} \right] \\ &\leq \gamma \left[T^{-\frac{N}{\kappa_r}} \sup_{0 \leq s \leq T} \left(\int_{B_{2r}} u^r(x,s) dx \right)^{\frac{2}{\kappa_r}} + \frac{T}{R^2} \right]. \end{aligned} \tag{24}$$

Then, by replacing T by t and R by ρ, (24) is exactly (12). □

2.2. Local integral estimates.

PROPOSITION 2.2. *Assume that u is a classical solution to* (9) *and that* $r > 1$. *Then there is a constant* γ *depending only on* N *and* r *such that*

$$\sup_{\tau \leq s \leq t} \int_{B_\rho(x_0)} u^r(x,s) dx \leq \gamma \left[\int_{B_{2\rho}(x_0)} u^r(x,\tau) dx + \frac{(t-\tau)^r}{\rho^{\kappa_r}} \right] \tag{25}$$

for all $\rho > 0$, $0 < \tau < t$, *and* $x_0 \in \mathbb{R}^N$.

Remark 2.3. An estimate like (25) is impossible for solutions of the heat equation. To see this, suppose that, for any solution v to the heat equation in $B_{2\rho} \times (0,T)$,

$$\int_{B_\rho} v^r(x,t) dx \leq \gamma(\|v(\cdot,0)\|_{r,B_{2\rho}}, N, r, t, \rho)$$

for any $0 < t < T$. Then fix $y \in \mathbb{R}^N$ such that $|y| > 0$ and consider

$$v_C(x,t) = \frac{C}{t^{\frac{N}{2}}} e^{-\frac{|x-y|^2}{4t}}.$$

Select $0 < \rho < |y|/2$. If the estimate above is true for some r, then

$$\frac{C^r}{t^{\frac{rN}{2}}} \int_{B_\rho} e^{-r\frac{|x-y|^2}{4t}} dx \leq \gamma(0, N, r, t, \rho).$$

Of course, this is impossible, since the left side → ∞ as $C \to \infty$, whereas the right side is independent of C.

Proof of Proposition 2.2. Again, we take $x_0 = 0$ and $\tau = 0$. Fix $\sigma \in (0,1]$ and choose $\zeta \in C_0^\infty(B_{\rho(1+\sigma)})$ satisfying

$$\begin{cases} 0 \leq \zeta \leq 1 & \text{in } B_{\rho(1+\sigma)}, \\ \zeta(x) = 1 & \text{in } B_\rho, \\ |D\zeta|^2, |\Delta\zeta| \leq \dfrac{\gamma}{(\sigma\rho)^2} & \text{in } B_{\rho(1+\sigma)}. \end{cases}$$

Multiply the first term of (9) by $u^{r-1}\zeta$ and integrate over $Q_s = B_{(1+\sigma)\rho} \times (0,s)$.

$$\begin{aligned} 0 &= \iint_{Q_s} (u_t - \Delta \ln u) u^{r-1} \zeta dx d\tau \\ &= \frac{1}{r}\iint_{Q_s} \frac{\partial u^r}{\partial t}\zeta dx d\tau + \iint_{Q_s} \frac{Du}{u}[Du^{r-1}\zeta + u^{r-1}D\zeta] dx d\tau \\ &= \frac{1}{r}\iint_{Q_s} \frac{\partial u^r}{\partial t}\zeta dx d\tau + (r-1)\iint_{Q_s} u^{r-3}|Du|^2\zeta dx d\tau \\ &\quad + \iint_{Q_s} u^{r-2} Du D\zeta dx d\tau \\ &= \frac{1}{r}T_1 + (r-1)T_2 + T_3. \end{aligned} \tag{26}$$

Since ζ is independent of t,

$$T_1 = \int_{B_{(1+\sigma)\rho}} u^r(x,s)\zeta(x)dx - \int_{B_{(1+\sigma)\rho}} u_0^r(x)\zeta(x)dx d\tau. \tag{27}$$

$$\begin{aligned} T_3 &= \frac{1}{r-1}\iint_{Q_s} D(u^{r-1})D\zeta dx d\tau \\ &= \frac{-1}{r-1}\iint_{Q_s} u^{r-1}\Delta\zeta dx d\tau \\ &\geq \frac{-\gamma}{(\sigma\rho)^2}\iint_{Q_s} u^{r-1} dx d\tau. \end{aligned} \tag{28}$$

By combining (26)–(28), using the fact that $T_2 \geq 0$, and taking the supremum for $0 \geq s \geq t$,

$$\sup_{0\leq s\leq t} \int_{B_\rho} u^r(x,s)dx \leq \int_{B_{(1+\sigma)\rho}} u_0^r dx + \frac{\gamma}{(\sigma\rho)^2}\iint_{Q_t} u^{r-1} dx d\tau. \tag{29}$$

Then, by applying Hölder's inequality to the last term of (29),

$$\begin{aligned} \sup_{0\leq s\leq t} \int_{B_\rho} u^r(x,s)dx &\leq \int_{B_{(1+\sigma)\rho}} u_0^r dx \\ &\quad + \frac{\gamma}{\sigma^2}\left(\frac{t^r}{\rho^{2r-N}}\right)^{\frac{1}{r}} \left(\sup_{0\leq s\leq t}\int_{B_{(1+\sigma)\rho}} u^r(x,s)dx\right)^{\frac{r-1}{r}}. \end{aligned} \tag{30}$$

Fix $R > 0$ and consider the sequence of radii

$$\rho_n = R\sum_{i=1}^{n} 2^{-i}.$$

Select σ_n so that $\rho_{n+1} = (1+\sigma_n)\rho_n$. Then

$$\sigma_n = \frac{\rho_{n+1} - \rho_n}{\rho_n} \geq 2^{-n-2}. \tag{31}$$

Set

$$Y_n = \sup_{0 \leq s \leq t} \int_{B_{\rho n}} u^r(x,s)dx.$$

Then, by (30) and (31),

$$Y_n \leq \int_{B_{2R}} u_0^r dx + \gamma 2^n \left(\frac{t^r}{\rho^{\kappa_r}}\right)^{\frac{1}{r}} Y_{n+1}^{\frac{r-1}{r}}.$$

To complete the proof, use an interpolation argument analogous to that in Proposition 2.1. □

COROLLARY 2.4. *If* $u_0 \in L^r(\mathbb{R}^N)$ *and* $r > \max\{1, \frac{N}{2}\}$, *then there is a constant* γ *depending only on* N *and* r *such that*

$$\sup_{0 \leq s \leq t} \int_{\mathbb{R}^N} u^r(x,s)dx \leq \gamma \int_{\mathbb{R}^N} u_0^r(x)dx. \tag{32}$$

Remark 2.5. Note that a closer analysis shows that we can take $\gamma = 1$ in (32).

2.3. Global estimates.

PROPOSITION 2.6. *Let* u *be a classical solution to* (9) *and let* $r > \max\{1, \frac{N}{2}\}$. *Then there is a constant* γ, *depending only on* r *and* N, *such that*

$$\sup_{B_\rho(x_0)} u(\cdot,t) \leq \gamma \left(t^{-\frac{N}{\kappa_r}} \|u_0\|_{r,B_{4\rho}(x_0)}^{\frac{2r}{\kappa_r}} + \frac{t}{\rho^2}\right). \tag{33}$$

Furthermore, if

$$|||u_0|||_r = \sup_{\rho > 1} \fint_{B_\rho} [|x|^2 u_0(x)]^r dx < \infty, \tag{34}$$

then, for all $t > 0$,

$$|||u(\cdot,t)|||_r < \infty \tag{35}$$

and, for all $x \in \mathbb{R}^N$,

$$|x|^2 u(x,t) \leq \gamma \left(t + t^{-\frac{N}{\kappa_r}} |||u_0|||_r^{\frac{2r}{\kappa_r}}\right). \tag{36}$$

Remark 2.7. For $N = 2$, estimate (33) is sharp in the sense that no estimate of this type is possible for $r = 1$. To see this, consider the one parameter family of solutions u_λ given in Example 1.3. As can easily be checked, $\|u_\lambda(\cdot,0)\|_{1,\mathbb{R}^2}$ is independent of λ, whereas $u_\lambda(0,\frac{T}{2}) = \frac{4T}{\lambda}$.

Remark 2.8. For $N \geq 3$, estimate (33) is almost sharp in the sense that no estimate of this type is possible for $r < \frac{N}{2}$. To see this, consider the one-parameter family of solutions given by example (1.5). For $r < \frac{N}{2}$, $\|u_\lambda(\cdot,0)\|_{r,B_1}$ is bounded independently of λ, whereas $u_\lambda(0,\frac{T}{2}) = \frac{\gamma}{\lambda}$ for some fixed number γ.

Proof of Proposition 2.6. (33) follows from Propositions 2.1 and 2.2. Fix $x_0 \in \mathbb{R}^N$ and $t > 0$. If $|x_0| < 2$, then (36) is obvious. So assume $|x_0| > 2$. Then, from (33),

$$\begin{aligned} |x_0|^2 u(x_0,t) &\leq |x_0|^2 \gamma \left(t^{-\frac{N}{\kappa_r}} \left[\int_{B_{\frac{|x_0|}{2}}(x_0)} u_0^r(x)dx \right]^{\frac{2}{\kappa_r}} + \frac{t}{|x_0|^2} \right) \\ &= \gamma \left(t^{-\frac{N}{\kappa_r}} \left[\frac{1}{|x_0|^N} \int_{B_{\frac{|x_0|}{2}}(x_0)} [|x_0|^2 u_0(x)]^r dx \right]^{\frac{2}{\kappa_r}} + t \right) \\ &\leq \gamma \left(t^{-\frac{N}{\kappa_r}} \left[\fint_{B_{2|x_0|}} [|x|^2 u_0(x)]^r dx \right]^{\frac{2}{\kappa_r}} + t \right), \end{aligned}$$

which proves both (35) and (36). □

Let Ω be a bounded domain in $\mathbb{R}^N$ with smooth boundary $\partial\Omega$. For $T > 0$, let $\Omega_T = \Omega \times (0,T)$ and $S_T = \partial\Omega \times (0,T)$. Consider classical solutions to the boundary value problem

$$(37) \qquad \begin{cases} u_t - \Delta \ln u = 0 & \text{in } \Omega_T, \\ u|_{S_T} = g, & \text{where } g \in C^\infty(S_T), \\ u(x,0) = u_0(x), & \text{where } u_0 \in C^\infty(\Omega). \end{cases}$$

We assume here that g and u_0 are strictly positive so that, by the maximum principle, $u > 0$ in Ω_T, and the PDE in (37) is well defined in the classical sense.

PROPOSITION 2.9. *Let u be a classical solution to* (37) *and let* $r > \max\{1, \frac{N}{2}\}$. *There exists a constant γ depending only on N and r such that for all $t > 0$,*

$$(38) \qquad \|u(\cdot,t)\|_{\infty,\Omega} \leq \|g\|_{\infty,S_T} + \gamma t^{-\frac{N}{\kappa_r}} \|u_0\|_{r,\Omega}^{\frac{2r}{\kappa_r}}.$$

The proof is very similar to that of Proposition 2.1. Select $k > \|g\|_{\infty,S_T}$ and multiply the PDE by the test functions

$$(u-k_n)_+^{r-1}\zeta_n^2, \quad \text{where } k_n = k(1-2^{-n}), \quad n = 0,1,2,\ldots.$$

Due to our choice of k, these test functions vanish on S_T, and the cutoff function ζ_n can be chosen to depend only on t. The proof can now be repeated with minor changes.

3. Local lower bounds. In this section, we consider positive classical solutions to (9). In addition, we assume that $u \leq M$ and that u satisfies the semiconvexity inequality,

$$(39) \qquad u_t \leq \frac{u}{t}.$$

Remark 3.1. If u is a classical solution to (37) with $g(x,t) = c > 0$, then the function $W = (\ln u), -\frac{1}{t}$ satisfies

$$W_t - \frac{1}{u}\Delta W = -W\left(W + \frac{2}{t}\right).$$

Clearly, $W \leq 0$ on the lateral portion of Ω_T and for t sufficiently small. If W has a positive maximum at a point P in the interior of Ω_T, then

$$0 < \left(W_t - \frac{1}{u}\Delta W\right)\Big|_P = -W(P)\left(W(P) + \frac{2}{t}\right) < 0.$$

Thus u satisfies (39). Thus any solution to the Cauchy problem which is constructed as the limit of solutions of this type satisfies (39) in the sense of distributions.

The PDE in (9) is well defined for $u > 0$. We will derive some quantitative lower bounds on u. We will do this in two stages. First, we establish some lower bounds in terms of some integral norms of $|\ln u|$, and then we convert these into pointwise estimates.

3.1. Local lower bounds in terms of integral norms.

PROPOSITION 3.2. *Let u be a classical solution of* (9) *satisfying the semiconvexity inequality* (39). *Fix $0 < \epsilon < 1$. Then there is a constant $\gamma(\epsilon)$ depending only on N and ϵ such that, for all $x_0 \in \mathbb{R}^N$ and for all $0 \leq \tau < t$,*

$$(40) \quad -\inf_{B_\rho(x_0)} \ln u(\cdot,t) \leq \max\left\{\gamma(\epsilon)\fint_{B_{2\rho}(x_0)} (\ln u(x,t))_- dx, \quad 1, \quad (1+\epsilon)\ln\frac{\rho^2}{t-\tau}\right\}.$$

Here $\gamma(\epsilon) = \gamma_0\epsilon^{-N(1+\frac{N}{4})}$, where γ_0 depends only on N.

Proof. Fix $0 < \epsilon < 1$. Again, without loss of generality, we set $x_0 = 0$ and $\tau = 0$. If

$$-\inf_{B_\rho(x_0)} \ln u(\cdot,t) \leq \max\left\{1, (1+\epsilon)\ln\frac{\rho^2}{t}\right\},$$

then the proof is complete. So assume the contrary. Set $v = \ln u$, that is, $u = e^v$. Then

$$(41) \qquad \frac{\partial e^v}{\partial t} - \Delta v = 0.$$

For $0 < \sigma \leq 1$, $k < -\max\{1, (1+\epsilon)\ln\frac{\rho^2}{t}\}$, and $\alpha = \frac{\epsilon}{1+\epsilon}$, set

$$\rho_n = \rho(1 + \sigma 2^{-n}) \quad \text{and} \quad k_n = k(1 - \alpha^n).$$

Then $k_{n+1} < k_n$, $k_0 = 0$, and $k_\infty = k$. Also, notice that for our choice of α, $k_{n+1} \leq k_1 = \frac{k}{1+\epsilon}$ for all $n \geq 0$. Furthermore, $\rho_{n+1} < \rho_n$, $\rho_0 = \rho(1+\sigma)$ and $\rho_\infty = \rho$. Then, with $B_n = B_{\rho n}$,

$$B_{\rho(1+\sigma)} \supseteq B_0 \supseteq B_1 \cdots \supseteq B_n \supseteq B_{n+1} \supseteq \cdots \supseteq B_\infty = B_\rho.$$

Let $\zeta_n \in C_0^\infty(B_n)$ be a smooth cutoff function satisfying

$$\begin{cases} 0 \leq \zeta_n \leq 1 & \text{in } B_n, \\ \zeta_n = 1 & \text{in } B_{n+1}, \\ |D\zeta_n|^2, |\Delta\zeta_n| \leq \gamma\dfrac{4^n}{(\sigma\rho)^2} & \text{in } B_n. \end{cases}$$

Multiply (41) by $(k_{n+1}-v)_+\zeta_n^2$ and integrate over B_n.

$$\begin{aligned} 0 &= \int_{B_n} (e^v v_t - \Delta v)(k_{n+1}-v)_+\zeta_n^2 dx \\ &= T_1 + T_2. \end{aligned} \tag{42}$$

$$T_2 = \int_{B_n} Dv\left[\zeta_n^2 D(k_{n+1}-v)_+ + 2(k_{n+1}-v)_+\zeta_n D\zeta_n\right] dx$$

$$\begin{aligned} &= -\int_{B_n} \left[|D(k_{n+1}-v)_+|^2\zeta_n^2 + D(k_{n+1}-v)_+^2\zeta_n D\zeta_n\right] dx \\ &= -\int_{B_n} |D(k_{n+1}-v)_+|^2\zeta_n^2 dx + 2\int_{B_n}(k_{n+1}-v)_+^2(|D\zeta_n|^2+\zeta_n\Delta\zeta_n)dx \\ &\leq -\int_{B_n} |D[(k_{n+1}-v)_+\zeta_n]|^2 dx + \gamma\frac{4^n}{(\sigma\rho)^2}\int_{B_n}(k_{n+1}-v)_+^2 dx. \end{aligned} \tag{43}$$

To estimate T_1, notice that

$$v_t = \frac{u_t}{u} \leq \frac{1}{t}.$$

Thus

$$\begin{aligned} T_1 &\leq \frac{1}{t}\int_{B_n} e^v(k_{n+1}-v)_+\zeta_n^2 dx \\ &\leq \frac{e^{k_{n+1}}}{t}\int_{B_n}(k_{n+1}-v)_+ dx. \end{aligned} \tag{44}$$

Thus, by combining (42)–(44) and using $k_{n+1} \leq k_1 \leq \frac{k}{1+\epsilon}$,

$$\begin{aligned} \int_{B_n} |D[(k_{n+1}-v)_+\zeta_n]|^2 dx &\leq \gamma\frac{4^n}{(\sigma\rho)^2}\int_{B_n}(k_{n+1}-v)_+^2 dx \\ &\quad + \frac{e^{\frac{k}{1+\epsilon}}}{t}\int_{B_n}(k_{n+1}-v)_+ dx. \end{aligned} \tag{45}$$

Notice that

$$\begin{aligned} \int_{B_n}(k_n-v)_+^2 dx &\geq \int_{B_n}(k_{n+1}-v)_+(k_n-v)_+ dx \\ &\geq (k_n-k_{n+1})\int_{B_n}(k_{n+1}-v)_+ dx \\ &\geq |k|\alpha^n(1-\alpha)\int_{B_n}(k_{n+1}-v)_+ dx. \end{aligned} \tag{46}$$

Thus, by combining (45) and (46), we have

$$(47) \qquad \int_{B_n} |D[(k_{n+1}-v)_+\zeta_n]|^2 dx \le \frac{\gamma}{(\sigma\rho)^2\alpha^n} \int_{B_n} (k_n - v)_+^2 dx.$$

Notice here we used the assumption that $-k > \max\{1, (1+\epsilon)\ln\frac{\rho^2}{t}\}$ to get that

$$\frac{e^{\frac{k}{1+\epsilon}}(1+\epsilon)\rho^2}{t|k|} \le 2.$$

By applying Hölder's inequality, the Sobolev inequality, and (47), we get

$$\begin{aligned}
\int_{B_{n+1}} (k_{n+1}-v)_+^2 dx &\le \int_{B_n} (k_{n+1}-v)_+^2 \zeta_n^2 dx \\
&\le \left(\int_{B_n} [(k_{n+1}-v)_+\zeta_n]^{\frac{2N}{N-2}} dx\right)^{\frac{N-2}{N}} |A_n(t)|^{\frac{2}{N}} \\
(48) \qquad &\le \gamma \int_{B_n} |D[(k_{n+1}-v)_+\zeta_n]|^2 dx |A_n(t)|^{\frac{2}{N}} \\
&\le \frac{\gamma b^n}{(\sigma\rho)^2} |A_n(t)|^{\frac{2}{N}} \int_{B_n} (k_n - v)_+^2 dx,
\end{aligned}$$

where $A_n(t) = \{x \in B_n \,|\, v(x,t) < k_{n+1}\}$. Note here that $b > 1$ is some constant depending on α. To estimate $A_n(t)$

$$\begin{aligned}
(49) \qquad \int_{B_n} (k_n - v)_+^2 dx &\ge |A_n(t)|(k_{n+1}-k_n)^2 \\
&= k^2\alpha^{2n}(1-\alpha)^2|A_n(t)|.
\end{aligned}$$

From (48) and (49),

$$(50) \qquad \int_{B_{n+1}} (k_{n+1}-v)_+^2 dx \le \gamma \frac{b^n}{(\sigma\rho)^2 k^{\frac{4}{N}}} \left(\int_{B_n} (k_n - v)_+^2 dx\right)^{1+\frac{2}{N}}.$$

Set

$$Y_n -\!\!\!\!\!\!\int_{B_n} (k_n - v)_+^2 dx.$$

Then, from (50),

$$Y_{n+1} \le \gamma \frac{b^n}{\sigma^2 |k|^{\frac{4}{N}}} Y_n^{1+\frac{2}{N}}.$$

Then, as in the proof of Proposition 2.1, $Y_n \to 0$ as $n \to \infty$ if

$$-k > \gamma \frac{b^{\frac{N^2}{8}}}{\sigma^{\frac{N}{2}}} Y_0^{\frac{1}{2}}.$$

Thus

$$-\inf_{B_\rho} v(x,t) \leq \frac{\gamma}{\sigma^{\frac{N}{2}}} \left(\fint_{B_{(1+\sigma)\rho}} v_-^2(x,t)dx \right)^{\frac{1}{2}}$$

$$\leq \frac{\gamma}{\sigma^{\frac{N}{2}}} \left(\sup_{B_{(1+\sigma)\rho}} v_-(\cdot,t) \right)^{\frac{1}{2}} \left(\fint_{B_{(1+\sigma)\rho}} v_-(x,t)dx \right)^{\frac{1}{2}}.$$

Consider the sequence of radii given by

$$\rho_n = \rho \sum_{j=0}^{n} 2^{-j}.$$

Select σ_n so that $\rho_{n+1} = (1+\sigma_n)\rho_n$. Then $\sigma_n \geq 2^{-n-1}$. Thus, with $M_n = \sup_{B_{\rho_n}} v_-(\cdot,t)$,

$$M_{n+1} \leq \gamma b^n M_n^{\frac{1}{2}} \left(\fint_{B_{(1+\sigma)\rho}} v_-(x,t)dx \right)^{\frac{1}{2}}.$$

To complete the proof, use an interpolation argument analogous to that in Proposition 2.1 to arrive at

$$-\inf_{B_\rho} v(x,t) \leq \gamma \fint_{B_{2\rho}} v_-(x,t)dx$$

when $-\inf_{B_\rho} v(x,t) > \max\{1, (1+\epsilon)\ln \frac{\rho^2}{t}\}$. This proves the proposition. □

3.2. A representation formula. Fix $x_0 \in \mathbb{R}^N$ and consider the Green's-type function for the Laplacian in the ball $B_\rho(x_0)$,

(51)

$$G(|x-x_0|;\rho) = \begin{cases} |x-x_0|^{2-N} - \rho^{2-N} + \dfrac{N-2}{2}\rho^{-N}(|x-x_0|^2 - \rho^2), & N \geq 3, \\ \ln\rho - \ln|x-x_0| + \dfrac{1}{2}\rho^{-2}(|x-x_0|^2 - \rho^2), & N = 2, \\ \rho - |x-x_0| + \dfrac{1}{2}\rho^{-1}(|x-x_0|^2 - \rho^2), & N = 1. \end{cases}$$

One can easily verify that

$$G(|x-x_0|;\rho)|_{\partial B_\rho(x_0)} = 0$$

and that

$$DG(|x-x_0|;\rho) \cdot \frac{x-x_0}{\rho}|_{\partial B_\rho(x_0)} = 0.$$

Moreover,

$$0 \leq G(|x-x_0|;\rho) \leq \begin{cases} |x-x_0|^{2-N} - \rho^{2-N}, & N \geq 3, \\ \ln\rho - \ln|x-x_0|, & N = 2, \\ \rho - |x-x_0|, & N = 1. \end{cases}$$

By direct calculation,

$$\Delta G(|x-x_0|;\rho)=\begin{cases} -\omega_N(N-2)\delta_{x_0}+N(N-2)\rho^{-N}, & N\geq 3,\\ -2\pi\delta_{x_0}+2\rho^{-2}, & N=2,\\ -2\delta_{x_0}+\rho^{-1}, & N=1,\end{cases}$$

where ω_N is the area of the unit sphere in $\mathbb{R}^N$ and δ_{x_0} is the Dirac delta measure with mass centered at x_0. Thus, for all $f\in C^2(\overline{B}_\rho(x_0))$,

$$\text{(52)}\qquad ⨍_{B_\rho(x_0)} f(x)dx = f(x_0)+C_N\int_{B_\rho(x_0)}\Delta f(x)G(|x-x_0|;\rho)dx,$$

where

$$C_N=\begin{cases}\dfrac{1}{\omega_N(N-2)}, & N\geq 3,\\ \dfrac{1}{2\pi}, & N=2,\\ \dfrac{1}{2}, & N=1.\end{cases}$$

Remark 3.3. Pointwise representations, such as (52), have been used by Gilding and Peletier (see [9]) in the context of the continuity of solutions to the porous-medium equation.

Suppose that u is a classical solution to (9) satisfying $u\leq M$. In (52), take $f(x)=\ln u(x,s)$. Then

$$\text{(53)}\qquad ⨍_{B_\rho(x_0)}\ln\frac{u(x,s)}{u(x_0,s)}dx = C_N\int_{B_\rho(x_0)} G(|x-x_0|;\rho)u_t(x,s)dx$$

for all $x_0\in\mathbb{R}^N$ and $s>0$.

From (39), the function

$$s\longrightarrow \ln\frac{u(x,s)}{s}$$

is decreasing. So, by integrating (53) from $\tau\leq s\leq t$,

$$\text{(54)}\qquad ⨍_{B_\rho(x_0)}\ln\frac{tu(x,\tau)}{\tau u(x_0,t)}dx\geq\frac{C_N}{t-\tau}\int_{B_\rho(x_0)}G(|x-x_0|;\rho)\int_\tau^t u_t(x,s)dsdx.$$

By appealing to Proposition 3.2 and selecting ρ so large so that

$$\frac{\rho^2}{\tau}>e,\quad\text{i.e.,}\quad \rho>\sqrt{e\tau},$$

we have, for $x\in B_\rho(x_0)$,

$$\text{(55)}\qquad \ln\frac{1}{u(x,\tau)}\leq\gamma(\epsilon)⨍_{B_{2\rho}(x_0)}\left(\ln\frac{1}{u(y,\tau)}\right)_+dy+(1+\epsilon)\ln\frac{\rho^2}{\tau}.$$

By combining this with (54), we arrive at the following proposition.

PROPOSITION 3.4. *There is a constant $\gamma(\epsilon)$, depending only on N and ϵ, such that, for all $x_0 \in \mathbb{R}^N$, $0 < \tau < t$, $\rho > \sqrt{e\tau}$, and $x \in B_\rho(x_0)$,*

$$\begin{aligned}\ln\frac{1}{u(x,\tau)} \le \frac{-\gamma(\epsilon)}{t-\tau}\int_{B_{2\rho}(x_0)} G(|y-x_0|;2\rho)\int_\tau^t u_t(y,s)ds dy\\ + \gamma(\epsilon)\ln\frac{Mt}{\tau u(x_0,t)} + (1+\epsilon)\ln\rho^2.\end{aligned} \tag{56}$$

Proof. By combining (54) and (55),

$$\begin{aligned}\ln\frac{1}{u(x,\tau)} &\le \frac{-\gamma}{t-\tau}\int_{B_\rho(x_0)} G(|y-x_0|;\rho)\int_\tau^t u_t(y,s)dsdy\\ &\quad + \gamma \fint_{B_{2\rho}(x_0)} (\ln u(y,\tau))_+ dy + (1+\epsilon)\ln\frac{\rho^2}{\tau} + \gamma\ln\frac{t}{\tau u(x_0,t)}\\ &\le \frac{-\gamma}{t-\tau}\int_{B_\rho(x_0)} G(|y-x_0|;\rho)\int_\tau^t u_t(y,s)dsdy\\ &\quad + \gamma\ln\frac{Mt}{\tau u(x_0,t)} + (1+\epsilon\ln\rho^2). \qquad \square\end{aligned}$$

3.3. Pointwise lower bounds. Starting from (56), we estimate

$$\begin{aligned}&-\frac{C_N}{t-\tau}\int_{B_{2\rho}(x_0)} G(|x-x_0|;2\rho)\int_\tau^t u_t(x,s)dsdx\\ &\le \frac{C_N}{t-\tau}\int_{B_{2\rho}(x_0)} G(|x-x_0|;2\rho)u(x,\tau)dx\\ &= I(x_0,\rho,t,\tau).\end{aligned}$$

With this notation and Proposition 3.4,

$$u(x,\tau) \ge \frac{1}{\rho^{2+2\epsilon}}\left(\frac{\tau u(x_0,t)}{Mt}\right)^\gamma e^{-\gamma I(x_0,\rho,t,\tau)} \tag{57}$$

for all

$$\rho \ge \sqrt{e\tau},$$

$x \in B_\rho(x_0)$, and $0 < \tau < t$. Note here that the constant γ still depends on ϵ.

Remark 3.5. If $u(x_0,t) > 0$, for some $x_0 \in \mathbb{R}^N$ and $t > 0$, then (57) guarantees that $u(x,\tau) > 0$ for all $x \in \mathbb{R}^N$ and all $0 < \tau < t$. This can be regarded as a global Harnack-type estimate.

3.4. Asymptotic behavior as $|x| \to \infty$: The case $N \ge 3$.

PROPOSITION 3.6. *Suppose that $N \ge 3$ and that u is a classical solution to* (9) *in $\mathbb{R}^N \times (0,T)$ satisfying* (39). *Suppose that, for some $0 < \tau < T$, $\alpha > 0$, $\gamma_0 > 0$, and $R > 0$,*

$$u(x,\tau) \le \frac{\gamma_0}{|x|^\alpha} \quad \textit{for } |x| > R. \tag{58}$$

Then

$$\lim_{\rho\to\infty}\int_{B_\rho} G(|x|;\rho)u(x,\tau)dx = \infty. \tag{59}$$

Proof. Let $R_0 < 0$ satisfy

$$\begin{cases} \ln\dfrac{R_0^2}{\tau} > 2\ln\dfrac{\gamma_0^{\frac{2}{\alpha}}}{\tau}, \\ R_0^\alpha > \gamma_0, \\ R_0 > \sqrt{e\tau}, \\ R_0 > R. \end{cases}$$

Notice here that R_0 depends on γ_0, α, R, and τ. Let $\rho > R_0$. Then

$$\begin{aligned} \fint_{B_{2\rho}} (\ln u(x,\tau))_- dx &\ge -\frac{\gamma}{\rho^N}\int_{\mathrm{Ann}(\rho;2\rho)} \ln u(x,\tau)dx \\ &\ge \gamma\left(\ln\frac{\rho^2}{\tau} - \ln\frac{\gamma_0^{\frac{2}{\alpha}}}{\tau}\right) \\ &\ge \gamma\ln\frac{\rho^2}{\tau}. \end{aligned}$$

Here

$$\mathrm{Ann}(\rho,2\rho) = \{x \in \mathbb{R}^N \mid \rho \le |x| \le 2\rho\}.$$

So, for $\rho > R_0$, there is a $\gamma_1 > 0$ not depending on ρ such that

$$\fint_{B_{2\rho}} (\ln u(x,\tau))_- dx \ge \gamma_1 \ln\frac{\rho^2}{\tau}.$$

Thus, from Proposition 3.2 with

$$\gamma \ge \max\left\{\gamma(1), \frac{2}{\gamma_1}\right\},$$

we have

$$-\inf_{B_\rho} \ln u(x,\tau) \le \gamma \fint_{B_{2\rho}} (\ln u(x,t))_- dx \tag{60}$$

for all $\rho > R_0$. This constant γ now depends on τ, N, α, R_0, and γ_0. Select t so that $\tau < t < T$. Then we can repeat the arguments beginning with equation (54) using (60) instead of (40). In so doing, we arrive at the following analogue of (57):

$$u(x,\tau) \ge \left(\frac{\tau u(0,t)}{Mt}\right)^\gamma e^{-\gamma I(0,\rho,t,\tau)} \tag{61}$$

for all $\rho > R_0$ and $|x| \le \rho$.

Now $I(0,\rho,t,\tau)$ is an increasing function of ρ. If (59) does not hold, then there is some constant C independent of ρ such that

$$u(x,\tau) \geq C\left(\frac{\tau u(0,t)}{Mt}\right)^{\gamma}$$

for all $x \in \mathbb{R}^N$. This contradicts (58). □

Remark 3.7. Proposition 3.6 is still true for $N = 1$ or 2. But, when $N = 1$ or 2, (59) is true for any $u(\cdot, t)$ unless $u(\cdot, t) = 0$ almost everywhere.

Assume that the initial data u_0 satisfies (34). In view of Proposition 2.6, the corresponding solution decays as $|x| \to \infty$ at least as fast as $|x|^{-2}$. Proposition 3.6 guarantees that $u(x,t)$, in some sense, decays no faster than $|x|^{-2}$. This is meant in the sense that a behavior of the type

$$u(x,\tau) \leq \frac{\gamma(\tau)}{(1+|x|^2)[\ln|x|]^{1+\epsilon}} \quad \text{for } |x| > 1$$

is impossible for any $\epsilon > 0$. Indeed, if behavior of this type did occur, then (58) would be satisfied, whereas (59) would not be satisfied for $0 < \tau < t$.

3.5. Asymptotic behavior as $|x| \to \infty$: The case $N = 2$. As can easily be checked,

$$\lim_{\rho\to\infty} \int_{B_\rho} \frac{G(|x-x_0|;\rho)}{(1+|x|^2)^k} dx = \infty \quad \text{for all } k > 0.$$

Thus, contrary to the case $N \geq 3$, there might exist solutions decaying faster than $|x|^{-2}$ as $|x| \to \infty$. Of course, this agrees with the explicit solution given by Example 1.2.

3.6. Some ill-posed problems for $N \geq 3$. Suppose that the initial data u_0 satisfies (34) and

$$\int_{\mathbb{R}^N} \frac{u_0(x)}{(1+|x|)^{N-2}} dx < \infty.$$

We claim that such a problem can have no solution satisfying (39). First, from Proposition 2.6, we know that any solution to (9) with initial data satisfying (34) must satisfy

$$u(x,t) \leq \frac{\gamma\left(t + t^{-\frac{N}{\kappa r}}|||u_0|||_r\right)^{\frac{2r}{\kappa r}}}{|x|^2}.$$

Now, if we had a solution in $\mathbb{R}^N \times (0,T)$ for some $T > 0$, then let $(x_0, t) \in \mathbb{R}^N \times (0,T)$ be a point such that $u(x_0,t) > 0$. Up to rescaling, we may assume that $u \leq 1$ in $\mathbb{R}^N \times (0,T)$. So $\ln u(x,t) \leq 0$. Then

$$\begin{aligned}\fint_{B_\rho(x_0)} \ln\frac{u(x_0,s)}{u(x,s)} dx &= \fint_{B_\rho(x_0)} \left(\ln\frac{u(x_0,s)}{s} + \ln s - \ln u(x,s)\right) dx \\ &\geq \ln\frac{u(x_0,s)}{s} + \ln s.\end{aligned}$$

Then, by applying (53) and integrating in ds from $0 < s < \tau$, we have

$$\begin{aligned}C_N \int_{B_\rho(x_0)} G(|x-x_0|;\rho) u(x,\tau) dx \leq C_N \int_{B_\rho(x_0)} & G(|x-x_0|;\rho) u_0(x) dx \\ &+ \tau(I - \ln u(x_0,\tau)).\end{aligned}$$

Thus

$$\lim_{\rho\to\infty}\int_{B_\rho(x_0)} G(|x-x_0|;\rho)u(x,\tau)dx < \infty$$

for all $0<\tau<t$. This, of course, contradicts Proposition 3.6. In particular, if $N\ge 3$ and $u_0\in C_0^\infty(\mathbb{R}^N)$, then (9) can have no classical solution satisfying (39). A weaker result, that (9) can have no solution when $u_0\in L^1(\mathbb{R}^N)N\ge 3$, appears in [14].

4. Finite extinction time. In this section, we examine classical solutions of (9) with a finite extinction time, i.e., solutions which become identically zero at some time T. In particular, we find sufficient conditions on the initial data to guarantee a finite extinction time. In addition, we find an upper bound for this extinction time. We have several examples of solutions with finite extinction times. See Examples 1.1, 1.3, and 1.5.

THEOREM 4.1. *Let $N\ge 3$ and $u_0\in L^{\frac{N}{2}}(\mathbb{R}^N)$. Then if u is a classical solution to (9) there is a $T>0$ such that $u(x,T)=0$ for all $x\in\mathbb{R}^N$. Furthermore,*

$$T\le\gamma(N)\|u_0\|_{\frac{N}{2},\mathbb{R}^N}.$$

Proof. Let $\zeta\in C_0^\infty(B_{2\rho})$ satisfy

$$\begin{cases} 0\le\zeta\le 1 & \text{in } B_{2\rho},\\ \zeta=1 & \text{in } B_\rho,\\ |D\zeta|^2, |\Delta\zeta|\le\dfrac{\gamma}{\rho^2} & \text{in } B_{2\rho}.\end{cases}$$

Multiply the first term of (9) by the test function $u^r\zeta^2$.

$$\begin{aligned} 0 &= \int_{\mathbb{R}^N}(u_t-\Delta\ln u)u^r\zeta^2dx\\ &= \frac{1}{r+1}\frac{d}{dt}\left[\int_{\mathbb{R}^N}u^{r+1}\zeta^2dx\right]+\int_{\mathbb{R}^N}\frac{Du}{u}(ru^{r-1}Du\zeta^2+2u^r\zeta D\zeta)dx \qquad (62)\\ &= T_1+T_2+T_3. \end{aligned}$$

$$\begin{aligned} T_2 &= \int_{\mathbb{R}^N}|Du|^2u^{r-2}\zeta^2dx\\ &\ge\frac{4}{r}\int_{\mathbb{R}^N}|D[u^{\frac{r}{2}}\zeta]|^2dx-\frac{\gamma}{\rho^2}\int_{\mathrm{An}(\rho,2\rho)}u^rdx. \end{aligned}\tag{63}$$

Here $\mathrm{An}(\rho,2\rho)=\{x\in\mathbb{R}^N\,|\,\rho\le|x|\le 2\rho\}$.

$$\begin{aligned} T_3 &= \int_{\mathbb{R}^N}Duu^{r-1}\zeta D\zeta dx\\ &= \frac{1}{r}\int_{\mathbb{R}^N}Du^r\zeta D\zeta dx\\ &= -\frac{1}{r}\int_{\mathbb{R}^N}u^r(\zeta\Delta\zeta+|D\zeta|^2)dx. \end{aligned}\tag{64}$$

From (62)–(64), we obtain

$$\frac{d}{dt}\left[\int_{\mathbb{R}^N}u^{r+1}\zeta^2dx\right]+\int_{\mathbb{R}^N}|D[u^{\frac{r}{2}}\zeta]|^2dx\le\frac{\gamma}{\rho^2}\int_{\mathrm{An}(\rho,2\rho)}u^rdx.\tag{65}$$

Since ζ vanishes on the boundary of $B_{2\rho}$, we can apply the Sobolev Embedding Theorem, i.e.,

$$\gamma(N,r)\left(\int_{B_{2\rho}}[u^{\frac{r}{2}}\zeta]^{\frac{2N}{N-2}}dx\right)^{\frac{N-2}{N}} \leq \int_{B_{2\rho}}|D[u^{\frac{r}{2}}\zeta]|^2dx. \tag{66}$$

Thus, with $r = \frac{N-2}{2}$,

$$\gamma(N)\left(\int_{B_{2\rho}}u^{\frac{N}{2}}\zeta^{\frac{2N}{N-2}}dx\right)^{\frac{N-2}{N}} \leq \int_{B_{2\rho}}|D[u^{\frac{r}{2}}\zeta]|^2dx. \tag{67}$$

From (65)–(67),

$$\frac{d}{dt}\left[\int_{B_{2\rho}}u^{\frac{N}{2}}\zeta^2dx\right] + \gamma_0\left[\int_{B_{2\rho}}u^{\frac{N}{2}}\zeta^{\frac{2N}{N-2}}dx\right]^{\frac{N-2}{N}} \leq \frac{\gamma}{\rho^2}\int_{\mathrm{An}(\rho,2\rho)}u^{\frac{N-2}{2}}dx.$$

By Hölder's inequality,

$$\int_{\mathrm{An}(\rho,2\rho)}u^{\frac{N-2}{2}}dx \leq \gamma\rho^2\left(\int_{\mathrm{An}(\rho,2\rho)}u^{\frac{N}{2}}dx\right)^{\frac{N-2}{N}}.$$

Thus

$$\frac{d}{dt}\left[\int_{B_{2\rho}}u^{\frac{N}{2}}\zeta^2dx\right] + \gamma(N)\left[\int_{B_{2\rho}}u^{\frac{N}{2}}\zeta^{\frac{2N}{N-2}}dx\right]^{\frac{N-2}{N}} \geq \gamma\left[\int_{\mathrm{An}(\rho,2\rho)}u^{\frac{N}{2}}dx\right]^{\frac{N-2}{N}}.$$

From Proposition 2.2, if $u_0 \in L^{\frac{N}{2}}(\mathbb{R}^N)$ then $u(\cdot,t) \in L^{\frac{N}{2}}(\mathbb{R}^N)$ for each $t > 0$. Thus, by letting $\rho \to \infty$, we have

$$\frac{df}{dt} + \gamma(N)f^{\frac{N-2}{N}} \leq 0,$$

where

$$f(t) = \int_{\mathbb{R}^N}u^{\frac{N}{2}}(x,t)dx.$$

As long as $f > 0$, this imples

$$f^{\frac{2}{N}}(t) \leq f^{\frac{2}{N}}(0) - \gamma(N)t.$$

Thus we have

$$u(\cdot,t) = 0 \quad \text{if} \quad t \geq \gamma(N)\|u_0\|_{\frac{N}{2},\mathbb{R}^N}.$$

This is the desired result. □

Theorem 4.1 is sharp in the sense that no theorem of this type is possible for $L^r(\mathbb{R}^N)$ for any $r > \frac{N}{2}$. To see this, let $0 < \alpha < 2$, and consider the following initial datum:

$$u_0(x) = \frac{1}{1+|x|^\alpha}.$$

For a wise choice of α, $u_0 \in L^r(\mathbb{R}^N)$ for some fixed $r > \frac{N}{2}$. Choose $\beta > \frac{2}{2-\alpha}$. By Young's inequality, we can show that there is a $C > 0$, depending only on α, β, and N, such that, for any $T > C$,

$$\frac{2(N-2)T}{T^\beta + |x|^2} \leq u_0(x).$$

For $T > C$, consider the explicit solution (see Example 1.5)

$$u_T(x,t) = \frac{2(N-2)(T-t)_+^{\frac{N}{N-2}}}{T^{\beta+\frac{2}{N-2}} + (T-t)_+^{\frac{2}{N-2}}|x|^2}.$$

Then

$$\begin{aligned} u_T(x,0) &= \frac{2(N-2)T}{T^\beta + |x|^2} \\ &\leq u_0(x). \end{aligned}$$

This implies that for each $T > C$ there exists a classical solution to (9) with initial data $u_0(x) = \frac{1}{1+|x|^\alpha}$ which does not vanish at least until time T. Theorem 4.1, however, is not entirely satisfactory because it does not predict an extinction time for the solutions given by Example 1.5.

REFERENCES

[1] G. I. Barenblatt, *On some unsteady motions of a liquid or a gas in a porous medium*, Prikl. Mat. Mekh., 16 (1952), pp. 67–78.

[2] P. Bènilan and M. G. Crandall, *The continuous dependence on ϕ of solutions of $u_t - \Delta\phi(u) = 0$*, Indiana Univ. Math. J., 30 (1981), pp. 161–177.

[3] A. Bertozzi, M. Brenner, T. DuPont, and L. Kadanoff, *Singularities and similarities in interface flows*, in Trends and Perspectives in Applied Mathematics, Springer-Verlag, New York, 1994, pp. 155–208.

[4] J. P. Burelbach, S. G. Bankoff, and S. H. Davis, *Nonlinear stability of evaporating/condensing liquid films*, J. Fluid Mech., 195 (1988), pp. 463–494.

[5] J. T. Chayes, S. J. Osher, and J. V. Ralston, *On singular diffusion equations with application to self-organized criticality*, Comm. Pure Appl. Math., 46 (1993), pp. 1363–1377.

[6] H. E. Conner, *Some general properties of a class of semilinear hyperbolic systems analogous to the differential-integral equations of gas dynamics*, J. Differential Equations, 10 (1971), pp. 188–203.

[7] E. DiBenedetto, *Degenerate Parabolic Equations*, Universitext series, Springer-Verlag, New York, 1993.

[8] E. DiBenedetto and Y. C. Kwong, *Intrinsic Harnack estimates and extinction profile for certain singular parabolic equations*, Trans. Amer. Math. Soc., 330 (1992), pp. 783–811.

[9] B. Gilding and L. A. Peletier, *Continuity of solutions of the porous medium equations*, Ann. Scuola. Norm. Sup. Pisa, 8 (1981), pp. 659–675.

[10] R. S. Hamilton, *The Ricci flow on surfaces*, Contemp. Math., 71 (1988), pp. 237–262.

[11] M. A. Herrero, *A limit case in nonlinear diffusion*, Nonlinear Anal., 13 (1989), pp. 611–628.

[12] T. G. Kurtz, *Convergence of sequences of semigroups of nonlinear operators with an application to gas kinetics*, Trans. Amer. Math. Soc., 186 (1973), pp. 259–272.

[13] R. E. Pattle, *Diffusion from an instantaneous point source with a concentration dependent coefficient*, Quart. J. Appl. Math., 12 (1959), pp. 407–409.

[14] J. L. Vazquez, *Nonexistence of solutions for nonlinear heat equations of fast diffusion type*, J. Math. Pures Appl., 71 (1992), pp. 503–526.

[15] M. B. Williams and S. H. Davis, *Nonlinear theory of film rupture*, J. Colloidal and Interface Sci., 90 (1982), pp. 220–228.

[16] L. Wu, *The Ricci flow on the complete* $\mathbb{R}^2$, Comm. Anal. Geom., 1 (1993), pp. 439–472.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 661–683, May 1996

004

ON THE DIRICHLET BOUNDARY VALUE PROBLEM FOR A DEGENERATE PARABOLIC EQUATION*

T. KILPELÄINEN[†] AND P. LINDQVIST[‡]

Abstract. The Perron method for degenerate parabolic equations like $u_t = \operatorname{div}(|\nabla u|^{p-2}\nabla u)$ is studied. The regular boundary points for the Dirichlet problem are characterized in terms of barriers. In the particular case of the space–time cylinder $G \times (0, T)$, a geometric characterization in terms of a Wiener-type test is given for regularity.

Key words. parabolic p-Laplacian, the Perron method, regular boundary points

AMS subject classifications. 35K65, 35K60, 31C45

1. Introduction.

It was observed by Sternberg in 1929 that the method developed by Perron [Pn] for solving the Dirichlet boundary value problem for the Laplace equation can readily be extended to the heat equation. It is nowadays well known that the Perron method applies to linear parabolic equations; cf. [F]. In this paper, we study potential theoretic aspects of certain nonlinear parabolic equations; one of our aims is to adapt the Perron method. The typical example that we have in mind is the p-parabolic equation

$$\frac{\partial u}{\partial t} = \operatorname{div}(|\nabla u|^{p-2}\nabla u),$$

where $1 < p < \infty$. This equation is degenerate when $2 < p < \infty$ and singular when $1 < p < 2$. For the regularity theory of such equations, the reader is asked to consult the recent monograph by DiBenedetto [DB].

In the elliptic case, the study of nonlinear potential theory of quasi-linear equations of the type

$$\operatorname{div}(|\nabla u|^{p-2}\nabla u) = 0$$

was initiated by Granlund, Lindqvist, and Martio in a series of papers, of which we mention only [GLM]. An account of the elliptic nonlinear potential theory is given in the monograph [HKM]. Needless to say, some parts of elliptic theory can be carried over to the parabolic situation as such, while for others, parabolic proofs present new difficulties.

In order to keep the presentation within reasonable limits, we have made a few simplifying assumptions. The most noteworthy of them is that we treat the Perron method only for bounded boundary values; this is not a serious restriction. We have not included all of our results in this short outline; for example, the counterpart to the celebrated condition of Petrowsky will be published elsewhere. On the other hand, there are some basic questions that we have not been able to answer yet. One urgent open problem is the parabolic counterpart to Wiener's resolutivity theorem; see [W, Thm. 32, p. 290] for the heat equation.

The paper is organized as follows. Section 2 is merely a discussion about the choice of equations and the admissible extensions for the theory. Section 3 contains

*Received by the editors June 15, 1994; accepted for publication October 3, 1994.

[†] Department of Mathematics, University of Jyväskylä, P.O. Box 35, 40351 Jyväskylä, Finland (terok@math.jyu.fi).

[‡] Department of Mathematics, Norwegian Institute of Technology, N-7034 Trondheim, Norway (lqvist@imf.unit.no).

basic definitions and expedient existence theorems. The p-superparabolic functions are introduced in §4. Finally, §5 contains the Perron method; barriers and regularity are discussed in §6.

2. Discussion. A genuine nonlinear parabolic potential theory should include equations like

$$\text{(I)} \qquad \frac{\partial u}{\partial t} = \operatorname{div}(|\nabla u|^{p-2}\nabla u),$$

$$\text{(II)} \qquad \frac{\partial u}{\partial t} = \Delta(|u|^{m-1}u),$$

and

$$\text{(III)} \qquad \frac{\partial(|u|^{p-2}u)}{\partial t} = \operatorname{div}(|\nabla u|^{p-2}\nabla u).$$

They reduce to the ordinary heat equation $u_t = \Delta u$ for $p = 2$ and $m = 1$. On the other hand, they all are special cases[1] of equations of the type

$$\frac{\partial u}{\partial t} = \operatorname{div}(|u|^{m-1}|\nabla u|^{p-2}\nabla u),$$

and so one would be led to construct a potential theory for a wide class of equations

$$\frac{\partial u}{\partial t} = \operatorname{div}\mathcal{A}(x, t, u, \nabla u).$$

In order to keep the presentation short and direct, we have modeled our approach according to equations of the p-parabolic type

$$\frac{\partial u}{\partial t} = \operatorname{div}\mathcal{A}_p(x, t, \nabla u),$$

where $\mathcal{A}_p(x,t,w) \approx |w|^{p-2}w$ and $1 < p < \infty$. This has the advantage that constants can be added to solutions. The precise assumptions about $\mathcal{A}_p$ are listed in the beginning of §3.

It is worth mentioning that, from a potential theoretic point of view, the third equation seems to be the most natural one, for there solutions can be multiplied with constants. As Trudinger has pointed out in [T, p. 225], the parabolic Harnack inequality holds for equation (III) in the same form as it does for the ordinary heat equation; see [M]. This is not the case for the p-parabolic equation (I); its Harnack estimate needs an *intrinsic* formulation depending on the solution in question; cf. [DB]. One of the reasons for basing our approach on the usual p-parabolic equation, rather than on its homogeneous variant, is that disturbances propagate, as it were, with infinite speed for the latter equation, while the p-parabolic equation enjoys the finite-speed propagation property when $p > 2$ but not when $1 < p \leq 2$. For $1 < p < 2$, there is an extinction phenomenon. We have favored this diversity. (See the note added in proof on p. 682.)

To give the reader a feeling for this fascinating phenomenon, we mention the fundamental solution obtained by Barenblatt in 1952 [B]. The Barenblatt solution to the p-parabolic equation (I) is

$$\mathcal{B}_p(x,t) = t^{-\frac{n}{\lambda}} \left(C - \frac{p-2}{p}\lambda^{\frac{1}{1-p}} \left(\frac{|x|}{t^{1/\lambda}} \right)^{\frac{p}{p-1}} \right)_+^{\frac{p-1}{p-2}}.$$

[1] Substitute $v = |u|^{p-2}u$ in (III).

It is defined for $x \in \mathbf{R}^n$ and $t > 0$ and is compactly supported in x. Here $p > 2$ and $\lambda = n(p-2)+p$. The constant C is usually chosen so that

$$\int \mathcal{B}_p(x,t)\,dx = 1,$$

i.e., so that $\mathcal{B}_p(x, 0+) = \delta(x)$, the Dirac delta function. When $p \to 2+$ this approaches the ordinary heat kernel,

$$\frac{1}{(4\pi t)^{n/2}} e^{-\frac{|x|^2}{4t}}.$$

However, the Barenblatt solution to the equation (III) is

$$\mathcal{W}_p(x,t) = Ct^{-\frac{n}{p(p-1)}} \exp\left(-\frac{p-1}{p}\left(\frac{|x|^p}{pt}\right)^{\frac{1}{p-1}}\right),$$

$1 < p < \infty$. It does not have a bounded support.[2]

Another point is that, if u is a solution to equation (III), then the function $v = |u|^k$ is a subsolution to the same equation when $p \geq 2$ and $k \geq 1$. Unfortunately, the corresponding situation is not that simple for equation (I).[3]

3. Solutions and supersolutions. Let us start by giving the precise assumptions about the structure of the equation

$$\frac{\partial u}{\partial t} = \operatorname{div}\mathcal{A}(x,t,\nabla u), \tag{3.1}$$

where $1 < p < \infty$ and $(x,t) \in \Omega$, Ω denoting a domain in $\mathbf{R}^n \times \mathbf{R}$. The solutions are understood in the weak sense, $u = u(x,t)$, and

$$\nabla u = \left(\frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_n}\right)$$

is the spatial gradient of u. The mapping

$$\mathcal{A} : \mathbf{R}^n \times \mathbf{R} \times \mathbf{R}^n \to \mathbf{R}^n$$

is assumed to satisfy the following conditions:

[2] While the solution $\mathcal{B}_p$ can be normalized so that

$$\int \mathcal{B}_p(x,t)\,dx = 1$$

for each $t > 0$, this is impossible to achieve for $\mathcal{W}_p$ when $p \neq 2$. The conservation law here is

$$\int \mathcal{W}_p(x,t)^{p-1}\,dx = 1$$

for all $t > 0$.

[3] In this respect, the "artificial equation"

$$\left|\frac{\partial u}{\partial t}\right|^{p-2}\frac{\partial u}{\partial t} = \operatorname{div}(|\nabla u|^{p-2}\nabla u),$$

not mentioned above, would be the most favorable.

(A) The mapping $(x,t) \mapsto \mathcal{A}(x,t,\xi)$ is measurable for all $\xi \in \mathbf{R}^n$, and the mapping $\xi \mapsto \mathcal{A}(x,t,\xi)$ is continuous for a.e. $(x,t) \in \mathbf{R}^n \times \mathbf{R}$.

(B) There are constants $0 < \alpha \leq \beta < \infty$ such that

$$\alpha|\xi|^p \leq \mathcal{A}(x,t,\xi) \cdot \xi \leq \beta|\xi|^p$$

for all $\xi \in \mathbf{R}^n$ and a.e. $(x,t) \in \mathbf{R}^n \times \mathbf{R}$.

(C) For a.e. $(x,t) \in \mathbf{R}^n \times \mathbf{R}$,

$$(\mathcal{A}(x,t,\xi) - \mathcal{A}(x,t,\zeta)) \cdot (\xi - \zeta) > 0$$

whenever $\xi \neq \zeta$ and $\xi, \zeta \in \mathbf{R}^n$.

It is sufficient that $\mathcal{A}$ is defined in $\Omega \times \mathbf{R} \times \mathbf{R}^n$. The role of (A) is merely to guarantee that composite functions like $\mathcal{A}(x,t,\nabla u(x,t))$ are measurable. Condition (C) implies uniqueness. For the p-Laplacian, the stronger inequality

$$\left(|\xi|^{p-2}\xi - |\zeta|^{p-2}\zeta\right) \cdot (\xi - \zeta) \geq 2^{1-p}|\xi - \zeta|^p$$

is valid for $p \geq 2$, and the inequality

$$\left(|\xi|^{p-2}\xi - |\zeta|^{p-2}\zeta\right) \cdot (\xi - \zeta) \geq (p-1)|\xi - \zeta|^2 \frac{|\xi|^{p-2} + |\zeta|^{p-2}}{2}$$

holds for $1 < p \leq 2$.

The solutions to equation (3.1) are functions in a local parabolic Sobolev space satisfying

$$\iint_{\Omega} \left(-u\frac{\partial \varphi}{\partial t} + \mathcal{A}(x,t,\nabla u) \cdot \nabla \varphi\right) dt\, dx = 0$$

for all test functions $\varphi \in C_0^1(\Omega)$. The a priori assumptions on u are the same as those in the definition below. It is essential that the time derivative u_t does not appear explicitly in the definition, for u_t is merely a distribution.

From now on, we shall, for instructional purposes, state the definitions and results only for the important p-parabolic equation. The general case is readily interpreted from this. The only exception is the explicit barrier constructed in §6.2, where the more general case is technically harder.

In what follows, Q will always stand for a parallelepiped

$$Q = (a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_n, b_n)$$

in $\mathbf{R}^n$, and the space–time sets

$$Q_T = Q \times (0,T), \quad Q_{t_1,t_2} = Q \times (t_1, t_2)$$

in $\mathbf{R}^n \times \mathbf{R}$ are called "boxes." In order to describe the appropriate function space, we introduce the abbreviation

$$V^p(t_1,t_2;Q) = C(t_1,t_2;L^2(Q)) \cap L^p(t_1,t_2;W^{1,p}(Q)).$$

Thus $u \in V^p(t_1,t_2;Q)$ means that the mapping

$$t \mapsto \int_Q |u(x,t)|^2\, dx$$

is continuous in $[t_1, t_2]$, the Sobolev derivative

$$\nabla u(x,t) = \left(\frac{\partial u(x,t)}{\partial x_1}, \dots, \frac{\partial u(x,t)}{\partial x_n}\right)$$

exists for a.e. $t \in [t_1, t_2]$, and

$$\int_{t_1}^{t_2} \left(\int_Q |\nabla u(x,t)|^p \, dx\right) dt < \infty.$$

In particular, the energy

$$\int_{t_1}^{t_2} \int_Q (|u|^2 + |\nabla u|^p) \, dx \, dt$$

is finite.

DEFINITION. *Let Ω be a domain in $\mathbf{R}^n \times \mathbf{R}$ and suppose that u belongs to $V^p(t_1, t_2; Q)$ whenever the closure of $Q \times (t_1, t_2)$ is contained in Ω. Then u is called a* solution *in Ω of the p-parabolic equation*

$$\frac{\partial u}{\partial t} = \operatorname{div}(|\nabla u|^{p-2} \nabla u) \tag{3.2}$$

if

$$\iint_\Omega (- u\varphi_t + |\nabla u|^{p-2} \nabla u \cdot \nabla \varphi) \, dt \, dx = 0$$

whenever $\varphi \in C_0^1(\Omega)$. If , in addition, u is continuous, then u is called p-parabolic *in Ω. Further, we say that u is a* supersolution *to* (3.2) *if the above integral is nonnegative whenever $\varphi \in C_0^1(\Omega)$ is nonnegative. A function v is a* subsolution *to* (3.2) *if $-v$ is a supersolution.*

By parabolic regularity theory, each locally bounded solution of (3.2) has a p-parabolic representative; for $1 < p < 2n/(2 + n)$, there are solutions that are not locally bounded (see [DB]). (In the pure p-parabolic case, even the spatial gradient ∇u, but not u_t, is Hölder continuous cf. [DB]. This is not true for solutions of the general equation (3.1). We shall not use this feature in what follows.) We will need a quantitative Hölder estimate. If u is p-parabolic in Ω and Ξ is a subdomain with compact closure in Ω, then the *interior Hölder estimate* has the form

$$|u(x, t_1) - u(y, t_2)| \le \gamma ||u||_{\infty,\Omega} (|x - y|^\alpha + |t_1 - t_2|^{\alpha/p}) \tag{3.3}$$

when (x, t_1), $(y, t_2) \in \Xi$. Here the positive exponent α depends only on n and p, while the constant γ depends, in addition, on the distance between Ξ and $\partial\Omega$, see [DB].

There is a principal, well-recognized difficulty with the definition. Namely, in proving any useful estimates, one usually needs a test function φ that depends on the solution u itself, for example, $\varphi = u\eta$, where η is a smooth function. Then the time derivative would contain u_t, but u_t does not necessarily exist as a function, so that φ is not, strictly speaking, admissible. This difficulty can be treated in two different ways: The first option is to use an equivalent definition in terms of Steklov averages as in [DB, pp. 18 and 25]. Second, one can proceed using convolutions of u with smooth mollifiers as in [AS, pp. 119–121]. Whichever the approach, the outcome is virtually the same as if the "forbidden" quantity u_t had been used at the intermediate stages, yet eliminated from the final formulation of the estimates. This remark concerns our proof of Lemma 3.1.

The parabolic boundary plays a crucial role for the Dirichlet boundary value problem. The *parabolic boundary* of $Q_T = Q \times (0,T)$ is

$$\Gamma_T = (\overline{Q} \times \{0\}) \cup (\partial Q \times (0,T]).$$

It consists of the bottom and the lateral sides, but the interior points of the top are excluded. A main feature, distinguishing the parabolic theory from the elliptic one, is that the p-parabolic functions are uniquely determined by their values on the parabolic boundary.

Often it is convenient to use test functions that vanish only on the parabolic boundary of Q_T. If φ is such a test function, then

$$-\int_Q u(x,T)\varphi(x,T)\,dx + \int_0^T \int_Q \big(- u\varphi_t + |\nabla u|^{p-2}\nabla u \cdot \nabla\varphi\big)\,dx\,dt = 0,$$

if u is p-parabolic. This follows easily from the definition by taking $\varphi\eta_\varepsilon$ as the test function, where η_ε is the usual cutoff function that depends only on t. In other words, η_ε is piecewise linear, $0 \le \eta_\varepsilon \le 1$, $\eta_\varepsilon(t) = 1$ for $t \le T-\varepsilon$, $\eta_\varepsilon(T) = 0$, and $|\eta_\varepsilon'(t)| \le 1/\varepsilon$.

We have a preliminary version of the comparison principle.

LEMMA 3.1. *Suppose that u is a supersolution and v a subsolution to* (3.2) *in $Q_T = Q \times (0,T)$. If u and $-v$ are lower semicontinuous on $\overline{Q}_T$ and $v \le u$ on the parabolic boundary of Q_T, then $v \le u$ a.e. in the box Q_T.*

Proof.[4] If $\varphi \in C_0^1(Q_T)$ is nonnegative, then

$$\int_0^T \int_Q \big(- u\varphi_t + |\nabla u|^{p-2}\nabla u \cdot \nabla\varphi\big)\,dx\,dt \ge 0$$

and

$$\int_0^T \int_Q \big(v\varphi_t - |\nabla v|^{p-2}\nabla v \cdot \nabla\varphi\big)\,dx\,dt \ge 0,$$

so that by adding them we have

$$\int_0^T \int_Q \big((v-u)\varphi_t + (|\nabla u|^{p-2}\nabla u - |\nabla v|^{p-2}\nabla v)\cdot\nabla\varphi\big)\,dx\,dt \ge 0.$$

Moreover, these inequalities remain true if u is replaced by $u + \varepsilon$, where ε is any constant.

To complete the proof we choose (formally) the test function φ to be

$$\varphi = (v - u - \varepsilon)_+\eta,$$

where $0 < \varepsilon < T$ is fixed and

$$\eta(t) = \begin{cases} \dfrac{T-\varepsilon-t}{T-\varepsilon} & \text{if } t \le T-\varepsilon, \\ 0 & \text{if } t \ge T-\varepsilon. \end{cases}$$

[4] In the case of two solutions with finite energies (they belong to $V^p(0,T;Q)$), a proof based on Steklov averages is given in [DB, Chap. VI, Lem. 3.1, pp. 160–161].

Then φ has compact support and we refer to the discussion above to justify the use of the time derivative of $(v-u-\varepsilon)_+$ in the calculations. Integrating by parts, we obtain

$$\begin{aligned}
&\int_0^T \int_{\{v\geq u+\varepsilon\}} \eta(|\nabla u|^{p-2}\nabla u - |\nabla v|^{p-2}\nabla v)\cdot(\nabla u - \nabla v)\,dx\,dt \\
&\leq \int_Q \int_0^T (v-u-\varepsilon)_+^2 \eta'\,dt\,dx + \frac{1}{2}\int_Q\int_0^T \eta\frac{\partial}{\partial t}(v-u-\varepsilon)_+^2\,dt\,dx \\
&= \frac{1}{2}\int_Q\int_0^T (v-u-\varepsilon)_+^2\eta'\,dt\,dx \\
&= -\frac{1}{2(T-\varepsilon)}\int_Q\int_0^{T-\varepsilon}(v-u-\varepsilon)_+^2\,dt\,dx \leq 0.
\end{aligned}$$

Since the first integral is nonnegative by the structural inequalities, we have that the last integral is zero, and hence $(v-u-\varepsilon)_+ = 0$ a.e. But this means that

$$v \leq u + \varepsilon$$

a.e. when $0 \leq t \leq T-\varepsilon$. Since $\varepsilon > 0$ was arbitrary, we have the desired inequality $v \leq u$ a.e. in Q_T. □

Remark. The above proof shows that a supersolution u is greater than a subsolution v on Q_T if $v(x,0) \leq u(x,0)$ and for each $t \in (0,T)$ the function $x \mapsto (v(x,t)-u(x,t))_+$ is in the Sobolev space $W_0^{1,p}(Q)$.

We shall need a basic existence theorem.[5]

LEMMA 3.2. *Suppose that θ is a continuous function on $\overline{Q}_T$. Then there is a unique p-parabolic function u that is continuous in $\overline{Q}_T$ and takes the boundary values $u=\theta$ on the parabolic boundary Γ_T. Moreover, if θ belongs to $V^p(0,T;Q)$, so does u.*

Proof. See [AL, Thm. 1.7, p. 318] or [L, pp. 162 and 166]. □

Remark 3.1. Using Lemma 3.2 repeatedly, one easily extends the existence result: *Suppose that Ω is a union of finitely many boxes*

$$Q_i \times (t_i, T_i)$$

and that θ is a continuous function on the parabolic boundary Γ of Ω. Then there is a unique p-parabolic function u, continuous on $\overline{\Omega}$, that coincides with θ on Γ. To verify this, one just has to begin with the earliest boxes. Here the parabolic boundary Γ is understood to be that part of the Euclidean boundary $\partial\Omega$ of Ω that lies in the union of the parabolic boundaries of the boxes $Q_i \times (t_i, T_i)$.

The obstacle problem is of fundamental importance to the nonlinear potential theory. Roughly speaking, one is looking for smallest supersolutions that lie above a given function ψ. Again we require an existence theorem for the box Q_T.

LEMMA 3.3. *Suppose that $\psi \in C^\infty(\overline{Q}_T)$ is given. Consider the class of all functions $v \in V^p(0,T;Q)$, continuous in $\overline{Q}_T$, such that $v \geq \psi$ in Q_T and $v = \psi$*

[5] In the recent literature, existence theorems are usually formulated for the Cauchy problem in $\mathbf{R}^n \times (0,T]$ or $\mathbf{R}^n \times (0,\infty)$; cf. [EV], [KV]. These results cannot easily be adapted to Q_T. A new existence proof should be constructed on the basis of the a priori estimates in [DB], but, to keep the presentation short, we give direct, though not quite adequate, references.

on the parabolic boundary Γ_T. Then there is a unique supersolution u in this class satisfying

$$\int_0^T \int_Q (v_t(v-u) + |\nabla u|^{p-2}\nabla u \cdot (\nabla v - \nabla u))\,dx\,dt$$

$$\geq \frac{1}{2}\int_Q |\psi(x,T) - u(x,T)|^2\,dx$$

for all smooth functions v in the aforementioned class.

Proof. See [AL, Thm. 3.2]. (To ensure that u is continuous, even when $t = T$, one may solve the problem in a slightly larger box, say $Q_{T+\varepsilon}$, $\varepsilon > 0$.) □

The unique supersolution given by the previous lemma is called *the solution to the obstacle problem.* It is p-parabolic in the set $\{u > \psi\}$.

The uniform Hölder estimate combined with the basic existence theorem leads to a convergence result.

LEMMA 3.4. *Suppose that u_k is a locally uniformly bounded sequence of p-parabolic functions in Ω. Then it has a subsequence that converges locally uniformly in Ω to a p-parabolic function.*

Proof. Let Ξ be a subdomain with compact closure in Ω. By the interior Hölder estimate (3.3), the sequence u_k is equicontinuous in $\overline{\Xi}$. Hence Ascoli's theorem allows us to select a subsequence that converges uniformly on Ξ to a continuous function. Exhausting Ω by an increasing sequence of such domains Ξ and performing a standard diagonalization process, we find a subsequence, denoted again by u_k, that converges locally uniformly in Ω to a continuous function, say h.

It sufficies to show that h is p-parabolic in each box $Q_{t_1,t_2} = Q \times (t_1, t_2)$ with closure in Ω. To this end, let H be the p-parabolic function in Q_{t_1,t_2}, continuous in $\overline{Q}_{t_1,t_2}$, taking the boundary values $H = h$ on the parabolic boundary Γ_{t_1,t_2} of Q_{t_1,t_2} (Lemma 3.2). Let $\varepsilon > 0$. For sufficiently large indices, we have

$$H - \varepsilon = h - \varepsilon < u_k < h + \varepsilon = H + \varepsilon$$

on Γ_{t_1,t_2}. Hence, by the comparison principle (Lemma 3.1),

$$H - \varepsilon \leq u_k \leq H + \varepsilon$$

in Q_{t_1,t_2}, so that

$$H - \varepsilon \leq h \leq H + \varepsilon$$

in Q_{t_1,t_2}. In conclusion, $h = H$ and hence h is p-parabolic. □

Remark 3.2. If $h_1 \leq h_2 \leq \cdots$ are p-parabolic in Q_T and

$$p > \frac{2n}{n+1},$$

then there are three possible alternatives for the limit function $h = \lim_{k\to\infty} h_k$:

(1) $h \equiv \infty$;

(2) h is p-parabolic in Q_T; and

(3) there is $\tau \in (0, T)$ such that h is p-parabolic in Q_τ, while $h(x,t) = \infty$ when $\tau < t \leq T$.

This result is obtained by using the intrinsic Harnack estimate [DB, pp. 157 and 184].

4. p-superparabolic functions. The supersolutions to equation (3.2) do not form a good closed class. In particular, the Barenblatt solutions

$$\mathcal{B}_p(x,t) = \begin{cases} t^{-\frac{n}{\lambda}}\left(C - \dfrac{p-2}{p}\lambda^{\frac{1}{1-p}}\left(\dfrac{|x|}{t^{1/\lambda}}\right)^{\frac{p}{p-1}}\right)_+^{\frac{p-1}{p-2}}, & t > 0, \\ 0, & t \leq 0, \end{cases}$$

where $\lambda = n(p-2)+p$, $p > 2$, and C is a real constant, are not supersolutions in any domain containing the origin $(0,0)$. It is the a priori integrability that fails to hold. In the elliptic theory an appropriate class, closed under monotone convergence, is provided by superharmonic functions. We study an analogous definition in this parabolic setting.

DEFINITION 4.1. *A function $u\,\Omega \to (-\infty, \infty]$ is called* p-superparabolic *if*

(i) *u is lower semicontinuous,*

(ii) *u is finite in a dense subset of Ω,*

(iii) *u satisfies the comparison principle on each box $Q_{t_1,t_2} = Q \times (t_1, t_2)$ with closure in Ω: if h is p-parabolic in Q_{t_1,t_2} and continuous on $\overline{Q}_{t_1,t_2}$ and if $h \leq u$ on the parabolic boundary of Q_{t_1,t_2}, then $h \leq u$ in the whole Q_{t_1,t_2}.*

The simplest interesting example is perhaps

$$u(x,t) = \begin{cases} 0 & \text{when } t \leq 0, \\ 1 & \text{when } t > 0. \end{cases}$$

This definition is the same as that used by Gehring [G] and Friedman [F] in their linear theory. The original definition of Sternberg used trapezoids; cf. [S]. Watson [W] has a different definition based on a representation formula, a counterpart to the Poisson formula, with "heat balls" playing the role of fundamental domains. All of these variants result in the same concept. To see this in our nonlinear theory, we will later prove that (iii) can be weakened considerably.

Let Ξ be any domain with compact closure in Ω. If h is p-parabolic in Ξ and continuous in $\overline{\Xi}$ and if $h \leq u$ on the (Euclidean) boundary $\partial\Xi$, then $h \leq u$ in the whole Ξ.

It is easy to see that (iii) implies this. It is clear if Ξ is a box, and it easily follows for a union of finitely many boxes. The case where Ξ is arbitrary is verified by covering the set where $h \geq u + \varepsilon$, $\varepsilon > 0$, with finitely many boxes. It is less obvious that the comparison property above will imply (iii). We will return to this question in Lemma 6.3.

It is immediately seen from the definition that if u and v are p-superparabolic in Ω, then so are the minimum $\min(u,v)$ and the functions $\lambda u + \mu$, where $\lambda \geq 0$ and $\mu \in \mathbf{R}$ are constants.

The result below implies that a proper version of a supersolution is p-superparabolic; in particular, the Barenblatt solution is p-superparabolic in the whole $\mathbf{R}^n \times (-\infty, \infty)$.

LEMMA 4.2. *If u is a lower semicontinuous supersolution to the p-parabolic equation (3.2) in Ω, then there is a p-superparabolic function v in Ω, defined by*

$$v(x,t) = \operatorname{ess} \liminf_{(y,s)\to(x,t)} u(y,s), \tag{4.1}$$

$(x,t) \in \Omega$, such that $v = u$ a.e. in Ω. In particular, if u is continuous, then u is p-superparabolic.

Proof. Since u is lower semicontinuous,

$$u(x,t) \leq \liminf_{(y,s)\to(x,t)} u(y,s) \leq \operatorname{ess} \liminf_{(y,s)\to(x,t)} u(y,s) = v(x,t)$$

for each $(x,t) \in \Omega$, whereas

$$u(x,t) = \lim_{r\to 0} \frac{1}{2r|B(x,r)|} \int_{t-r}^{t+r} \int_{B(x,r)} u(y,s)\,dy\,ds$$
$$\geq \operatorname{ess} \liminf_{(y,s)\to(x,t)} u(y,s) = v(x,t)$$

for a.e. $(x,t) \in \Omega$ by the Lebesgue differentiation theorem. Hence $u = v$ a.e. in Ω. Moreover, it is clear that v is lower semicontinuous and that v is finite a.e.. To see that v is indeed p-superparabolic, fix a box $Q \times (t_1, t_2)$ with closure in Ω. If h is any p-parabolic function, continuous in $\overline{Q} \times [t_1, t_2]$, such that $h \leq v$ on the parabolic boundary of $Q \times (t_1, t_2)$, then h is a subsolution in $Q \times (t_1, t_2)$ and, therefore, $h \leq v$ a.e. on $Q \times (t_1, t_2)$ by the comparison principle (Lemma 3.1). By (4.1), this inequality holds everywhere on $Q \times (t_1, t_2)$. □

We next present two versions of the comparison principle. The first one is "elliptic."

LEMMA 4.3. *Suppose that u is p-superparabolic and v is p-subparabolic in a bounded open set Ω. If u and v are bounded and if*

$$\limsup_{\xi\to\xi_0} v(\xi) \leq \liminf_{\xi\to\xi_0} u(\xi)$$

at each point ξ_0 on the Euclidean boundary $\partial\Omega$ of Ω, then $v \leq u$ in Ω.

Proof. For fixed $\varepsilon > 0$, the set

$$K_\varepsilon = \{\xi \in \Omega : v(\xi) \geq u(\xi) + \varepsilon\}$$

is a compact subset of Ω. Therefore, there is an open set $D_\varepsilon \subset \Omega$ such that $K_\varepsilon \subset D_\varepsilon$, where D_ε is a union of finitely many boxes $Q_i \times (t_{i_1}, t_{i_2})$ and

$$\partial D_\varepsilon \subset \Omega \setminus K_\varepsilon.$$

Because v is upper semicontinuous, u is lower semicontinuous, and the parabolic boundary Γ_ε of D_ε is compact, we find a continuous function θ on Γ_ε such that $v \leq \theta \leq u + \varepsilon$ on Γ_ε. If h is the p-parabolic function in D_ε that coincides with θ on Γ_ε (see Remark 3.1), then we infer from the definition of p-superparabolic and p-subparabolic functions that

$$v \leq h \leq u + \varepsilon$$

in D_ε. Hence $v \leq u + \varepsilon$ in Ω, and the lemma follows by letting $\varepsilon \to 0$. □

In a similar manner, the reader easily establishes the following "parabolic" comparison principle:

LEMMA 4.4. *Suppose that u is p-superparabolic and v is p-subparabolic in a cylinder $\Omega = G \times (t_1, t_2)$, where $G \subset \mathbf{R}^n$ is a bounded domain. If u and v are bounded and if*

$$\limsup_{\xi\to\xi_0} v(\xi) \leq \liminf_{\xi\to\xi_0} u(\xi)$$

at each point ξ_0 on the parabolic boundary[6] *of Ω, then $v \leq u$ in Ω.*

[6] The parabolic boundary is $(\overline{G} \times \{t_1\}) \cup (\partial G \times [t_1, t_2])$.

4.6. Parabolic modification The main tool in the Perron method is the modification of p-superparabolic functions. Let $Q_T = Q \times (0,T)$ be a box with closure in Ω. If u is p-superparabolic in Ω and bounded on Q_T, we define the *p-parabolic modification*

$$U = \begin{cases} u & \text{in } \Omega \setminus \overline{Q_T}, \\ v & \text{in } Q \times [0,T], \end{cases}$$

where

$$v(\xi) = \sup\{h(\xi) : h \in C(\overline{Q_T}) \text{ is } p\text{-parabolic and } h \le u \text{ on } \Gamma_T\}.$$

Then it is clear that $U \le u$ on Ω. Moreover, *U is p-superparabolic in Ω and p-parabolic in Q_T.* To see this, choose an increasing sequence θ_j of continuous functions on Γ_T such that

$$u = \lim_{j \to \infty} \theta_j$$

on Γ_T. Let h_j be the p-parabolic function in Q_T that coincides with θ_j on Γ_T. Then it follows from the comparison principle that the sequence h_j is increasing on $\overline{Q_T}$ and that the limit function is v. Moreover, since the sequence h_j is bounded, the limit function v is p-parabolic by Lemma 3.4. Now it is rather immediate that U is p-superparabolic in Ω.

5. The Perron method. In what follows, we let Ω be a bounded open set in $\mathbf{R}^n \times \mathbf{R}$. Let $f : \partial\Omega \to \mathbf{R}$ be any bounded function; note that f is defined on the Euclidean boundary of Ω. The Perron method aims at constructing a p-parabolic function H in Ω that takes the boundary values f on $\partial\Omega$. Of course, this is not possible in this generality. In this section, we shall construct two functions, the upper and the lower Perron solutions $\overline{H}_f$ and $\underline{H}_f$, that both correspond in a sense to f on $\partial\Omega$.

A function u is said to belong to the *upper class* $\mathcal{U}_f$ if u is p-superparabolic in Ω and bounded below and

$$\liminf_{\eta \to \xi} u(\eta) \ge f(\xi)$$

at each point $\xi \in \partial\Omega$. Observe that the upper class $\mathcal{U}_f$ is never empty, for f is bounded so that large constants are members of $\mathcal{U}_f$.

The *lower class* $\mathcal{L}_f$ is defined analogously. It consists of p-subparabolic functions v, bounded above, satisfying

$$\limsup_{\eta \to \xi} v(\eta) \le f(\xi)$$

at each point $\xi \in \partial\Omega$; also, the constant $-\infty$ is in $\mathcal{L}_f$.

Next, the *upper solution* $\overline{H}_f$ and the *lower solution* $\underline{H}_f$ are defined by

$$\overline{H}_f(\xi) = \inf\{u(\xi) : u \in \mathcal{U}_f\}$$

and

$$\underline{H}_f(\xi) = \sup\{v(\xi) : v \in \mathcal{L}_f\}.$$

It follows from the comparison principle (Lemma 4.3) that $v \le u$ whenever $u \in \mathcal{U}_f$ and $v \in \mathcal{L}_f$. Hence

$$\underline{H}_f \le \overline{H}_f$$

in Ω. Since the boundary function f is bounded, both $\overline{H}_f$ and $\underline{H}_f$ are bounded by the same constants as f.

THEOREM 5.1. *If the boundary function $f : \partial\Omega \to \mathbf{R}$ is bounded, then the Perron solutions $\underline{H}_f$ and $\overline{H}_f$ are p-parabolic.*

Proof. Fix a box $Q_{t_1,t_2} = Q \times (t_1, t_2)$ with closure in Ω; here Q is a rectangle in $\mathbf{R}^n$ as usual. Next, choose a countable, dense subset

$$\Xi = \{\xi_1, \xi_2, \dots\}$$

of Q_{t_1,t_2}. For each $j = 1, 2, \dots$ we choose a sequence of functions $u_{i,j}$ in $\mathcal{U}_f$ such that

$$\lim_{i\to\infty} u_{i,j}(\xi_j) = \overline{H}_f(\xi_j).$$

Moreover, we are free to replace $u_{i,j+1}$ by $\min(u_{i,j}, u_{i,j+1})$, and hence we have that

$$\lim_{i\to\infty} u_{i,j}(\xi_k) = \overline{H}_f(\xi_k) \tag{5.1}$$

for each $k = 1, 2, \dots, j$ and for each j. For the p-parabolic modification $U_{i,j}$ of $u_{i,j}$ in Q_{t_1,t_2}, it holds that

$$\overline{H}_f \le U_{i,j} \le u_{i,j}$$

and $U_{i,j}$ is p-parabolic in Q_{t_1,t_2}. By passing to a subsequence, if necessary, we infer from Lemma 3.4 that $U_{i,j}$ converges locally uniformly to a p-parabolic function v_j in Q_{t_1,t_2}. By again employing Lemma 3.4, we find a subsequence of v_j that converges locally uniformly to a p-parabolic function h in Q_{t_1,t_2}. By the construction, it is clear that

$$h \ge \overline{H}_f$$

in Q_{t_1,t_2}. On the other hand, by (5.1), we have that $h = \overline{H}_f$ in the dense subset Ξ of Q_{t_1,t_2}. Therefore, if u is any function from $\mathcal{U}_f$, then its p-parabolic modification U in Q_{t_1,t_2} is not greater than u, and by continuity, $U \ge h$ in Q_{t_1,t_2}. Hence

$$\overline{H}_f \ge h$$

in Q_{t_1,t_2}. It follows that $\overline{H}_f = h$ is p-parabolic in Q_{t_1,t_2} and hence in Ω.

The lower solution $\underline{H}_f$ is treated completely analogously. □

Remark. Alternative proofs for Theorem 5.1 could be based on the Choquet topological lemma as in [HKM] or on the argument in [GLM].

Example. Let $\Omega = Q \times (0, T)$ and suppose that $f : \partial\Omega \to \mathbf{R}$ is continuous. Then the upper and the lower Perron solutions coincide and

$$H_f = \underline{H}_f = \overline{H}_f$$

is the p-parabolic function that coincides with f on the parabolic boundary of Ω. As anticipated, the values of f at the top of the box Ω do not have any influence on the solution. Indeed, if h is p-parabolic in Ω and takes the values f on the *parabolic boundary* (Lemma 3.2), then the function

$$h(x,t) + \frac{\varepsilon}{T-t}$$

belongs to $\mathcal{U}_f$ for $\varepsilon > 0$ and to $\mathcal{L}_f$ for $\varepsilon < 0$. The result can be restated by saying that f is *resolutive*.

6. Barriers and boundary regularity. We shall define the barrier function for the boundary value problem as in classical theory. It gives a necessary and sufficient condition for the regularity of boundary points. This leads to some useful conditions for regularity; for instance, an exterior sphere condition is established. Furthermore, the regular boundary points of space–time cylinders are completely characterized in terms of a Wiener-type test.

DEFINITION. *Suppose that ξ_0 is a boundary point of a bounded domain $\Omega \subset \mathbf{R}^n \times \mathbf{R}$. A function w is a* barrier in Ω at the point ξ_0 *if*

$$\begin{aligned}
&\text{(i)} && w \text{ is positive and } p\text{-superparabolic in } \Omega,\\
&\text{(ii)} && \liminf_{\zeta\to\xi} w(\zeta) > 0 \quad \text{if } \xi \in \partial\Omega, \quad \xi \neq \xi_0,\\
&\text{(iii)} && \liminf_{\zeta\to\xi_0} w(\zeta) = 0.
\end{aligned}$$

Although we have, for convienence, assumed that the barrier is defined in the whole Ω, *this is completely a local question*: Let $\tilde{\Omega}$ be another domain such that

$$\overline{B} \cap \tilde{\Omega} = \overline{B} \cap \Omega$$

for some open ball B centered at ξ_0. Suppose that there is a barrier, say w, in $\tilde{\Omega}$ at ξ_0. Let

$$m = \inf\{w(\xi) : \xi \in \partial B \cap \tilde{\Omega}\}.$$

Then $m > 0$ and it easily follows that the function

$$v = \begin{cases} \min(w, m) & \text{in } B \cap \tilde{\Omega}, \\ m & \text{in } \Omega \setminus B \end{cases}$$

is a barrier in Ω. Hence there is a barrier in Ω at ξ_0 exactly when there is a barrier in $\tilde{\Omega}$.

THEOREM 6.1. *Suppose that $f : \partial\Omega \to \mathbf{R}$ is bounded and continuous at $\xi_0 \in \partial\Omega$. If there is a barrier in Ω at ξ_0, then*

$$\lim_{\xi\to\xi_0} \underline{H}_f(\xi) = f(\xi_0) = \lim_{\xi\to\xi_0} \overline{H}_f(\xi).$$

Proof. The proof is very classical. Let $|f(\xi) - f(\xi_0)| < \varepsilon$ for $\xi \in \partial\Omega$ with $|\xi - \xi_0| < \delta$ and choose, by the aid of the lower semicontinuity of the barrier w, a constant $M > 0$ such that

$$M\, w(\xi) \geq 2 \sup |f|$$

for $\xi \in \overline{\Omega}$, $|\xi - \xi_0| < \delta$. Then the function $M\,w + \varepsilon + f(\xi_0)$ belongs to the upper class $\mathcal{U}_f$ and has the limit $f(\xi_0) + \varepsilon$ at ξ_0. Similarly, the function $-M\,w - \varepsilon + f(\xi_0)$ belongs to the lower class $\mathcal{L}_f$ and has the limit $f(\xi_0) - \varepsilon$ at ξ_0. The theorem follows. □

We call a boundary point $\xi_0 \in \partial\Omega$ *p-regular* if

$$\lim_{\xi\to\xi_0} \overline{H}_f(\xi) = f(\xi_0)$$

whenever $f : \partial\Omega \to \mathbf{R}$ is continuous. This concept depends heavily on the equation. For instance, it may happen that a point $\xi_0 \in \partial\Omega$ is regular for the equation $\Delta u = u_t$, while it is not regular for the equation $\Delta u = \frac{1}{2}u_t$. Thus the term $\mathcal{A}_p$*-regular* is appropriate in the general case.

Because $\underline{H}_f = -\overline{H}_{-f}$, we could replace $\overline{H}_f$ by $\underline{H}_f$ in the definition above.

We have the classical characterization for regularity in terms of barriers.

THEOREM 6.2. *A boundary point ξ_0 of a bounded open set Ω is p-regular if and only if there is a barrier in Ω at ξ_0.*

Proof. The sufficiency part was already established. To prove the necessity, let $\xi_0 = (x_0, t_0)$ and define

$$\psi(x,t) = \frac{p-1}{p}|x-x_0|^{p/(p-1)} + \varepsilon(t-t_0)^2,$$

where ε is a constant with

$$0 < 2\varepsilon \operatorname{diam}(\Omega) < n.$$

The function ψ is p-subparabolic in Ω (Lemma 4.2). Then $w = \underline{H}_\psi$ is a barrier in Ω at ξ_0 because

$$\lim_{\xi \to \xi_0} w(\xi) = \psi(\xi_0) = 0$$

by the p-regularity of ξ_0, and the other properties are immediate, for $w \geq \psi$ in Ω. The theorem follows. □

Since the existence of a barrier is a local property, so is the regularity of a boundary point. Moreover, if ξ_0 is a p-regular boundary point of Ω, then ξ_0 is p-regular with respect to each subdomain to whose boundary it belongs.

Example. Exterior sphere condition. Let $\xi_0 = (x_0, t_0) \in \partial\Omega$. Suppose that there exists a closed ball

$$\{(x,t) : |x-x'|^2 + (t-t')^2 \leq R_0^2\}$$

that intersects with the closure $\overline{\Omega}$ exactly at ξ_0. We claim that ξ_0 is regular if $x_0 \neq x'$.

For the construction of the barrier, we define

$$w(x,t) = e^{-\alpha R_0^2} - e^{-\alpha R^2}, \quad R = (|x-x'|^2 + (t-t')^2)^{1/2},$$

where the constant $\alpha > 0$ is to be fixed later. For $(x,t) \in \Omega$ sufficiently near to ξ_0, we have that

$$0 < \delta \leq |x-x'|, \quad -2R_0 \leq t-t', \quad R_0 \leq R < 2R_0,$$

where we can take $2\delta \leq |x_0 - x'|$, for instance. Keep in mind that it suffices to construct a local barrier. An easy calculation yields

$$\begin{aligned}
&\operatorname{div}(|\nabla w|^{p-2}\nabla w) \\
&= (2\alpha)^{p-1}|x-x'|^{p-2}e^{-(p-1)\alpha R^2}(n+p-2-2\alpha(p-1)|x-x'|^2) \\
&\leq (2\alpha)^{p-1}(n+p-2-2\alpha(p-1)\delta^2)e^{-(p-1)\alpha R^2}
\end{aligned}$$

and

$$w_t = 2\alpha e^{-\alpha R^2}(t-t') \geq -4\alpha R_0 e^{-\alpha R^2}.$$

Consider first the case $1 < p < 2$. Choose $\alpha > 0$ so large that $n+p-2-2\alpha(p-1)\delta^2 \leq -1$. In order to have

$$\operatorname{div}(|\nabla w|^{p-2}\nabla w) < w_t$$

near ξ_0 in Ω, we have to verify that

$$(2\alpha)^{p-2}e^{-(p-2)\alpha R^2} \geq 2R_0,$$

which is clearly the case if α is large enough, since $p-2<0$. Thus w is p-superparabolic in a neighborhood of ξ_0 in Ω by Lemma 4.2. Consequently, w is a local barrier at ξ_0, and ξ_0 is therefore p-regular.

The case $p=2$ is even simpler, and we leave it for the reader to explore.

Finally, let $p>2$ and choose α so that

$$n+p-2-2\alpha(p-1)\delta^2 = -1.$$

Then

$$\operatorname{div}(|\nabla w|^{p-2}\nabla w) \leq -(2\alpha)^{p-1}e^{-(p-1)4\alpha R_0^2}$$

and

$$w_t \geq -4\alpha R_0 e^{-\alpha R_0^2}$$

so that

$$\operatorname{div}(|\nabla w|^{p-2}\nabla w) \leq Cw_t,$$

where $C = C(p,\alpha,R_0) > 0$. Now, for $\lambda > 0$ small enough, the function λw is p-superparabolic and hence the desired local barrier.

The restriction that $x_0 \neq x'$ is essential only to exclude the south pole $(x', t'-R_0)$ as a tangent point (cf. the example after Theorem 5.1). Indeed, if

$$\xi_0 = (x', t'+R_0)$$

is the north pole (the latest moment on the exterior sphere), then the function w above is a barrier at ξ_0, for

$$w_t = 2\alpha e^{-\alpha R^2}(t-t') > 0$$

near ξ_0 in this case, while

$$\operatorname{div}(|\nabla w|^{p-2}\nabla w) \leq 0$$

for α large enough.

As an application of the exterior sphere condition, we prove the characterization of p-superparabolic functions, referred to in the discussion after Definition 4.1.

LEMMA 6.3. *Suppose that $u : \Omega \to (-\infty, \infty]$ is lower semicontinuous in $\Omega \subset \mathbf{R}^n \times \mathbf{R}$ and finite in a dense subset of Ω. Then u is p-superparabolic if and only if for each domain Ξ with compact closure in Ω and each $h \in C(\overline{\Xi})$, p-parabolic in Ξ, the condition $h \leq u$ on $\partial\Xi$ implies $h \leq u$ in Ξ.*

Proof. The necessity of the condition was discussed right after the definition of p-superparabolic functions. To prove the sufficiency, let $Q_{t_1,t_2} = Q \times (t_1,t_2)$ be a box with closure in Ω and let $h \in C(\overline{Q}_{t_1,t_2})$ be p-parabolic in Q_{t_1,t_2} such that $h \leq u$ on the parabolic boundary Γ_{t_1,t_2} of Q_{t_1,t_2}. Suppose that

$$Q = (a_1,b_1) \times (a_2,b_2) \times \cdots \times (a_n,b_n).$$

For $\delta > 0$, $\delta < t_2 - t_1$, choose a hyperplane P_δ in $\mathbf{R}^n \times \mathbf{R}$ such that the points $(x, t_2-\delta)$, where $x_1 = a_1$, and (y, t_2), where $y_1 = b_1$, belong to P_δ. Then let Ξ be the subset of Q_{t_1,t_2} that contains the points (x,t), where $t < s$ is such that $(x,s) \in P_\delta$. We easily infer from the exterior sphere condition that Ξ is p-regular. For fixed $\varepsilon > 0$, we may choose $\delta > 0$ so small that

$$u(x,t) \geq h(x,t) - \frac{\varepsilon}{t_2 + \frac{\delta}{2} - t}$$

for $(x,t) \in P_\delta \cap \partial\Xi$. Let $\overline{H}$ be the Perron solution in Ξ with

$$h - \frac{\varepsilon}{t_2 + \frac{\delta}{2} - t}$$

as a boundary function. Then $\overline{H}$ is continuous up to the boundary of Ξ. Since we are free to assume that $u \geq 0$, we have that

$$u \geq H$$

in Ξ, since the same inequality holds on $\partial\Xi$. This means that

$$u(x,t) \geq h(x,t) - \frac{\varepsilon}{t_2 + \frac{\delta}{2} - t}$$

for $(x,t) \in \Xi$, and by letting $\delta \to 0$ and $\varepsilon \to 0$, we arrive at the desired inequality

$$u \geq h$$

in Q_{t_1,t_2}. □

Before constructing more barriers, let us recall that it may happen that a point $\xi_0 \in \partial\Omega$ is regular for the equation $\Delta u = u_t$, while it is not regular for $\Delta u = \frac{1}{2}u_t$.

Example: *Heat balls.* The heat balls play a central role for the more refined parts of linear theory. They are defined through the inequality

$$\frac{1}{(4\pi(t_0 - t))^{n/2}} \exp\left(-\frac{|x_0 - x|^2}{4(t_0 - t)}\right) > c,$$

where c is a positive constant and $t < t_0$. In other words, heat balls are level sets of the fundamental solution of the heat equation. It is an essential feature that the "center" (x_0, t_0) is an irregular boundary point of the heat ball, while all the other boundary points are regular.

The p-parabolic balls are, by analogy, defined through the Barenblatt solutions $\mathcal{B}_p$. The inequality is now

$$\mathcal{B}_p(x_0 - x, t_0 - t) > c,$$

where $t < t_0$ and $p > 2$. Then p-parabolic balls are bounded and from the exterior sphere condition, we infer that its boundary points except the "center" (x_0, t_0) are regular. Indeed, *the "center" (x_0, t_0) is an irregular boundary point of the p-parabolic ball.*

To prove the irregularity we are free to assume that (x_0, t_0) is the origin. We have the inequality

$$\left(1 - \frac{p-2}{p}\lambda^{\frac{1}{1-p}}\left(\frac{|x|}{(-t)^{1/\lambda}}\right)^{\frac{p}{p-1}}\right)_+^{\frac{p-1}{p-2}} > c(-t)^{\frac{n}{\lambda}}$$

and $t < 0$. For simplicity, we have taken the normalizing constant in the Barenblatt solution to be 1. Since the irregularity is a local question, we may restrict the range of t further, say,

$$-\left(\frac{c}{2}\right)^{-\lambda/n} < t < 0.$$

It suffices to show that the origin is an irregular boundary point of the subdomain

$$\left(1-\frac{p-2}{p}\lambda^{\frac{1}{1-p}}\left(\frac{|x|}{(-t)^{1/\lambda}}\right)^{\frac{p}{p-1}}\right)_+^{\frac{p-1}{p-2}} > \frac{1}{2},$$

where $-T < t < 0$. This can be written as

$$\left(1+\frac{p-2}{p}\lambda^{\frac{1}{1-p}}\left(\frac{|x|}{(-t)^{1/\lambda}}\right)^{\frac{p}{p-1}}\right)^{\frac{p-1}{p-2}} < \left(2-\left(\frac{1}{2}\right)^{\frac{p-2}{p-1}}\right)^{\frac{p-1}{p-2}}.$$

For a sufficiently small value of the constant K, this contains the subdomain

$$u(x,t) < K,$$

where

$$u(x,t) = \left(1+\frac{p-2}{p}\lambda^{\frac{1}{1-p}}\left(\frac{|x|}{(-t)^{1/\lambda}}\right)^{\frac{p}{p-1}}\right)^{\frac{p-1}{p-2}} - 1 - B(-t)^{\frac{n(p-2)}{\lambda}},$$

Here B is a positive constant, to be specified later. The function u will do as "eine Irregularitätsbarriere," to quote an expression from [Py]. That is,

(i) u is p-superparabolic in the domain, where $u(x,t) < K$ and $-T < t < 0$ (T depending on p),

(ii) $u(0,t) < 0$, when $t < 0$.

Here (ii) is evident. Before proving (i), let us show how this implies that the origin is irregular.

Consider the domain bounded by the surface $u(x,t) = K$ and the hyperplane $t = -T$; in this domain

$$u(x,t) < K.$$

If we assign the boundary values f to be K when $-T < t \leq 0$ and $u(x,-T)$ on the plane $t = -T$, then the function

$$u(x,t) + \frac{\varepsilon}{-t}, \qquad \varepsilon > 0,$$

is in the upper class for f. Thus

$$\overline{H}_f(x,t) \leq u(x,t) + \frac{\varepsilon}{-t},$$

and hence $\overline{H}_f \leq u$. But this implies that

$$\limsup_{t\to 0-} \overline{H}_f(0,t) \leq \limsup_{t\to 0-} u(0,t) \leq 0,$$

so that $\overline{H}_f$ cannot attain the right boundary value K at the origin.

The proof of (i) is a calculation. We may assume that $K \leq 1$. Direct derivations and some minor manipulations with the resulting expressions yield

$$\begin{aligned}
&u_t - \operatorname{div}(|\nabla u|^{p-2}\nabla u) = B\frac{n(p-2)}{\lambda}(-t)^{\frac{n(p-2)}{\lambda}-1}\\
&\quad -\frac{p}{\lambda(p-2)}\left(\frac{1}{(-t)} - \frac{1}{(-t)^{p/\lambda}}\right)\left(1+\frac{p-2}{p}\lambda^{\frac{1}{1-p}}\left(\frac{|x|}{(-t)^{1/\lambda}}\right)^{\frac{p}{p-1}}\right)^{\frac{1}{p-2}}\\
&\quad +\left(\frac{p}{\lambda(p-2)}\left(\frac{1}{(-t)} - \frac{1}{(-t)^{p/\lambda}}\right) - \frac{n}{\lambda(-t)^{p/\lambda}}\right)\left(1+\frac{p-2}{p}\lambda^{\frac{1}{1-p}}\left(\frac{|x|}{(-t)^{1/\lambda}}\right)^{\frac{p}{p-1}}\right)^{\frac{p-1}{p-2}}\\
&\geq B\frac{n(p-2)}{\lambda}(-t)^{\frac{n(p-2)}{\lambda}-1}\\
&\quad -\frac{p}{\lambda(p-2)}\left(\frac{1}{(-t)} - \frac{1}{(-t)^{p/\lambda}}\right)\left(1+\frac{p-2}{p}\lambda^{\frac{1}{1-p}}\left(\frac{|x|}{(-t)^{1/\lambda}}\right)^{\frac{p}{p-1}}\right)^{\frac{p-1}{p-2}}\\
&\quad +\left(\frac{p}{\lambda(p-2)}\left(\frac{1}{(-t)} - \frac{1}{(-t)^{p/\lambda}}\right) - \frac{n}{\lambda(-t)^{p/\lambda}}\right)\left(1+\frac{p-2}{p}\lambda^{\frac{1}{1-p}}\left(\frac{|x|}{(-t)^{1/\lambda}}\right)^{\frac{p}{p-1}}\right)^{\frac{p-1}{p-2}}\\
&= B\frac{n(p-2)}{\lambda}\frac{1}{(-t)^{p/\lambda}} - \frac{n}{\lambda}\frac{1}{(-t)^{p/\lambda}}\left(1+\frac{p-2}{p}\lambda^{\frac{1}{1-p}}\left(\frac{|x|}{(-t)^{1/\lambda}}\right)^{\frac{p}{p-1}}\right)^{\frac{p-1}{p-2}}\\
&\geq B\frac{n(p-2)}{\lambda}\frac{1}{(-t)^{p/\lambda}} - \frac{n}{\lambda}\frac{1}{(-t)^{p/\lambda}}\left(2 + B(-t)^{\frac{n(p-2)}{\lambda}}\right).
\end{aligned}$$

The last expression is nonnegative if

$$B(p-2) \geq 2 + B(-T)^{\frac{n(p-2)}{\lambda}}.$$

Take

$$B = \frac{2}{p-2-(-T)^{n(p-2)/\lambda}}$$

and choose T small enough for this to make sense.

This concludes our proof. □

What is to happen in the future will have no influence on the present time. The regularity or irregularity of a boundary point $\xi_0 = (x_0, t_0) \in \partial\Omega$ is completely determined by times $t < t_0$. Let

$$\Omega_- = \{(x,t) \in \Omega : t < t_0\}$$

and

$$\Omega_+ = \{(x,t) \in \Omega : t > t_0\}.$$

Note that both Ω_- and Ω_+ may be disconnected; however, it is easy to see that the barrier characterization for regularity remains true for disconnected open sets.

THEOREM 6.4. *Let $\xi_0 = (x_0, t_0) \in \partial\Omega$. Then ξ_0 is a regular boundary point of the domain Ω if and only if ξ_0 is a regular boundary point of Ω_- or $\xi_0 \notin \partial\Omega_-$.*

Proof. The necessity of the condition follows since $\Omega_- \subset \Omega$, and therefore, either $\xi_0 \notin \partial\Omega_-$ or a barrier at ξ_0 in Ω is also a barrier in Ω_-.

To establish the converse, suppose first that $\xi_0 \notin \partial\Omega_-$. Then the exterior sphere condition ensures that ξ_0 is a regular boundary point of Ω.

To complete the proof, suppose that ξ_0 is a regular boundary point of Ω_-. Given the p-subparabolic boundary values

$$\psi(x,t) = \frac{p-1}{p}|x - x_0|^{p/(p-1)} + \varepsilon(t - t_0)^2$$

on $\partial\Omega$, where ε is a constant with $0 < 2\varepsilon\mathrm{diam}(\Omega) < n$, the lower Perron solution $H = \underline{H}_\psi$ satisfies the inequality

$$H \geq \psi > 0$$

in Ω. Moreover,

$$\lim_{\substack{\zeta\to\xi_0\\ \zeta\in\Omega_-}} H(\zeta) = \psi(\xi_0) = 0. \tag{6.1}$$

Indeed, let u be any bounded function in the upper class for ψ in Ω_-. Then for each $\varepsilon > 0$, the function

$$v(x,t) = \begin{cases} \sup u & \text{if } t > t_0 - \varepsilon, \\ u(x,t) & \text{if } t \leq t_0 - \varepsilon \end{cases}$$

is in the upper class for ψ in Ω. It follows that the restriction to Ω_- of H coincides with the upper Perron solution of ψ in Ω_-. Hence (6.1) follows, since ξ_0 is a regular boundary point of Ω_-.

This means that H is a barrier at ξ_0 in Ω_-. We claim that H will do as a barrier in Ω. To this end, we have to show that

$$\lim_{\substack{\zeta\to\xi_0\\ \zeta\in\Omega}} H(\zeta) = 0. \tag{6.2}$$

This is easy. If $\xi_0 \notin \partial\Omega_+$, there is nothing to prove. So assume that $\xi_0 \in \partial\Omega_+$ and define

$$\varphi = \begin{cases} \psi & \text{in } \partial\Omega, \\ H & \text{in } \Omega \cap \partial\Omega_+. \end{cases}$$

Then the restriction to $\partial\Omega_+$ of φ is continuous at ξ_0. Let

$$h = \underline{H}_\varphi$$

be the lower Perron solution of φ in Ω_+. Then $h = H$ in Ω_+. Indeed, if $u \in \mathcal{L}_\psi(\Omega)$, then $u|_{\Omega_+} \in \mathcal{L}_\varphi(\Omega_+)$ so that

$$u|_{\Omega_+} \leq h$$

and hence

$$H \leq h$$

in Ω_+. For the reverse inequality, let $v \in \mathcal{L}_\varphi(\Omega_+)$. Then

$$\limsup_{\zeta\to\xi} v(\zeta) \leq \varphi(\xi) \leq \liminf_{\zeta\to\xi} H(\zeta)$$

for each $\xi \in \partial\Omega_+$, so that $v \leq H$ in Ω_+ by the comparison principle. Thus $h \leq H$ and we conclude that $h = H$ in Ω_+.

Since, by the exterior sphere condition, the earliest points are regular, ξ_0 is a regular boundary point of Ω_+. Hence

$$\lim_{\substack{\zeta \to \xi_0 \\ \zeta \in \Omega_+}} h(\zeta) = \varphi(\xi_0) = 0.$$

Since $H = h$ in Ω_+, this and (6.1) yield (6.2). □

Remark 6.1. The previous proof reveals an interesting fact. The restriction to Ω_- of the Perron solution $\overline{H}_f$ in Ω coincides with the upper Perron solution of f in Ω_- no matter how the bounded function f is defined at points on $\partial\Omega_- \setminus \partial\Omega$.

6.1. Space–time cylinders. The regularity in the case of the elliptic p-Laplacian can be characterized by a Wiener-type test [KM], while in the p-parabolic case, no such characterization is known in general domains except for $p = 2$ [EG]. In this section, we characterize p-regular boundary points for space–time cylinders.

Consider the cylinder $\Omega = G \times (0, T)$, where G is a bounded domain in $\mathbf{R}^n$. Then each boundary point lying at the "bottom" of Ω is p-regular. Indeed, if $\xi_0 = (x_0, 0) \in G \times \{0\} \subset \partial\Omega$, then the p-parabolic function

$$w(x,t) = \frac{p-1}{p}|x - x_0|^{p/(p-1)} + nt$$

is a barrier at ξ_0. The boundary points of the bottom, i.e., points on $\partial G \times \{0\}$, are easily seen to be p-regular by employing the exterior sphere condition. Moreover, it is easily seen that none of the points (x_0, T), $x_0 \in G$, on the interior of the "top" is p-regular (cf. the example after Theorem 5.1). To give a geometric characterization for the regularity of the points (x_0, t) on the "sides" $\partial G \times (0, T]$ of the cylinder, we recall that in the elliptic theory, a boundary point $x_0 \in \partial G$ is *regular for the p-Laplacian* if

$$\lim_{x \to x_0} h_f(x) = f(x_0)$$

whenever f is a continuous function on ∂G; here h_f is the Perron solution corresponding to the boundary function of the p-Laplacian

$$\operatorname{div}(|\nabla h_f|^{p-2}\nabla h_f) = 0;$$

see [HKM]. Then a boundary point $x_0 \in \partial G$ is regular for the p-Laplacian if and only if

$$\int_0^1 \left(\frac{\operatorname{cap}_p(\overline{B}(x_0,t) \setminus G, B(x_0, 2t))}{t^{n-p}} \right)^{1/(p-1)} \frac{dt}{t} = \infty, \tag{6.3}$$

where the p-capacity $\operatorname{cap}_p(K, D)$ of a compact set K in D is defined as

$$\operatorname{cap}_p(K, D) = \inf \int_D |\nabla\varphi|^p \, dx,$$

where the infimum is taken over all $\varphi \in C_0^\infty(D)$ such that $\varphi \geq 1$ on C; see [KM], [Ma].

THEOREM 6.5. *Let $x_0 \in \partial G$ and $0 < t_0 \leq T$. Then $\xi_0 = (x_0, t_0)$ is a p-regular boundary point of G_T if and only if x_0 is regular for the p-Laplacian.*

Proof. Since the future does not have any effect on the regularity (Theorem 6.4), we are free to assume that $t_0 < T$.

Suppose first that ξ_0 is p-regular and let φ be a continuous function on ∂G. Let h_φ be the p-harmonic Perron solution in G with boundary values φ. Then the function

$$f(x,t) = \begin{cases} \varphi(x) & \text{if } 0 < t < T, \\ h_\varphi(x) & \text{if } t = 0 \text{ or } t = T \end{cases}$$

is bounded on ∂G_T and continuous at ξ_0. Moreover, if u is a p-superharmonic function from the "elliptic" upper class for φ in G, then the function $v(x,t) = u(x)$ belongs to the "parabolic" upper class for f in G_T. Hence

$$\liminf_{(x,t)\to\xi_0} \overline{H}_f(x,t) \le \liminf_{x\to x_0} h_\varphi(x).$$

By the barrier characterization for regularity, we infer from Theorem 6.1 that

$$\lim_{(x,t)\to\xi_0} \overline{H}_f(x,t) = f(\xi_0);$$

thus

$$\liminf_{x\to x_0} h_\varphi(x) \ge \varphi(x_0)$$

whenever φ is a continuous function on ∂G. Consequently,

$$\varphi(x_0) \le \liminf_{x\to x_0} h_\varphi(x) \le \limsup_{x\to x_0} h_\varphi(x) = -\liminf_{x\to x_0} h_{-\varphi}(x) \le \varphi(x_0)$$

so that x_0 is regular for the p-Laplacian.

To prove the converse, suppose that x_0 is regular for the p-Laplacian. Let $\varphi(x) = |x - x_0|$ and let u be the solution of

$$\begin{cases} \operatorname{div}(|\nabla u|^{p-2}\nabla u) = -1 \text{ in } G, \\ u - \varphi \in W_0^{1,p}(G). \end{cases}$$

Then u is p-superharmonic in G and $u(x) \ge |x - x_0|$, for φ is p-subharmonic. Because the Wiener test (6.3) is satisfied, we infer from [GZ] that

$$\lim_{x\to x_0} u(x) = \varphi(x_0) = 0;$$

therefore u is a barrier for the p-Laplacian in G.

Define

$$v(x,t) = u(x) + (t_0 - t).$$

Then

$$\operatorname{div}(|\nabla v|^{p-2}\nabla v) = -1 = v_t$$

and it follows that v is a barrier at ξ_0 with respect to

$$G_{t_0} = G \times (0, t_0).$$

Hence ξ_0 is a p-regular boundary point of G_{t_0}. Theorem 6.4 now implies that ξ_0 is a p-regular boundary point of G_T. $\square$

6.2. Petrowsky's condition. In 1933, Petrowsky established a sharp condition for the regularity of the latest moment in time on $\partial\Omega$ in connection with the heat equation; cf. [Py]. For example, for the one-dimensional heat equation $u_t = u_{xx}$, the origin is a regular boundary point of the domain defined by

$$\frac{x^2}{-4t} < \log|\log(-t)|, \quad -T < t < 0,$$

while the origin is not a regular boundary point of any domain defined by

$$\frac{x^2}{-4t} < (1+\varepsilon)\log|\log(-t)|, \quad -T < t < 0,$$

if $\varepsilon > 0$.

For the p-parabolic equation with $p > 2$ a good, but perhaps not sharp, condition is that the origin is a regular boundary point of the domain

$$\left(\frac{|x|}{(-t)^{1/\lambda}}\right)^{\frac{p}{p-1}} < A(-t)^{\frac{n(p-2)}{\lambda}}|\log(-t)|^{\alpha(p-2)}, \quad -T < t < 0,$$

where $A > 0$ and $\alpha > 0$ are any constants, and $\lambda = n(p-2)+p$ is as in the Barenblatt solution. (As $p \to 2+$ more elaborate calculations must be done if one wants to achieve the logarithms in Petrowsky's condition.)

The construction of the barrier is based on the choice

$$u(x,t) = f(t)\left(1 + \frac{p-2}{p}\lambda^{\frac{1}{1-p}}\left(\frac{|x|}{(-t)^{1/\lambda}}\right)^{\frac{p}{p-1}}\right)^{\frac{p-1}{p-2}} + \varphi(t),$$

where

$$f(t) = -a|\log(-t)|^{\alpha},$$

$a > 0$, and

$$\varphi(t) = a|\log(-t)|^{\alpha} + C(-t)^{1-\frac{p}{\lambda}}|\log(-t)|^{\alpha(p-1)}.$$

A rather straightforward but lengthy calculation shows that

$$u_t - \operatorname{div}(|\nabla u|^{p-2}\nabla u) \geq 0$$

in the domain, where $u(x,t)$ is positive. Hence u is p-superparabolic in a convenient domain.

This indicates how the barrier is constructed. However, the proof must be omitted here, since the actual calculations would lead us way astray.

Note added in proof. It has come to our attention that the p-parabolic equation has a very strong physical application. The Barenblatt solution describes the propagation of the heat after the explosion of a hydrogen bomb in the atmosphere.

REFERENCES

[AL] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.

[AS] D. G. Aronson and J. Serrin, *Local behavior of solutions of quasilinear parabolic equations*, Arch. Rational Mech. Anal., 25 (1967), pp. 81–122.

[B] G. I. BARENBLATT, *On selfsimilar motions of compressible fluids in a porous medium*, Prikl. Mat. Mekh., 16 (1952), pp. 679–698 (in Russian).

[DB] E. DIBENEDETTO, *Degenerate Parabolic Equations*, Springer-Verlag, New York, 1993.

[EV] J. R. ESTEBAN AND J. L. VÁSQUEZ, *Homogeneous diffusion in* **R** *with power- like nonlinear diffusivity*, Arch. Rational Mech. Anal., 103 (1988), pp. 39–80.

[EG] L. C. EVANS AND R. F. GARIEPY, *Wiener's criterion for the heat equation*, Arch. Rational Mech. Anal., 78 (1982), pp. 293–314.

[F] A. FRIEDMAN, *Parabolic equations of the second order*, Trans. Amer. Math. Soc., 93 (1959), pp. 509–530.

[GZ] R. GARIEPY AND W. P. ZIEMER, *A regularity condition at the boundary for solutions of quasilinear elliptic equations*, Arc. Rational Mech. Anal., 67 (1977), pp. 25–39.

[G] F. W. GEHRING, *On solutions of the equation of heat conduction*, Michigan Math. J., 5 (1958), pp. 191–202.

[GLM] S. GRANLUND, P. LINDQVIST, AND O. MARTIO, *Note on the PWB-method in the nonlinear case*, Pacific J. Math., 125 (1986), pp. 381–395.

[HKM] J. HEINONEN, T. KILPELÄINEN, AND O. MARTIO, *Nonlinear Potential Theory of Degenerate Elliptic Equations*, Oxford University Press, Oxford, 1993.

[KV] S. KAMIN AND J. L. VÁSQUEZ, *Fundamental solutions and asymptotic behaviour for the p-Laplacian equation*, Rev. Math. Iberoamericana, 4 (1988), pp. 339–354.

[KM] T. KILPELÄINEN AND J. MALÝ, *The Wiener test and potential estimates for quasilinear elliptic equations*, Acta Math., 172 (1994), pp. 137–161.

[L] J. L. LIONS, *Quelques Méthodes de Résolution des Problémes aux Limites Non Linéaires*, Dunod, Paris, 1969.

[Ma] V. G. MAZ'YA, *On the continuity at a boundary point of solutions of quasi-linear elliptic equations*, Vestnik Leningrad Univ. Mat. Mekh. Astronom., 25 (1970), pp. 42–55 (in Russian); Vestnik Leningrad Univ. Math., 3 (1976), pp. 225–242 (English translation).

[M] J. MOSER, *A Harnack inequality for parabolic differential equations*, Comm. Pure Appl. Math., 27 (1964), pp. 101–134; *Correction*, Comm. Pure Appl. Math., 20 (1967). pp. 231–236.

[Pn] O. PERRON, *Eine neue Behandlung der Randwertaufgabe der* $\Delta u = 0$, Math. Z., 18 (1923), pp. 42–54.

[Py] I. PETROWSKY, *Zur ersten Randwertaufgabe der Wärmeleitungsgleichung*, Compositio Math., 1 (1935), pp. 383–419.

[S] W. STERNBERG, *Über die Gleichung der Wärmeleitung*, Math. Ann., 101 (1929), pp. 394–398.

[T] N. S. TRUDINGER, *Pointwise estimates and quasilinear parabolic equations*, Comm. Pure. Appl. Math., 21 (1968), pp. 205–226.

[W] N. A. WATSON, *Green functions, potentials, and the Dirichlet problem for the heat equation*, Proc. London Math. Soc., (3), 33 (1976), pp. 251–298; *Corrigendum*, Proc. London Math. Soc. (3), 37 (1977), pp. 32–34.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 684–707, May 1996

005

ON THE SIZE AND SMOOTHNESS OF SOLUTIONS TO NONLINEAR HYPERBOLIC CONSERVATION LAWS*

RONALD A. DEVORE[†] AND BRADLEY J. LUCIER[‡]

Abstract. We address the question of which function spaces are invariant under the action of scalar conservation laws in one and several space dimensions. We establish two types of results. The first result shows that if the initial data is in a rearrangement-invariant function space, then the solution is in the same space for all time. Secondly, we examine which smoothness spaces among the Besov spaces are invariant for conservation laws. Previously, we showed in one dimension that if the initial data has bounded variation and the flux is convex and smooth enough, then the Besov spaces $B^\alpha_q(L_q)$, $\alpha > 1$, $q = 1/(\alpha+1)$, are invariant smoothness spaces. Now, in one space dimension, we show that no other Besov space with $\alpha > 1$ is invariant. In several space dimensions, we show that no Besov space $B^\alpha_q(L_q)$ with $\alpha > 1$ is invariant. Combined with previous results, our theorems completely characterize for $\alpha > 1$ which Besov spaces are smoothness spaces for scalar conservation laws.

Key words. conservation laws, regularity, rearrangement-invariant spaces, Besov spaces

AMS subject classifications. 36L65, 35B65, 35D10, 46E30, 46E35

1. Introduction.

We are interested in the size and smoothness, as measured in certain function spaces, of solutions $u(x,t)$, $x \in \mathbb{R}^d$, $t \geq 0$, to the scalar hyperbolic conservation law

$$\begin{aligned} &u_t + \nabla_x \cdot f(u) = 0, \quad x \in \mathbb{R}^d,\ t > 0, \\ &u(x,0) = u_0(x), \qquad x \in \mathbb{R}^d. \end{aligned} \tag{1.1}$$

Here, the flux f maps $\mathbb{R}$ into $\mathbb{R}^d$ and $\nabla_x \cdot f(u)$ denotes the divergence of $f(u(x,t))$ with respect to the spatial variables $x \in \mathbb{R}^d$. In general, classical solutions to (1.1) do not exist for all time $t > 0$; indeed, at some time $t > 0$, which depends on f and u_0, the solution u to (1.1) will generally develop discontinuities known as "shocks" even if the flux and the initial condition are smooth. One defines weak solutions to (1.1) as functions $u(x,t)$ that satisfy

$$-\int_0^t \int_{\mathbb{R}^d} [u(x,t)\phi_t(x,t) + f(u(x,t)) \cdot \nabla_x\phi(x,t)]\,dx\,dt + \int_{\mathbb{R}^d} u(x,T)\phi(x,T)\,dx - \int_{\mathbb{R}^d} u(x,0)\phi(x,0)\,dx = 0$$

for all $\phi \in C^1(\mathbb{R}^{d+1})$ with compact support. Weak solutions are not unique; however, by imposing restrictions, known as entropy conditions, on weak solutions u, it is possible to select from these weak solutions the physically relevant solution to (1.1). See [10], [16]. When we speak about the solution to (1.1), we shall mean this entropy solution.

* Received by the editors April 29, 1994; accepted for publication September 21, 1994. This research was supported by Office of Naval Research contracts N00014-91-J-1076 and N00014-91-J-1152.

[†] Department of Mathematics, University of South Carolina, Columbia, SC 29208 (devore@math.scarolina.edu).

[‡] Department of Mathematics, Purdue University, West Lafayette, IN 47907-1395 (lucier@math.purdue.edu).

We are interested in the question of which function spaces X, such as the spaces of p-integrable functions $L_p(\mathbb{R}^d)$ or the space of functions of bounded variation $\mathrm{BV}(\mathbb{R}^d)$, are invariant under the differential equation. This means that if we denote the mapping $u_0 \mapsto u(\,\cdot\,,t)$ for fixed $t>0$ by $E(t)$ (i.e., $E(t)u_0 = u(\,\cdot\,,t)$), then we are interested in function spaces X for which there exists a constant C such that for for all $u_0 \in X$

$$\|E(t)u_0\|_X \le C\|u_0\|_X.$$

If the X norm (or quasinorm, or seminorm) in some sense measures smoothness of functions, then we call X a regularity space for (1.1).

We first address the question of how to measure the size of solutions to (1.1). For all convex functions $\eta\colon \mathbb{R} \to \mathbb{R}$ and under suitable conditions on f, entropy solutions $u(x,t)$ of (1.1) satisfy

$$\int_{\mathbb{R}^d} \eta(u(x,t))\,dx \le \int_{\mathbb{R}^d} \eta(u_0(x))\,dx$$

for all $t>0$; see, e.g., [11]. By setting $\eta(u) = |u|^p$ for $1 \le p < \infty$, one sees immediately that

$$\|u(\,\cdot\,,t)\|_{L_p(\mathbb{R}^d)} \le \|u_0\|_{L_p(\mathbb{R}^d)}. \tag{1.2}$$

One can show independently that (1.2) holds also for $p=\infty$. Thus, $L_p(\mathbb{R}^d)$, $1 \le p \le \infty$, are invariant spaces for solutions $u(x,t)$ of (1.1). Actually, one can prove somewhat more, as we now discuss.

The solution operator $E(t)$ of (1.1) is not only bounded on $L_1(\mathbb{R}^d)$, but is a contraction in $L_1(\mathbb{R}^d)$; i.e., if u_0 and v_0 are two initial conditions for (1.1), then

$$\|u(\,\cdot\,,t) - v(\,\cdot\,,t)\|_{L_1(\mathbb{R}^d)} \le \|u_0 - v_0\|_{L_1(\mathbb{R}^d)}, \quad t>0. \tag{1.3}$$

Thus, the nonlinear mapping $E(t)$ is a contraction on $X := L_1(\mathbb{R}^d)$ and is bounded on $Y := L_\infty(\mathbb{R}^d)$. A simple argument, given in §4, shows that if a possibly nonlinear mapping $E(t)$ defined on $X+Y := \{\, f+g \mid f \in X,\ g \in Y \,\}$ is a contraction on X and is bounded on Y, then $E(t)$ is a bounded mapping on all interpolation spaces between X and Y as determined by the method of real interpolation. First results on conservation laws show that $E(t)u_0$ is defined for each locally integrable function u_0, and, in particular, for each function u_0 in $L_1(\mathbb{R}^d) + L_\infty(\mathbb{R}^d)$, if f is globally Lipschitz continuous. In our case, this means that $E(t)$ is bounded on all interpolation spaces between $L_1(\mathbb{R}^d)$ and $L_\infty(\mathbb{R}^d)$, and, in particular on $L_p(\mathbb{R}^d)$ for $1<p<\infty$. In addition, many other spaces, such as the Lorentz spaces $L_{p,q}(\mathbb{R}^d)$, $1 \le p \le \infty$, $1 < q \le \infty$, and the Orlicz spaces defined on $\mathbb{R}^d$, are interpolation spaces for $L_1(\mathbb{R}^d)$ and $L_\infty(\mathbb{R}^d)$. Calderón [2] characterized the interpolation spaces between $L_1(\mathbb{R}^d)$ and $L_\infty(\mathbb{R}^d)$ as the set of all *rearrangement-invariant* function spaces on $\mathbb{R}^d$; see §4. Thus, solutions of (1.1) are bounded on all rearrangement-invariant function spaces on $\mathbb{R}^d$.

The main focus of this paper is the smoothness, or regularity, of solutions to (1.1), which we next describe. Because of the appearance of shocks, $E(t)$ does not map $C(\mathbb{R}^d)$ into itself. The question arises whether there is any other sense in which the solution of (1.1) retains smoothness.

One should note that solutions of (1.1) are translation invariant, i.e., $E(t)(u_0)(x+h) = E(t)(u_0(\cdot + h))(x)$ for all x and h in $\mathbb{R}^d$. Thus, from (1.3) we see that for all $h \in \mathbb{R}^d$,

$$\|u(\cdot + h, t) - u(\,\cdot\,,t)\|_{L_1(\mathbb{R}^d)} \le \|u_0(\cdot + h) - u_0\|_{L_1(\mathbb{R}^d)}. \tag{1.4}$$

Therefore, with the usual norm for the Lipschitz spaces $\mathrm{Lip}(\alpha, L^p(\mathbb{R}^d))$ given by

$$\|v\|_{\mathrm{Lip}(\alpha,L^p(\mathbb{R}^d))} := \sup_{0\neq h\in\mathbb{R}^d} |h|^{-\alpha}\|v(\cdot+h)-v\|_{L_p(\mathbb{R}^d)}, \quad 0<\alpha\leq 1,\ 0<p\leq\infty,$$

(1.4) implies immediately that

$$\|u(\,\cdot\,,t)\|_{\mathrm{Lip}(\alpha,L^1(\mathbb{R}^d))} \leq \|u_0\|_{\mathrm{Lip}(\alpha,L^1(\mathbb{R}^d))}$$

for all $0<\alpha\leq 1$. In particular, the set of functions of bounded variation on $\mathbb{R}^d$, $\mathrm{BV}(\mathbb{R}^d)$, is invariant under $E(t)$, since, by definition, $\mathrm{Lip}(1, L^1(\mathbb{R}^d)) = \mathrm{BV}(\mathbb{R}^d)$.

The Lipschitz spaces $\mathrm{Lip}(\alpha, L^p(\mathbb{R}^d))$, $0<\alpha<1$, are special cases of the more general Besov spaces $B^\alpha_q(L_p(\mathbb{R}^d))$ (see §3 for a definition), which depend on three parameters $0<\alpha<\infty$, $0<p\leq\infty$, and $0<q\leq\infty$. In fact, $\mathrm{Lip}(\alpha, L^p(\mathbb{R}^d)) = B^\alpha_\infty(L_p(\mathbb{R}^d))$ for $0<\alpha<1$ and $0<p\leq\infty$. For a Besov space, the parameter α determines the order of smoothness (roughly speaking, the number of derivatives). The second parameter p specifies the space in which smoothness is measured, namely $L_p(\mathbb{R}^d)$. The third parameter q allows one to make subtle distinctions in smoothness. Of special interest are the spaces $B^\alpha_p := B^\alpha_p(L_p(\mathbb{R}^d))$. These are sometimes called fractional-order Sobolev spaces because of their similarity to the classical Sobolev spaces; in fact, B^r_2 is identical to the Sobolev space W^r_2, $r=1,2,\dots$, consisting of functions from $L_2(\mathbb{R}^d)$ that have all of their rth distributional derivatives in $L_2(\mathbb{R}^d)$.

The main interest of the present paper is the classification of all Besov spaces $X = B^\alpha_p$, $0<p\leq\infty$, $0<\alpha<\infty$, for which $u_0\in X$ implies $u(\,\cdot\,,t)\in X$ for all later times $t>0$. We shall determine all the regularity spaces of conservation laws among these Besov spaces except for a certain set of values of α and p (with $0\leq\alpha\leq 1$). We use the remainder of this introduction to formulate and explain our results.

We have noted that $\mathrm{Lip}(\alpha, L^1(\mathbb{R}^d))$, $0<\alpha\leq 1$, is a regularity space for (1.1) with constant $C=1$. Once one knows the definitions, it is easily shown that all the Besov spaces $B^\alpha_q(L_1(\mathbb{R}^d))$, $0<\alpha<1$, $0<q\leq\infty$, are regularity spaces, again with $C=1$.

Perhaps somewhat more surprising is the fact established in [6], [7], and [14] that the spaces $\mathrm{BV}\cap B^\alpha_{\tau(\alpha)}$, $\tau(\alpha) := (\alpha+1)^{-1}$, $\alpha>0$, are regularity spaces in one space dimension provided that the flux f is suitably smooth. In some cases, for example for the inviscid Burgers equation, the space $B^\alpha_{\tau(\alpha)}$ is itself a regularity space (i.e., it is not necessary to intersect this space with BV).

With these results in hand, the question arises whether any other Besov spaces are regularity spaces for (1.1), to which we now attend. To explain the results of the present paper, it is useful to give a diagram that organizes our knowledge of smoothness spaces. We identify any smoothness space with smoothness α in $L_p(\mathbb{R}^d)$, and in particular the Besov space B^α_p, with the point $(1/p, \alpha)$ in the upper-right quadrant of $\mathbb{R}^2$. The classification of Besov spaces as regularity spaces in one space dimension can then be visualized as in Figure 1. The line segment connecting $(0,0)$ to $(1,0)$ represents the $L_p(\mathbb{R})$ spaces, or more generally the rearrangement-invariant spaces (which we show in §4 are invariant spaces for (1.1)). The line segment with endpoints $(1,0)$ and $(1,1)$ represents the regularity spaces $\mathrm{Lip}(\alpha, L^1(\mathbb{R}))$ or $B^\alpha_q(L_1(\mathbb{R}^d))$, $0<\alpha<1$, $0<q\leq\infty$. The half-line $\mathcal{L}_1$ with slope one emanating from $(1,0)$ represents the regularity spaces B^α_τ, $\tau=\tau(\alpha)=(\alpha+1)^{-1}$ already discussed. Each space to the right of the line $\mathcal{L}_1$ contains some functions that are not locally integrable and hence the conservation law does not have a solution for all initial values from these spaces (i.e., the regularity question does not have a meaning). The line $\mathcal{L}_0$ emanating from the origin with slope one separates spaces embedded in $C(\mathbb{R})$ (those above the line)

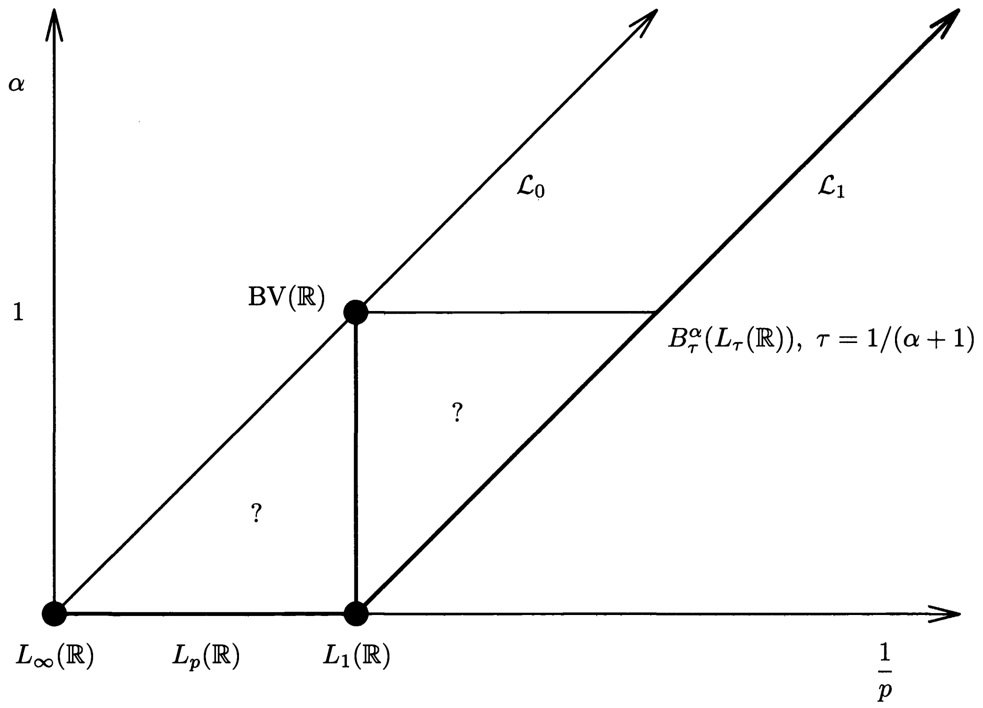


FIG. 1. *Smoothness spaces in one space dimension. The line segment from* $(1,0)$ *to* $(1,1)$ *represents the spaces* $B^\alpha_q(L_1(\mathbb{R}))$, $0 < \alpha < 1$, $0 < q \le \infty$. *The Besov spaces in the open regions not marked by question marks cannot be regularity spaces for* (1.1).

and those not embedded in $C(\mathbb{R})$(below the line). It follows that any space above the line $\mathcal{L}_0$ cannot be a regularity space for (1.1) since, in general, continuous data u_0 generate solutions with shocks. We prove in the present paper that for $\alpha > 1$, none of the Besov spaces $B^\alpha_p(L_p(\mathbb{R}))$ are regularity spaces except for the spaces $B^\alpha_{\tau(\alpha)}$, which corresponds to points on $\mathcal{L}_1$.

In one space dimension, we have not determined whether the spaces represented by points in the parallelogram with vertices $(0,0)$, $(1,0)$, $(1,1)$, and $(2,1)$ are regularity spaces for (1.1). All the same, we conjecture that all these spaces are regularity spaces for (1.1). We are able to use techniques from approximation theory and interpolation of operators to prove this for certain points in this region but will not report on this here since the results are not complete and the arguments are quite technical. One would hope that some nonlinear interpolation argument would settle all these cases.

Our results concerning regularity in one space dimension described above are, for the most part, negative. It is possible, however, to prove some positive results for regularity in one space dimension. We show, using a general argument, that if f is uniformly convex and smooth enough and if $u_0 \in \mathrm{BV}(\mathbb{R}) \cap B^\alpha_\sigma(L_\sigma(\mathbb{R}))$ for $\alpha > 1$ and $\sigma > \tau(\alpha) = 1/(\alpha+1)$, then for all $t > 0$, $u(\cdot, t)$ is in every Besov space $B^\beta_q(L_p(\mathbb{R}))$ with $(1/p, \beta)$ strictly inside the quadrilateral with corners $(0,0)$, $(1,0)$, $(1,1)$, and $(1/\tau(\alpha), \alpha)$.

The situation regarding regularity spaces in several space dimension is quite different. One might suspect that the Besov spaces B^α_τ, $\tau := (\alpha/d + 1)^{-1}$ are regularity spaces for space dimension $d > 1$, since their one-dimensional counterparts ($d = 1$) are. (These are precisely the spaces of minimal smoothness that are embedded into

$L_1(\mathbb{R}^d)$, by a variant of the Sobolev embedding theorem.) However, we shall show in this paper that under very general conditions on the flux f, none of the Besov spaces $B_q^\alpha(L_p(\mathbb{R}^d))$ with $\alpha > 1$ are regularity spaces for (1.1). Again, we are not able to completely settle the case $0 < \alpha \leq 1$. In $\mathbb{R}^d$ the lines $\mathcal{L}_0$ and $\mathcal{L}_1$ are to be replaced by $\mathcal{L}_0'$ and $\mathcal{L}_1'$, which emanate from $(0,0)$ and $(1,0)$, respectively, with slope d. We already noted that the spaces on the line segment with endpoints $(1,0)$, $(1,1)$ are regularity spaces.

We feel that our negative results on regularity spaces in several space dimensions give useful information about the structure of solutions to (1.1) and the behavior of numerical methods for their solutions. In order to bring out this point, we first make a few remarks about the connections between regularity and numerical methods. A typical numerical method creates for discrete time values t_n an approximation u_n to $u(\,\cdot\,,t_n)$. The approximants u_n will lie in certain linear or nonlinear spaces Σ_n associated with the numerical method. Usually, Σ_n is a space of piecewise polynomials with either a fixed or variable grid. The approximation power of such a method is often (although not always) associated with the approximation power from the space Σ_n. In any case, no numerical method can approximate better than the best approximation from Σ_n. The order of best approximation by elements in Σ_n is characterized by the smoothness of the function $u(\,\cdot\,,t_n)$ being approximated. It is therefore important to understand when the solution $u(\,\cdot\,,t)$ is in the smoothness space associated with a given order of best approximation by elements in Σ_n. For example, the regularity spaces $B^\alpha_{\tau(\alpha)}$ characterize the classes of functions that can be approximated with a given approximation order $N^{-\alpha}$ in $L_1(\mathbb{R})$ by free-knot spline functions with N knots. This shows that numerical methods based on moving grids (in one space dimension) should be effective in recovering solutions to scalar conservation laws. This is indeed the case, as is shown in [12], [6], and [14], where numerical methods based on moving grid finite elements are constructed that provide approximation order α for any $\alpha > 0$.

In the multivariate case, there are no regularity spaces among the Besov spaces $B_p^\alpha(L_p(\mathbb{R}^d))$, $\alpha > 1$. The Besov spaces are homogeneous: they measure regularity the same way in all coordinate directions. From another viewpoint, these Besov spaces are characterized by very regular approximation processes such as approximation by wavelets or splines with regular partitions. The elements in these spaces behave the same in all coordinate directions. On the other hand, the "fluid transport" in conservation laws can be very directionally dependent. To approximate well such a solution at later time $t > 0$ requires finer resolution in directions where mass is accumulating. For example, if we were approximating by piecewise constants, we would need elements that are finer in certain directions and coarser in others. This is not possible with splines on regular partitions or wavelets. This is reflected in the fact that their approximation spaces (the Besov spaces) are not regularity spaces for conservation laws in several space dimensions.

2. Properties of entropy solutions to conservation laws. We begin by recalling certain properties of the solution to (1.1) that will be used in what follows. Let $E(t)$ denote the evolution operator that associates to u_0 the solution $E(t)u_0 := u(\,\cdot\,,t)$ of (1.1) at time $t > 0$. Then $E(t)$ maps $L_1(\mathbb{R}^d) + L_\infty(\mathbb{R}^d)$ into itself. Moreover, $E(t)$ is a norm-one, bounded operator on $L_1(\mathbb{R}^d)$ and $L_\infty(\mathbb{R}^d)$:

$$\|E(t)u_0\|_{L_p(\mathbb{R}^d)} \leq \|u_0\|_{L_p(\mathbb{R}^d)}, \quad p = 1, \infty. \tag{2.1}$$

Actually, $E(t)$ is a bounded operator with norm one on each of the $L_p(\mathbb{R}^d)$ spaces

$1 \le p \le \infty$. This can be derived from the study of entropy–entropy flux pairs or proved as in §4.

The operator $E(t)$ is monotone in the sense that $E(t)u_0 \ge E(t)v_0$ whenever $u_0 \ge v_0$ (see, for example, [3]) and preserves integrals

$$\int_{\mathbb{R}^d} E(t)u_0\, dx = \int_{\mathbb{R}^d} u_0\, dx.$$

From these properties, one derives that $E(t)$ is a contraction (see [3])

$$\|E(t)(u_0) - E(t)(v_0)\| \le \|u_0 - v_0\|.$$

Here and later, the unsubscripted norm $\| \cdot \|$ always denotes the $L_1(\mathbb{R}^d)$ norm.

There is no simple description of the solution u of (1.1). However, in one space dimension, the following method of Lax [11] gives a useful analytic method for obtaining u. We assume that the flux f is strictly convex. It follows that the transport velocity $a(u) := f'(u)$ is strictly increasing on $\mathbb{R}$ and is therefore invertible (under composition of functions) on $\mathbb{R}$. We assume further that the initial condition u_0 is continuous with compact support. Any initial condition can be approximated to arbitrary accuracy in the $L_1(\mathbb{R})$ norm by such functions. Under these conditions, Lax [11] shows that the solution $u(x,t)$ of (1.1) can be described by

$$u(x,t) = u_0(y(x,t)), \tag{2.2}$$

where $y = y(x,t)$ satisfies

$$\frac{x-y}{t} = a(u_0(y)). \tag{2.3}$$

In general, there are many solutions y to (2.3). The one that satisfies (2.2) is determined as the solution to an extremal problem (cf. Theorem 3.1 in [11]).

Lax establishes various properties of the selection $y(x,t)$. In particular, he shows that for each fixed $t > 0$,

$$y(\,\cdot\,,t) \text{ is increasing on } \mathbb{R}. \tag{2.4}$$

Shocks occur in the solution to (1.1) at points where $y(\,\cdot\,,t)$ discontinuous. This can occur when there is more than one solution y to (2.3) and we jump down from one piece of the graph of u_0 to another.

3. Besov spaces. In this section, we give the definition of Besov spaces and several equivalent norms for these spaces, which will be used in the sequel. For $\alpha > 0$ and $0 < p, q \le \infty$, the Besov space $B_q^\alpha(L_p(\mathbb{R}^d))$ is a space of functions with smoothness α in L_p. The secondary parameter q gives a finer gradation of these spaces that is important in many applications.

To describe these spaces, we use the difference operators Δ_h^r, $r = 1, 2, \dots$, with step $h \in \mathbb{R}^d$. These are defined inductively with $\Delta_h(v,x) := v(x+h) - v(x)$ and $\Delta_h^r := \Delta_h \Delta_h^{r-1}$. It follows that

$$\Delta_h^r(v,x) := \sum_{j=0}^{r} (-1)^{r+j} \binom{r}{j} v(x+jh). \tag{3.1}$$

With these differences, we can define the moduli of smoothness

$$\omega_r(v,s)_p := \sup_{0 \le |h| \le s} \|\Delta_h^r(v)\|_{L_p(\mathbb{R}^d)}, \quad s \ge 0,$$

for each $r = 1, 2, \ldots$. The rate at which $\omega_r(v,t)_p$ tends to zero gives information about the smoothness of v in $L_p(\mathbb{R}^d)$.

The Besov spaces are defined for $0 < \alpha < r$ and $0 < p, q \le \infty$ as the set of all functions $v \in L_p(\mathbb{R}^d)$ for which

$$|v|_{B_q^\alpha(L_p(\mathbb{R}^d))} := \begin{cases} \left(\int_0^\infty \left[s^{-\alpha} \omega_r(v,s)_p \right]^q \frac{ds}{s} \right)^{\frac{1}{q}}, & 0 < q < \infty, \\ \sup_{s \ge 0} s^{-\alpha} \omega_r(v,s)_p, & q = \infty, \end{cases} \tag{3.2}$$

is finite. The conditions (3.2) require that $\omega_r(v,s)_p$ behave like $O(s^\alpha)$ as $s \to 0$; the exact requirement on $\omega_r(v,s)_p$ varies with q and becomes stronger as q gets smaller. We define the following "norm" for $B_q^\alpha(L_p(\mathbb{R}^d))$:

$$\|v\|_{B_q^\alpha(L_p(\mathbb{R}^d))} := \|v\|_{L_p(\mathbb{R}^d)} + |v|_{B_q^\alpha(L_p(\mathbb{R}^d))}.$$

Because we allow p and q to be less than 1, this "norm" does not always satisfy the triangle inequality, but it is always a quasinorm, i.e., there exists a constant C such that for all u and v in $B_q^\alpha(L_p(\mathbb{R}^d))$,

$$\|u + v\|_{B_q^\alpha(L_p(\mathbb{R}^d))} \le C(\|u\|_{B_q^\alpha(L_p(\mathbb{R}^d))} + \|v\|_{B_q^\alpha(L_p(\mathbb{R}^d))}).$$

It can be shown that the above definition of Besov spaces does not depend on the choice of r, since all values of $r > \alpha$ give rise to equivalent norms and hence the same space. We note that since $\omega_r(v,t)_p \le 2^r \|v\|_{L_p(\mathbb{R}^d)}$, $1 \le p \le \infty$ (the same inequality holds with 2^r replaced by $2^{r/p}$ when $p < 1$), we obtain an equivalent norm for $B_q^\alpha(L_p(\mathbb{R}^d))$ if we take the integral or supremum in (3.2) over only the interval $[0,1]$. Thus, membership of v in $B_q^\alpha(L_p(\mathbb{R}^d))$ is determined only by the integral or supremum on $[0,1]$.

For certain values of the parameters, the Besov spaces are identical with other smoothness spaces. For example, if $0 < p \le \infty$ and $0 < \alpha$ is not an integer, then $B_\infty^\alpha(L_p(\mathbb{R}^d)) = \mathrm{Lip}(\alpha, L_p(\mathbb{R}^d))$ are the classical Lipschitz spaces. When $\alpha = k$ is an integer, we obtain the generalized Lipschitz spaces, for which $\omega_r(v,s)_p$, $r > k$, is used in place of $\omega_k(v,s)_p$ in the definition of the usual Lipschitz spaces. For $p = 2$ and $\alpha > 0$, the Besov spaces $B_2^\alpha(L_2(\mathbb{R}^d)) = W^\alpha(L_2(\mathbb{R}^d))$ are the Sobolev spaces. The Besov spaces with $q = p$, which we shall denote by $B_p^\alpha := B_p^\alpha(L_p(\mathbb{R}^d))$, are of particular interest. These are sometimes called generalized Sobolev spaces.

The application of Besov spaces to approximation theory and interpolation of linear operators leads to alternate characterizations of these spaces. We shall mention two of these alternate characterizations that hold in the univariate case.

The first characterization describes the Besov spaces in terms of approximation by linear spaces of spline functions. Let $\mathcal{S}_{n,r}$ denote the set of all univariate piecewise polynomials of degree $< r$ that have global smoothness $C^{(r-2)}$ and have break points only at the dyadic integers $j2^{-n}$, $j \in \mathbb{Z}$. For each $f \in L_p(\mathbb{R})$, we define

$$s_n(f)_p := s_{n,r}(f)_p := \inf_{S \in \mathcal{S}_{n,r}} \|f - S\|_{L_p(\mathbb{R})}.$$

Then (see [8]), for $0 < \alpha < r$, a function f is in $B_q^\alpha(L_p(\mathbb{R}))$ if and only if

$$\left(\sum_{n=0}^\infty [2^{n\alpha} s_n(f)_p]^q \right)^{\frac{1}{q}} < \infty, \tag{3.3}$$

with the usual change to a supremum when $q = \infty$. Moreover, (3.3) gives an equivalent seminorm for $B_q^\alpha(L_p(\mathbb{R}))$. It follows directly from (3.3) that $B_\sigma^\alpha(L_p(\mathbb{R}))$ is continuously embedded in $B_\tau^\beta(L_p(\mathbb{R}))$ if $\alpha > \beta$ or $\alpha = \beta$ and $\sigma < \tau$.

Our second characterization describes the Besov spaces in terms of univariate nonlinear approximation. For this, we let $\Sigma_n := \Sigma_{n,r}$ denote the collection of all piecewise-polynomial functions of degree $< r$ on $\mathbb{R}$ that consist of at most $n+1$ pieces. Thus $S \in \Sigma_n$ if and only if there exist n breakpoints $x_1 < x_2 < \cdots < x_n$ such that with $x_0 := -\infty$, $x_{n+1} := \infty$, the function S is a polynomial of degree $< r$ on each of the intervals (x_{i-1}, x_i), $i = 1, \ldots, n$. No assumption is made about the smoothness of S at the breakpoints. The set Σ_n is not a linear space, but it can be considered a nonlinear manifold parameterized by the breakpoints and the coefficients of the polynomial pieces.

We can describe certain Besov spaces in terms of their approximation by the elements of Σ_n. For this, we define for $f \in L_p(\mathbb{R})$, $0 < p \leq \infty$,

$$\sigma_n(f)_p := \sigma_{n,r}(f)_p := \inf_{S \in \Sigma_n} \|f - S\|_{L_p(\mathbb{R})}, \tag{3.4}$$

which is the error in approximating f in the $L_p(\mathbb{R})$ norm by the elements of Σ_n.

Nonlinear spline approximation can be used to characterize certain of the spaces B_τ^α (see [15] and [9]). If $r > \alpha > 0$ and $\tau > 0$ are given, and if there is a p with $0 < p < \infty$ such that $\tau = \tau(\alpha, p) := (\alpha + 1/p)^{-1}$, then $f \in B_\tau^\alpha(L_\tau(\mathbb{R}))$ if and only if

$$\left(\sum_{n=1}^{\infty} [2^{n\alpha} \sigma_{2^n}(f)_p]^\tau\right)^{\frac{1}{\tau}} < \infty \tag{3.5}$$

and (3.5) when added to $\|\cdot\|_{L_\tau(\mathbb{R})}$ gives an equivalent norm for B_τ^α:

$$\|f\|_{B_\tau^\alpha(L_\tau(\mathbb{R}))} \sim \|f\|_{L_\tau(\mathbb{R})} + \left(\sum_{n=1}^{\infty} [2^{n\alpha} \sigma_{2^n}(f)_p]^\tau\right)^{\frac{1}{\tau}}. \tag{3.6}$$

Furthermore, (3.6) implies that $B_{\tau(\alpha,p)}^\alpha$ is continuously embedded in $B_{\tau(\beta,p)}^\beta$ if $\alpha > \beta$. Indeed, $B_{\tau(\alpha,p)}^\alpha$ is continuously embedded in $L_p(\mathbb{R})$ (see [8]), and the family of spaces $B_{\tau(\alpha,p)}^\alpha$ lies on the half-line with slope one emanating from the point $(1/p, 0)$ in Figure 1.

It may be useful to say a few words about the differences in the two characterizations (3.3) and (3.5). The characterization (3.3) describes the space $B_q^\alpha(L_p(\mathbb{R}^d))$ in terms of approximation in $L_p(\mathbb{R}^d)$. Thus the approximation is taking place in the same space ($L_p(\mathbb{R}^d)$) in which the smoothness is measured. In contrast, in (3.5) the approximation takes place in the space $L_p(\mathbb{R}^d)$ but the smoothness is measured in the space $L_\tau(\mathbb{R}^d)$; this is characteristic of nonlinear approximation. Since $\tau < p$, the class of functions that can be approximated by the nonlinear family Σ_n is larger than the class that is approximated by the linear spaces $\mathcal{S}_n$.

4. Rearrangement invariant spaces. In this section, we shall give an elementary approach to finding invariant spaces based on interpolation of operators. We begin by recalling some basic facts about the K-functional and its application to the theory of interpolation of operators.

If (X_0, X_1) is a pair of complete, quasinormed spaces embedded in a Hausdorff space $\mathcal{X}$, then for each $v \in X_0 + X_1$, we form the K-functional

$$K(v,s) := K(v,s;X_0,X_1) := \inf_{v=v_0+v_1} \{\|v_0\|_{X_0} + s\|v_1\|_{X_1}\}, \quad s \geq 0,$$

where the infimum is taken over all decompositions $v = v_0 + v_1$ with $v_i \in X_i$, $i = 0, 1$. It is easy to see that for fixed $s > 0$, $K(\,\cdot\,, s)$ is a quasinorm and for fixed v, $K(v, \,\cdot\,)$ is an increasing concave function on $\mathbb{R}_+$.

The K-functional was introduced by Peetre as a tool for obtaining interpolation spaces X for the pair (X_0, X_1). We recall that a complete quasinormed space X contained in $X_0 + X_1$ is called an *interpolation space* for the pair (X_0, X_1) if each linear operator T that boundedly maps X_0 and X_1 into themselves also maps X boundedly into itself. For such a T, it follows that

$$K(Tv, s; X_0, X_1) \leq M\, K(v, s; X_0, X_1), \quad s > 0,$$

with M the maximum of the norm of T on the two spaces X_0, X_1.

In view of (4.1), we can obtain interpolation spaces for (X_0, X_1) by applying to $K(v, \,\cdot\,)$ a quasinorm defined for functions on $\mathbb{R}_+$. We mention in particular the (θ, q) norms, $0 < \theta < 1$, $0 < q \leq \infty$, which give the spaces $X_{\theta,q}$ that consist of all $v \in X_0 + X_1$ for which

$$(4.1) \qquad |v|_{X_{\theta,q}} := \left(\int_0^\infty [s^{-\theta} K(v,s)]^q \frac{ds}{s} \right)^{1/q}$$

is finite (with the usual change to a supremum when $q = \infty$). It follows from (4.1) that T maps $X_{\theta,q}$ into itself for each $0 < \theta < 1$, $0 < q \leq \infty$, with a norm not exceeding M.

It is not possible to apply this interpolation directly to the operator $E(t)$ associated with (1.1) since it is not linear. However, the following simple remark can be used in place of linearity. We say that an operator T is X_0-Lipschitz on $X_0 + X_1$ if

$$\|T(v_0) - T(v_1)\|_{X_0} \leq M_0 \|v_0 - v_1\|_{X_0}$$

for each $v_0, v_1 \in X_0 + X_1$ for which $v_1 - v_0 \in X_0$.

LEMMA 4.1. *If T is a (possibly nonlinear) operator that is X_0-Lipschitz with constant M_0 on $X_0 + X_1$ and is bounded with norm M_1 on X_1, then T satisfies*

$$(4.2) \qquad K(Tv, s) \leq M\, K(v, s), \quad v \in X_0 + X_1,\ s > 0,$$

with $M := \max(M_0, M_1)$.

Proof. The proof is almost a triviality. Let $s > 0$. For a given $\epsilon > 0$, let $v = v_0 + v_1$ be a decomposition of v that satisfies $\|v_0\|_{X_0} + s\|v_1\|_{X_1} \leq K(v,s) + \epsilon$. Then $Tv = (Tv - Tv_1) + Tv_1$ and $v - v_1 = v_0 \in X_0$. Hence,

$$\|Tv - Tv_1\|_{X_0} + s\|Tv_1\|_{X_1} \leq M_0\|v - v_1\|_{X_0} + sM_1\|v_1\|_{X_1} \leq M(K(f,s) + \epsilon).$$

Since $\epsilon > 0$ is arbitrary, (4.2) follows from the definition of the K-functional. □

We now apply this lemma to the solution operator $E := E(t)$ for the conservation law (1.1). The contractivity of the operator E on $L_1(\mathbb{R}^d)$ has a local variant. For this, we assume that the flux f is in Lip 1. It follows that the transport velocity vector $f'(u)$ satisfies

$$\Lambda := \sup_{u \in \mathbb{R}} \|f'(u)\|_{\ell_1} < \infty,$$

where $\|x\|_{\ell_1} := \sum_{i=1}^d |x_i|$. Under this assumption, we have (see [10]) for any locally integrable functions u_0 and v_0 and any ball $B(x, R)$ of radius $R > 0$ centered at the point x,

$$\int_{B(x,R)} |E(u_0) - E(v_0)| \leq \int_{B(x,R+\Lambda t)} |u_0 - v_0|. \tag{4.3}$$

In particular, if $u_0, v_0 \in L_1(\mathbb{R}^d) + L_\infty(\mathbb{R}^d)$ and $u_0 - v_0 \in L_1(\mathbb{R}^d)$, then taking a limit as $R \to \infty$ in (4.3) shows that E is L_1-Lipschitz on $L_1 + L_\infty$ with constant 1. Since E is a norm-one operator on L_∞, we can apply Lemma 4.1 and find

$$K(E(u_0), s; L_1(\mathbb{R}^d), L_\infty(\mathbb{R}^d)) \leq K(u_0, s; L_1(\mathbb{R}^d), L_\infty(\mathbb{R}^d)). \tag{4.4}$$

The K-functional for the pair $(L_1(\mathbb{R}^d), L_\infty(\mathbb{R}^d))$ can be described in terms of rearrangements of functions. We refer the reader to the book of Bennett and Sharpley [1] and Calderón's paper [2] for a discussion of rearrangements and the material that follows in this section. The rearrangement v^* of a function $v \in L_1(\mathbb{R}^d) + L_\infty(\mathbb{R}^d)$ is a nonincreasing function defined on $\mathbb{R}_+$ that is equimeasurable with v, i.e,

$$\operatorname{meas}\{\, x \in \mathbb{R}^d \mid |v(x)| > y \,\} = \operatorname{meas}\{\, x \in \mathbb{R}_+ \mid |v^*(x)| > y \,\}$$

for all $y \geq 0$. It follows that the rearrangement v^* of any $v \in L_p(\mathbb{R}^d)$ is in $L_p(\mathbb{R}_+)$ and

$$\|v\|_{L_p(\mathbb{R}^d)} = \|v^*\|_{L_p(\mathbb{R}_+)}, \quad 1 \leq p \leq \infty. \tag{4.5}$$

It can readily be verified that the K-functional for (L_1, L_∞) is

$$K(v, s; L_1(\mathbb{R}^d), L_\infty(\mathbb{R}^d)) = \int_0^s v^*(y)\, dy. \tag{4.6}$$

The Hardy–Littlewood maximal function v^{**} of v^* is related to the K-functional $K(v, s; L_1(\mathbb{R}^d), L_\infty(\mathbb{R}^d))$ in the following way:

$$v^{**}(s) := \frac{1}{s} \int_0^s v^*(y)\, dy = \frac{1}{s} K(v, s; L_1(\mathbb{R}^d), L_\infty(\mathbb{R}^d)).$$

Using the functions v^{**}, Calderón defines a normed space $X \subset L^1(\mathbb{R}^d) + L^\infty(\mathbb{R}^d)$ as *rearrangement invariant* if

$$f \in X \text{ and } g^{**}(s) \leq f^{**}(s),\ s \geq 0, \implies g \in X \text{ and } \|g\|_X \leq \|f\|_X.$$

Calderón showed that the interpolation spaces for the pair $(L_1(\mathbb{R}^d), L_\infty(\mathbb{R}^d))$ consist precisely of the set of rearrangement-invariant spaces.

As a consequence, from Lemma 4.1 we obtain the following.

THEOREM 4.2. *If the flux f is in* Lip 1, *then for any $t > 0$, the evolution operator $E = E(t)$ for the conservation law* (1.1) *when applied to an arbitrary function $u_0 \in L_1(\mathbb{R}^d) + L_\infty(\mathbb{R}^d)$ satisfies*

$$E(u_0)^{**}(s) \leq u_0^{**}(s), \quad s \geq 0. \tag{4.7}$$

In particular, E maps every rearrangement-invariant space X on $\mathbb{R}^d$ into itself with norm 1.

Inequality (4.7) gives precise information about the relative sizes of u_0 and Eu_0. For example, if we let $d = 1$, $f(u) = u^2$, u_0 equal the characteristic function of $[0, 1]$,

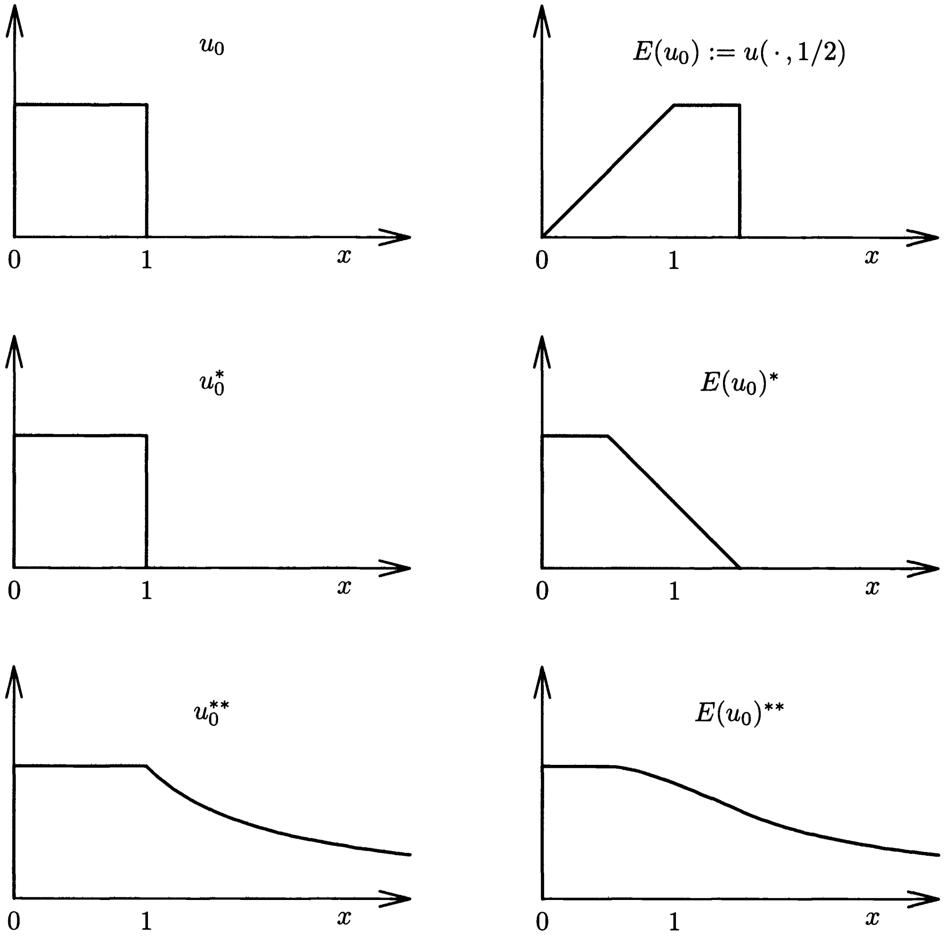


FIG. 2. *For these figures, the flux is $f(u) = u^2$. The initial data $u_0 = \chi_{[0,1]}$, its nonincreasing rearrangement u_0^*, and u_0^{**}, the maximal function of u_0^*. Similarly for $E(u_0) := u(\cdot, 1/2)$.*

and $t = 1/2$, then $u_0 = u_0^*$,

$$E(u_0)(x) = \begin{cases} 0, & x \le 0, \\ x, & 0 \le x \le 1, \\ 1, & 1 \le x \le 3/2, \\ 0, & 1 < x, \end{cases} \quad \text{and} \quad E(u_0)^*(x) = \begin{cases} 0, & x < 0, \\ 1, & 0 \le x \le 1/2, \\ 3/2 - x, & 1/2 \le x \le 3/2, \\ 0, & 3/2 \le x. \end{cases}$$

Note that for $1 < x < 3/2$, $u_0^*(x) < E(u_0)^*(x)$, yet, by Theorem 4.2 (or an easy direct calculation), we have that

$$E(u_0)^{**}(x) \le u_0^{**}(x) \tag{4.8}$$

for all $x > 0$. See Figure 2.

Some examples of rearrangement-invariant spaces are the $L_p(\mathbb{R}^d)$ spaces, $1 \le p \le \infty$, and the Lorentz spaces $L_{p,q}(\mathbb{R}^d)$, $1 \le p \le \infty$, $1 \le q \le \infty$; for $p > 1$, these consist of all $v \in L_1(\mathbb{R}^d) + L_\infty(\mathbb{R}^d)$ for which

$$\|v\|_{L_{p,q}(\mathbb{R}^d)} := \left(\int_0^\infty [s^{1/p}v^{**}(s)]^q \frac{ds}{s}\right)^{\frac{1}{q}}$$

is finite. Other rearrangement-invariant function spaces include $L \log L$ and Orlicz spaces.

We note that this analysis applies to any evolution equation that both satisfies a maximum principle and is a contraction on $L_1(\mathbb{R}^d)$, and, in fact, to any translation-invariant, integral-preserving, contraction semigroup on $L_1(\mathbb{R})$ (see, e.g., [13]).

5. Regularity in Besov spaces, part I: Positive results. The remainder of this paper concerns itself with the regularity of the solution $u(\,\cdot\,,t)$ of (1.1) as measured in Besov spaces. We continue to use the notation $\tau(\alpha,p) := (\alpha + 1/p)^{-1}$ and $B_s^\alpha := B_s^\alpha(L_s(\mathbb{R}^d))$.

It is well known that if the initial data u_0 is of bounded variation, then the solution $u(\,\cdot\,,t)$ of (1.1) is of bounded variation for all positive time t and

$$|u(\,\cdot\,,t)|_{\mathrm{BV}} \le |u_0|_{\mathrm{BV}} \tag{5.1}$$

for all $t > 0$. In addition, we showed in one space dimension [6], [14] that if $f \in C^{r+1}$ is globally Lipschitz continuous and uniformly convex, and $u_0 \in \mathrm{BV} \cap B^\alpha_{\tau(\alpha,1)}$ for some $1 < \alpha < r$, then for all $t > 0$, $u(\,\cdot\,,t) \in \mathrm{BV} \cap B^\alpha_{\tau(\alpha,1)}$ and

$$\|u(\,\cdot\,,t)\|_{B^\alpha_{\tau(\alpha,1)}} \le C(\|u_0\|_{B^\alpha_{\tau(\alpha,1)}} + 1), \tag{5.2}$$

where C depends only on r, t, $\|f^{(r+1)}\|_{L_\infty}$, and $|u_0|_{\mathrm{BV}}$; i.e., $\mathrm{BV} \cap B^\alpha_{\tau(\alpha,1)}$ is a regularity space for (1.1). (For the inviscid Burgers equation,

$$\|u(\,\cdot\,,t)\|_{B^\alpha_{\tau(\alpha,1)}} \le C\|u_0\|_{B^\alpha_{\tau(\alpha,1)}},$$

where C depends only on α.) One can ask whether any other spaces $B_q^\alpha(L_\sigma(\mathbb{R})) \cap \mathrm{BV}(\mathbb{R})$ are also regularity spaces for (1.1). As we have explained in the introduction, this cannot hold for Besov spaces that correspond to points above the line $\mathcal{L}_0$ or below the line $\mathcal{L}_1$ of Figure 1. Thus we can restrict our attention to spaces corresponding to points in the region bounded by $\mathcal{L}_0$, $\mathcal{L}_1$, and the x-axis. Each Besov space in this region is of the form $B_q^\alpha(L_{\tau(\alpha,p)})$ for some p between 1 and infinity, $0 < q \le \infty$, and $\alpha > 0$.

In this section, we show that if $u_0 \in \mathrm{BV} \cap B^\alpha_{\tau(\alpha,p)}$, $\alpha > 1$ and $1 \le p \le \infty$, then $u(\,\cdot\,,t)$ is in $\mathrm{BV} \cap B^\beta_{\tau(\beta,q)}$ for all $1 < q < \infty$ and $\beta < 1 + (\alpha - 1)/q$. The points $(1/\tau(\beta,q), \beta)$, $\beta = 1 + (\alpha - 1)/q$, lie on the line segment joining the points $(1,1)$ and $(1/\tau(\alpha,1), \alpha)$ in Figure 1; i.e., these spaces are intermediate to BV and $B^\alpha_{\tau(\alpha,1)}$. In the next two sections, we show that, in general, $u(\,\cdot\,,t)$ may not be in $\mathrm{BV} \cap B^\beta_{\tau(\beta,q)}$ for $1 < q < \infty$ whenever $\beta > 1 + (\alpha - 1)/q$. In particular, we can say that the $B^\alpha_{\tau(\alpha,p)}$, $\alpha > 1$, are not regularity spaces for (1.1) for any values of $p \ne 1$. As in [7], we can remove the restriction of BV functions if $f(u) = u^2$.

We first prove the following lemma, which holds for general functions v, not necessarily solution of (1.1).

LEMMA 5.1. *If $v \in \mathrm{BV} \cap B^\alpha_{\tau(\alpha,1)}$, $\alpha > 1$, then $v \in B^\beta_\sigma(L_{\tau(\beta,p)})$ for $\beta = 1 + (\alpha - 1)/p$, $1 < p < \infty$, and $\sigma = p\tau(\alpha,1)$.*

Proof. The proof will be based on a standard interpolation argument. We take as the BV-seminorm

$$|v|_{\mathrm{BV}} = \sup_{s>0} \frac{\omega_1(v,s)_1}{s}.$$

Since for any $r \geq 1$, $\omega_r(v,s)_1 \leq 2^{r-1}\omega_1(v,s)_1$, we have

$$\omega_r(v,s)_1 \leq C\omega_1(v,s)_1 \leq C\, s|v|_{\mathrm{BV}}$$

with C depending only on r.

We now assume that $1 < \alpha < r$, and we write $\mu := \tau(\alpha, 1)$ and $\lambda := \tau(\beta, p)$, i.e.,

$$\frac{1}{\mu} = \alpha + 1 \text{ and } \frac{1}{\lambda} = \beta + \frac{1}{p} = \frac{\alpha}{p} + 1.$$

Notice that

$$\Big(\frac{1}{\lambda}, \beta\Big) = \frac{1}{p}\Big(\frac{1}{\mu}, \alpha\Big) + \Big(1 - \frac{1}{p}\Big)(1,1),$$

that is, $(1/\lambda, \beta)$ is a convex combination of $(1/\mu, \alpha)$ and $(1,1)$. We shall estimate $\omega_r(v,s)_\lambda$. We use the abbreviated notation $\Delta(x) := |\Delta_h^r(v,x)|$. We let p' be the conjugate exponent to p, i.e., $1/p + 1/p' = 1$. It follows from Hölder's inequality that

$$\begin{aligned}\int_{\mathbb{R}} \Delta(x)^\lambda\, dx = \int_{\mathbb{R}} \Delta(x)^{\frac{\lambda}{p'}} \Delta(x)^{\frac{\lambda}{p}}\, dx &\leq \Big(\int_{\mathbb{R}} \Delta(x)\, dx\Big)^{\frac{\lambda}{p'}} \Big(\int_{\mathbb{R}} \Delta(x)^{\frac{\lambda}{p}\frac{1}{1-\lambda/p'}}\, dx\Big)^{1-\frac{\lambda}{p'}} \\ &\leq C|v|_{\mathrm{BV}}^{\frac{\lambda}{p'}} |h|^{\frac{\lambda}{p'}} \omega_r(v,|h|)_\mu^{\mu(1-\frac{\lambda}{p'})}\end{aligned}$$

since $\mu = \frac{\lambda}{p}\frac{1}{1-\lambda/p'}$.

By taking a supremum for $|h| \leq s$, we can replace the left side by $\omega_r(v,s)_\lambda^\lambda$ and raise both sides to the power $1/\lambda$ to obtain

$$\omega_r(v,s)_\lambda \leq C|v|_{\mathrm{BV}}^{\frac{1}{p'}} s^{\frac{1}{p'}} \omega_r(v,s)_\mu^{\frac{1}{p}}, \tag{5.3}$$

where we have used the fact that $(\frac{1}{\lambda} - \frac{1}{p'})\mu = \frac{1}{p}$. Therefore,

$$\begin{aligned}\int_0^\infty [s^{-\beta}\omega_r(v,s)_\lambda]^\sigma \frac{ds}{s} &\leq C|v|_{\mathrm{BV}}^{\frac{\sigma}{p'}} \int_0^\infty s^{-\beta\sigma + \frac{\sigma}{p'}} \omega_r(v,s)_\mu^{\frac{\sigma}{p}} \frac{ds}{s} \\ &= C|v|_{\mathrm{BV}}^{\frac{\sigma}{p'}} \int_0^\infty [s^{-\alpha}\omega_r(v,s)_\mu]^\mu \frac{ds}{s}\end{aligned}$$

since $\mu = \tau(\alpha,1) = \sigma/p$ and $\alpha\mu = \beta\sigma - \sigma/p'$. Raising both sides of this inequality to the power $1/\sigma$ shows that

$$|v|_{B_\sigma^\beta(L_{\tau(\beta,p)}(\mathbb{R}))} \leq C|v|_{\mathrm{BV}}^{\frac{1}{p'}} |v|_{B_{\tau(\alpha,1)}^\alpha}^{\frac{1}{p}}. \qquad \square$$

In the next theorem, we apply Lemma 5.1 to solutions of (1.1).

THEOREM 5.2. *Let the flux f be strictly convex and in C^{r+1}. If u_0 is a function of compact support in $B_{\tau(\alpha,p)}^\alpha \cap \mathrm{BV}(\mathbb{R})$ for some $\alpha > 1$ and $1 < p \leq \infty$, then for any later time t the solution $u(\cdot, t)$ to (1.1) is in every Besov space $B_{\tau(\beta,q)}^\beta \cap \mathrm{BV}(\mathbb{R})$ for all $1 \leq q < \infty$ and $0 < \beta < 1 + (\alpha - 1)/q$.*

Proof. It may help the reader to refer to Figure 3 during the course of this proof. We first assume that $1 < q < \infty$ and $1 \leq \beta < 1 + (\alpha - 1)/q$, and we choose auxiliary

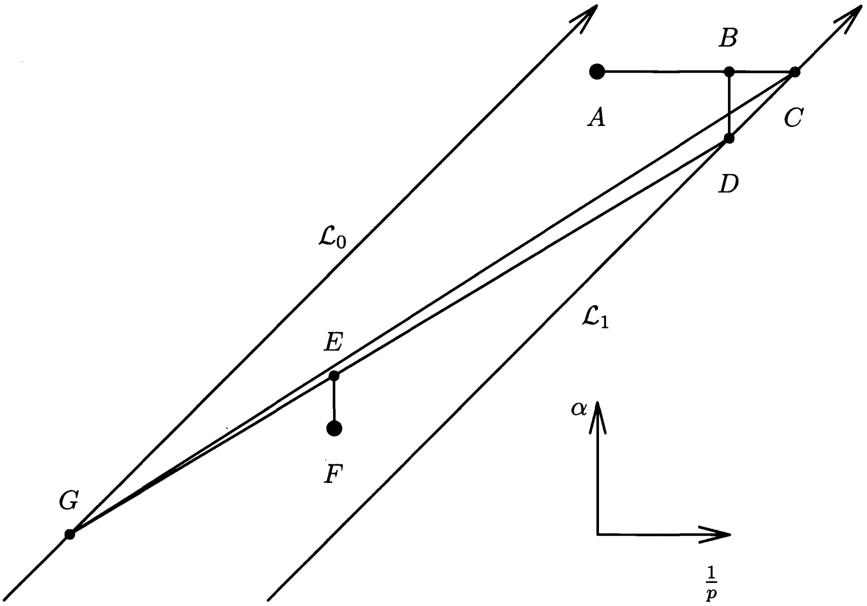


FIG. 3. *The parameters and spaces of Theorem* 5.2. *The points and the spaces are* A: $(\frac{1}{\tau(\alpha,p)},\alpha)$, $B^{\alpha}_{\tau(\alpha,p)}(L_{\tau(\alpha,p)}(\mathbb{R}))$; B: $(\frac{1}{\tau(\alpha',1)},\alpha)$, $B^{\alpha}_{\tau(\alpha,p)}(L_{\tau(\alpha',1)}(\mathbb{R}))$; C: $(\frac{1}{\tau(\alpha,1)},\alpha)$; D: $(\frac{1}{\tau(\alpha',1)},\alpha')$, $B^{\alpha'}_{\tau(\alpha',1)}(L_{\tau(\alpha',1)}(\mathbb{R}))$; E: $(\frac{1}{\tau(\beta',r)},\beta') = (\frac{1}{\tau(\beta,q)},\beta')$, $B^{\beta'}_{r\tau(\alpha',1)}(L_{\tau(\beta',r)}(\mathbb{R}))$; F: $(\frac{1}{\tau(\beta,q)},\beta)$, $B^{\beta}_{\tau(\beta,q)}(L_{\tau(\beta,q)}(\mathbb{R}))$; *and* G: $(1,1)$, $\mathrm{BV}(\mathbb{R})$). *The lines* $\mathcal{L}_0$ *and* $\mathcal{L}_1$ *are as in Figure* 1.

parameters α', β', and s that satisfy $1 < \alpha' < \alpha$, $\beta < \beta' < 1 + (\alpha' - 1)/q$, $\tau(\alpha', 1) < \tau(\alpha, p)$ and

$$\frac{1}{\tau(\beta,q)} = \beta' + \frac{1}{s},$$

i.e., $\tau(\beta, q) = \tau(\beta', s)$. Note that it is always possible to choose these parameters by first choosing α' close enough to α. It follows that $1 < s < \infty$.

We note that because u_0 has compact support, $u_0 \in B^{\alpha}_{\tau(\alpha,p)}(L_{\tau(\alpha,p)}(\mathbb{R}))$ (point A) implies that $u_0 \in B^{\alpha}_{\tau(\alpha,p)}(L_{\tau(\alpha',1)}(\mathbb{R}))$ (point B), since $\tau(\alpha', 1) < \tau(\alpha, p)$. Now, by the embedding theorems mentioned in §3, u_0 is in $B^{\alpha'}_{\tau(\alpha',1)}(L_{\tau(\alpha',1)}(\mathbb{R}))$ (point D), since $\alpha' < \alpha$.

Inequalities (5.1) and (5.2) and the fact that $u_0 \in B^{\alpha'}_{\tau(\alpha',1)}(L_{\tau(\alpha',1)}(\mathbb{R})) \cap \mathrm{BV}(\mathbb{R})$ imply that $u(\cdot\,,t)$ is in the same space for all $t > 0$. We can apply Lemma 5.1 to see that $u(\cdot\,,t)$ is in $B^{\beta'}_{s\tau(\alpha',1)}(L_{\tau(\beta',s)}(\mathbb{R})) = B^{\beta'}_{s\tau(\alpha',1)}(L_{\tau(\beta,q)}(\mathbb{R}))$ (point E). Finally, because $\beta < \beta'$, a standard embedding theorem implies that $u(\cdot\,,t)$ is in $B^{\beta}_{\tau(\beta,q)}(L_{\tau(\beta,q)}(\mathbb{R}))$ (point F), as required.

When $q = 1$ or $\beta < 1$, the theorem follows from what we have already shown and the fact that $B^{\tilde{\beta}}_{\tau(\tilde{\beta},q)}$ is embedded in $B^{\beta}_{\tau(\beta,q)}$ for any $\beta < \tilde{\beta}$. □

6. Regularity in Besov spaces, part II: Limits on regularity. Recall that a smoothness space X is a regularity space for (1.1) if $u_0 \in X \implies u(\cdot\,,t) \in X$ for all $t > 0$. We have remarked that the spaces $\mathrm{BV}(\mathbb{R}) \cap B^{\alpha}_{\tau(\alpha,1)}$ are regularity spaces for (1.1), and we gave simple arguments in the introduction to show that spaces on or above the line $\mathcal{L}_0$ with $\alpha > 1$ or below the line $\mathcal{L}_1$ cannot be regularity spaces for X. We went on to show that if u_0 is in any space $X = B^{\alpha}_{\tau(\alpha,p)} \cap \mathrm{BV}(\mathbb{R})$ for $\alpha > 1$ and $1 < p < \infty$ (i.e., between the lines $\mathcal{L}_0$ and $\mathcal{L}_1$), then the solutions $u(\,\cdot\,,t)$ remained in $B^{\beta}_{\tau(\beta,q)}$ for $1 \le q < \infty$ and $\beta < 1+(\alpha-1)/q$ for all time; we now show that the solution of (1.1) is, in general, not in any space $B^{\beta}_{\tau(\beta,q)}$ with $1 \le q < \infty$ and $\beta > 1+(\alpha-1)/q$. In particular, no Besov space B^{α}_{σ} for $\alpha > 1$ and any $\sigma \ne \tau(\alpha,1)$ is a regularity space for the one-dimensional conservation law (1.1). In this section, we analyze the inviscid Burgers equation, and in the next section we treat general strictly convex fluxes f. In §8, we generalize these results to multivariate problems.

So we consider in this section only $f(u) = u^2/2$, $\alpha > 1$, and $\tau = \tau(\alpha,p) := (\alpha+1/p)^{-1}$ for $1 < p < \infty$. For any α and p, we construct initial data $u_0 \in \mathrm{BV} \cap B^{\alpha}_{\tau(\alpha,p)}$ such that the solution $u(x,1)$ of (1.1) at time 1 is not in any space $B^{\beta}_{\tau(\beta,q)}$ for $1 \le q < \infty$ and $\beta > 1+(\alpha-1)/q$.

It is perhaps easier to first describe the solution $u(x,1)$ that we want to achieve at time 1 and then explain how to find suitable initial data u_0 that yields it. We construct a set of functions ϕ_k, $k = 1,2,\dots$, each of compact support on intervals I_k such that $\sum_k |I_k|$ is finite; our solution will be $u(\,\cdot\,,1) = u_1 := \sum_k \phi_k(\,\cdot - x_k)$, where the increasing sequence of points x_k is chosen such that $u(x,1)$ has bounded support and the supports of $\phi_k(\,\cdot - x_k)$ and $\phi_j(\,\cdot - x_j)$ don't overlap if $j \ne k$. The graph of $\phi_k(x - x_k)$ is given in Figure 4. The rightmost portion of its graph has N_k steps with height $H_k := 2^{-k}$ and width $W_k := 2^{-\alpha k}$; on the left, a linear piece with slope $1/2$ connects the top of the steps with the x axis. Precisely,

$$\phi_k(x) := \begin{cases} 0, & x \le -2N_kH_k, \\ \frac{1}{2}x + N_kH_k, & -2N_kH_k \le x \le 0, \\ (N_k - n)H_k, & nW_k \le x < (n+1)W_k,\ 0 \le n < N_k, \\ 0, & N_kW_k \le x. \end{cases}$$

We choose N_k as the greatest integer such that $N_kH_k \le k^{-r}$. If N_k is zero, then ϕ_k is defined to be zero. If $N_k > 0$ then

$$\frac{1}{2}k^{-r} \le N_kH_k \le k^{-r}. \tag{6.1}$$

The integer r will be given later. We have

$$|I_k| = 2N_kH_k + N_kW_k \le 3N_kH_k \le 3k^{-r}, \tag{6.2}$$

so $\sum_k |I_k| < \infty$ as claimed, and a set of points $\{x_k\}_{k=1}^{\infty}$ can easily be chosen with the required properties.

We obtain $u_0(x)$ by solving the associated backward problem

$$\begin{aligned} &v_t - \left(\frac{1}{2}v^2\right)_x = 0, && x \in \mathbb{R}^d,\ t > 0, \\ &v(x,0) = v_0(x) := u(x,1), && x \in \mathbb{R}^d. \end{aligned}$$

Then we take $u_0 := v(\,\cdot\,,1)$. Each jump in v_0 smooths into a linear rarefaction wave with slope -1, and the linear piece on the left of each ϕ_k evolves into a steeper profile with slope 1, but not yet a shock, at time 1. Thus, $u_0(x) := v(x,1) = \sum_k \psi_k(x - x_k)$,

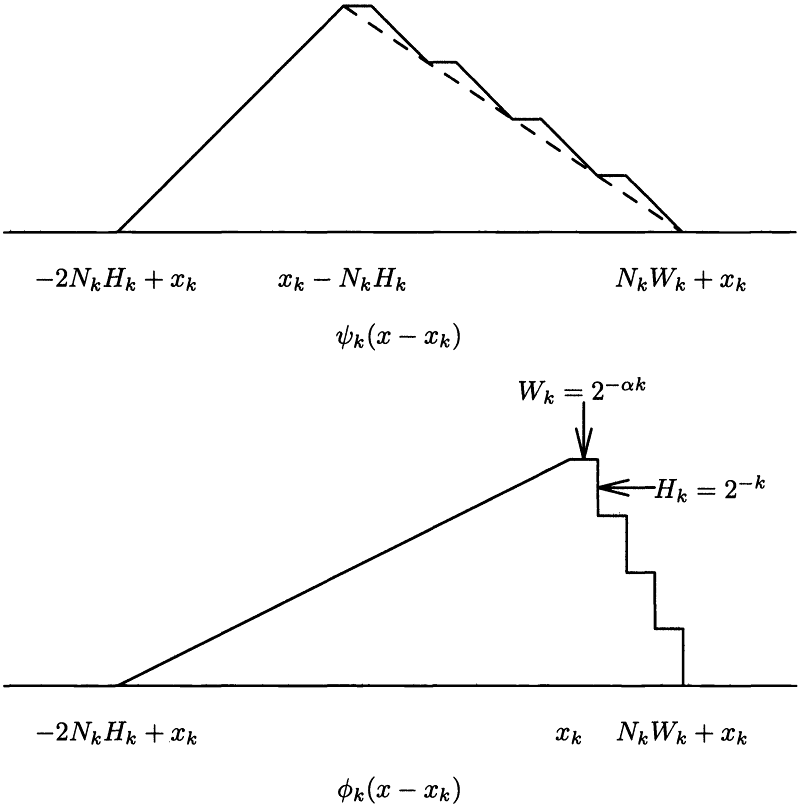


FIG. 4. *The functions $\phi_k(x - x_k)$ and $\psi_k(x - x_k)$, from which $u_0(x) = \sum_k \psi_k(x - x_k)$ and $u(x,1) = \sum_k \phi_k(x - x_k)$. The dashed line indicates the linear approximation to $\psi_k(x - x_k)$ in Theorem* 6.1.

where the continuous function ψ_k takes the values

$$\begin{cases} 0, & x = -2N_kH_k, \\ (N_k - n)H_k, & x = nW_k - (N_k - n)H_k \text{ and} \\ & x = (n+1)W_k - (N_k - n)H_k, \quad 0 \le n < N_k, \\ 0, & x = N_kW_k, \end{cases}$$

is linear between these values, and is zero outside the interval $[-2N_kH_k, N_kW_k]$. Thus, on the right, ψ_k consists of linear pieces with slopes alternating between 0 and -1. See Figure 4. Note that ψ_k has $2N_k + 1$ linear pieces and ϕ_k has $N_k + 1$ linear pieces.

It is easy to justify that u_1 is the solution to (1.1) when $t = 1$ with initial data u_0. For example, in the description of the solution given by Lax (see our §2), for $t < 1$ there is a unique solution $y(x,t)$ to (2.3). When $t = 1$ there is a unique solution to (2.3) except at breakpoints of u_1. These correspond to the jumps in u_1.

We fix $\alpha > 1$ and $1 < p < \infty$ and let r be the smallest integer that satisfies $(r-1)\tau(\alpha,p)/p > 1$, or, equivalently, $r > \alpha p + 2$.

THEOREM 6.1. *Let $1 < p < \infty$ and $\alpha > 1$ be fixed. The function u_0 defined above is in $B^{\alpha}_{\tau(\alpha,p)}$ for $\tau(\alpha,p) := (\alpha + 1/p)^{-1}$, while the function $u_1 = u(\,\cdot\,,1)$ with u the solution to* (1.1) *for this u_0 and $f(u) := u^2/2$ is not in any space $B^{\beta}_{\tau(\beta,q)}$ for any $1 < q < \infty$ and any $\beta > 1 + (\alpha - 1)/q$.*

Proof. (i) We first show that $u_0 \in B^\alpha_{\tau(\alpha,p)}$. For this purpose, we use the seminorm (3.5) for $B^\alpha_{\tau(\alpha,p)}$ that is defined using the approximation errors $\sigma_{2^j}(u_0)_p$. We fix a value of $n \geq 1$ such that $2^n \geq n^r$ and we bound $\sigma_m(u_0)$ for $m := 2^{n+3}$ by constructing a piecewise-linear approximant S to u_0 as follows.

Recall that

$$\operatorname{supp}(\phi_k) = \operatorname{supp}(\psi_k) = I_k = [-2N_kH_k, N_kW_k].$$

Outside of $\bigcup_k(x_k + I_k)$, we define S to be zero. Let $z_k := x_k + N_kW_k$ be the right endpoint of the support interval $x_k + I_k$ of $\psi_k(x - x_k)$. On each $x_k + I_k$ for $1 \leq k < n$ (i.e., where u_0 is largest), we define $S(x) := u_0(x) = \psi_k(x - x_k)$. Then S has at most $\sum_{j=1}^{n-1}(2N_j + 2) \leq 2^{n+1} + 2n \leq 2^{n+2}$ linear pieces to the left of the point z_{n-1}. To the right of z_{n-1}, we define S as follows. On any interval $x_k + I_k$, with $n \leq k \leq 2^n$ (where u_0 is of moderate size), we define S to be the continuous, piecewise-linear function that passes through the points $(x_k - 2N_kH_k, 0)$, $(x_k - N_kH_k, N_kH_k)$, and $(x_k + N_kW_k, 0)$ (the dashed line in Figure 4). To the right of z_{2^n} (where u_0 is smallest), we define S to be identically zero; we call this semi-infinite interval I_∞. Then S has at most $3 \cdot 2^n$ breakpoints to the right of z_{n-1} and hence at most $m = 2^{n+3}$ breakpoints in all.

We consider next the error $E(x) := |u_0(x) - S(x)|$ at points where E is not identically zero. On any interval $x_k + I_k$, $k = n, \dots, 2^n$, the error is no greater than W_k (since the slope of the dashed line in Figure 4 is greater than -1) and E is nonzero on a set of measure at most $N_k(W_k + H_k) \leq 2N_kH_k \leq 2k^{-r}$. Hence

$$\int_{x_k+I_k} E(x)^p\,dx \leq 2W_k^p k^{-r} \leq 2 \cdot 2^{-k\alpha p}k^{-r}. \tag{6.3}$$

On the other hand, on I_∞, $E(x) \leq N_{2^n}H_{2^n} \leq 2^{-nr}$ and E is nonzero on a set of measure not exceeding

$$\sum_{k=2^n}^\infty |I_k| \leq 3\sum_{k=2^n}^\infty \frac{1}{k^r} \leq C2^{-n(r-1)},$$

(see (6.2)) with C (here and later in this proof) depending only on r, since $r > 1$. This gives

$$\int_{I_\infty} E(x)^p\,dx \leq C2^{-nrp}2^{-n(r-1)}. \tag{6.4}$$

Adding the estimates (6.3) and (6.4), we obtain for $m = 2^{n+3}$ and n sufficiently large,

$$\begin{aligned}\sigma_m(u_0)_p^p \leq \|u_0 - S\|_p^p &\leq C2^{-nrp}2^{-n(r-1)} + 2\sum_{k=n}^{2^n} k^{-r}2^{-k\alpha p}\\ &\leq 2^{-nrp}2^{-n(r-1)} + C2^{-n\alpha p}n^{-r}\\ &\leq C2^{-n\alpha p}n^{-r}\end{aligned}$$

since $r \geq \alpha$ and $r > 1$. This inequality also holds (trivially) for all n by simply adjusting the constant C. Using this and the monotonicity of $\sigma_j(u_0)_p$, we obtain that

$$\sum_{n=1}^\infty [2^{n\alpha}\sigma_{2^n}(u_0)_p]^{\tau(\alpha,p)} \leq C\sum_{n=1}^\infty n^{-r\tau(\alpha,p)/p} < \infty$$

since $r\tau(\alpha,p)/p > 1$ by our definition of r. This shows that u_0 is in $B^{\alpha}_{\tau(\alpha,p)}$ and completes the proof of (i).

(ii) We show next that $u_1 = u(\cdot,1) \notin B^{\beta}_{\tau(\beta,q)}$ for any $1 \leq q < \infty$ and $\beta > 1 + (\alpha - 1)/q$ by giving a lower bound on $\omega_{r'}(u_1,t)_{\tau(\beta,q)}$ for any fixed $r' > \beta$. We consider any k for which ϕ_k is not identically zero, and we examine $\Delta_h^{r'}(\phi_k,x)$ for $h := h_k := W_k/r'$ and $\nu W_k - h < x < \nu W_k$, $\nu = 1,\dots,N_k$, i.e., just to the left of each jump of height H_k in ϕ_k. Since $\Delta_h^{r'}(g,x) = 0$ for any constant function g, we have for $\nu W_k - h < x < \nu W_k$,

$$\Delta_h^{r'}(\phi_k,x) = \Delta_h^{r'}(\phi_k - (N_k - \nu)H_k, x) = (-1)^{r'}H_k = (-1)^{r'}2^{-k} \tag{6.5}$$

since all values of $\phi_k(x+jh) - (N_k - \nu)H_k = 0$ in (3.1) except for $j = 0$. This holds for x on a set of measure $N_k h = N_k W_k/r' \geq C2^k k^{-r}2^{-\alpha k} = Ck^{-r}2^{-(\alpha-1)k}$.

From (6.5), we derive for $\tau := \tau(\beta,q) := (\beta + 1/q)^{-1}$,

$$\omega_{r'}(u_1,h_k)_\tau^\tau \geq |H_k|^\tau k^{-r}2^{-(\alpha-1)k} = C2^{k(1-\tau-\alpha)}k^{-r}, \tag{6.6}$$

for all k sufficiently large, say $k \geq k_0$. Using the monotonicity of $\omega_{r'}(u_1,t)_\tau$ in t, dividing the interval of integration in (3.2) into intervals $[h_{k+1},h_k)$ and discretizing the integral yields

$$|u_1|^\tau_{B^\beta_\tau} \geq C\sum_{k=k_0}^{\infty} h_k^{-\beta\tau}\omega_{r'}(u_1,h_k)_\tau^\tau \geq C\sum_{k=k_0}^{\infty} k^{-r}2^{k(\beta\tau\alpha+1-\tau-\alpha)} = \infty \tag{6.7}$$

because

$$\beta\tau\alpha + 1 - \tau - \alpha = \beta\frac{q}{\beta q+1}\alpha + 1 - \frac{q}{\beta q+1} - \alpha = \frac{\beta q + 1 - q - \alpha}{\beta q + 1} > 0.$$

Hence $u_1 \notin B^\beta_{\tau(\beta,q)}$. □

The previous theorem can be used together with embedding theorems for Besov spaces to show that none of the Besov spaces $B^\alpha_s(L_\tau)$ with $(\alpha+1)^{-1} < \tau < \alpha^{-1}$ are regularity spaces. That is, we can allow any value of s. A modification of the construction of the theorem allows this conclusion for $\tau = \alpha^{-1}$. We already remarked that no such space with $\tau > 1/\alpha$ or $\tau < 1/(\alpha+1)$ is a regularity space. We leave these details to the reader.

7. More general fluxes. We shall next show that the results of the previous section are valid for more general fluxes f. We shall assume in this section that f is a strictly convex function on $\mathbb{R}$. Then $a(u) := f'(u)$ is strictly increasing and has an inverse $b := a^{-1}$ under composition of functions. We shall assume that $a(0) = 0$ and therefore $b(0) = 0$ (this assumption could be removed with a suitable change in the construction below).

If $\alpha > 1$, let r, u_0, and u_1 be defined as in the previous section. We consider the solution $v(x,t)$ to (1.1) for the flux f and the initial condition

$$v_0(x) := b(u_0(x)), \quad x \in \mathbb{R}. \tag{7.1}$$

The same argument we have given in §6 can be applied here to show that the solution $v(x,t)$ of (1.1) for $t \leq 1$ with data v_0 is the same as $b(u(x,t))$, where $u(x,t)$ is defined in the previous section.

The next two theorems will show that for every $1 < p < \infty$ and $\alpha > 1$, the function v_0 is in $B^\alpha_{\tau(\alpha,p)}$, but $v(\cdot,1) = v_1$ is not in any space $B^\beta_{\tau(\beta,q)}$ for any $1 < q < \infty$ and

$\beta > 1 + (\alpha - 1)/q$. We shall assume in these theorems that the flux f is in C^{r+1}; it follows that a and b are in C^r.

THEOREM 7.1. *Let α, r, u_0, and $\tau := \tau(\alpha, p)$ be defined as above, and let $1 < p < \infty$. If b is in $C^r[0,1]$, then v_0 is in $B^{\alpha}_{\tau(\alpha,p)}$.*

Proof. This proof is similar to part (i) of Theorem 6.1.

It is reasonable to expect that v_0 will be as smooth as u_0 since b is smooth, but the proof is not completely trivial. We shall estimate the error in approximating v_0 by piecewise polynomials of order r with m pieces. That is, we shall estimate the error $\sigma_m(v_0)_p$ in approximating v_0 by the elements of $\Sigma_{m,r}$ in the $L_p(\mathbb{R})$ norm. We first note that since $b(0) = 0$ and $0 \le u_0(x) \le 1$, for all x, we have $0 \le v_0(x) \le \|b'\|_{L_\infty[0,1]} u_0(x)$ for all $x \in \mathbb{R}$. This implies that $v_0 \in L_p(\mathbb{R})$.

We fix an integer $n \ge 1$ and we estimate $\sigma_m(v_0)_p$ for $m \ge C_0 2^n$ with C_0 an absolute constant that is specified in the course of the proof. We let $\Psi_1 := \sum_{k=1}^{n-1} \psi_k(\,\cdot - x_k)$, $\Psi_2 := \sum_{k=n}^{2^n} \psi_k(\,\cdot - x_k)$, and $\Psi_3 := \sum_{k>2^n} \psi_k(\,\cdot - x_k)$. These functions have disjoint supports. From the definition of v_0, we have

$$v_0 = b(\Psi_1) + b(\Psi_2) + b(\Psi_3) =: b_1 + b_2 + b_3.$$

We first estimate $\sigma_m(b_1)_p$. Recall that Ψ_1 is a piecewise-linear function on $\mathbb{R}$ with no more than $C2^n$ breakpoints. We shall now show that one can add at most an additional $C_0 2^n$ breakpoints so that for any interval I in the resulting partition, $\Psi_1(I)$ is contained in an interval of length 2^{-n}. Indeed, the variation of $\psi_k \le 2/k^r$ for $k = 1, \dots, n$; hence, we need only insert at most $2k^{-r}2^n + 2$ new breakpoints for each $k = 1, \dots, n-1$ to obtain the desired partition. For each of these intervals I, we let P_I be the Taylor polynomial to b of order r expanded at the center of $\Psi_1(I)$. Then $P_I(\Psi_1)$ is a polynomial of order r on I. We define the piecewise-polynomial function S_1 by $S_1 := P_I(\Psi_1)$ for each I. Then,

$$\|b_1 - S_1\|_{L_\infty(I)} = \|b(\Psi_1) - P_I(\Psi_1)\|_{L_\infty(I)} \le \|b^{(r)}\|_{L_\infty[0,1]} 2^{-nr}.$$

Since b_1 and S_1 have compact support, it follows that

$$\sigma_m(b_1)_p \le C2^{-nr}, \quad m \ge C_0 2^n \tag{7.2}$$

with C not depending on n.

We can estimate $\sigma_m(b_2)$ in a similar way. We have shown in the proof of Theorem 6.1 that there is a piecewise-linear function $\tilde{\Psi}_2$ with at most $3 \cdot 2^n$ pieces that satisfies

$$\|\Psi_2 - \tilde{\Psi}_2\|_{L_\infty(\mathbb{R})} \le 2^{-n\alpha}.$$

Moreover, $\tilde{\Psi}_2 = \sum_{k=n}^{2^n} \tilde{\psi}_k(\,\cdot - x_k)$ with each $\tilde{\psi}_k(\,\cdot - x_k)$ a piecewise linear function with 4 pieces and $\mathrm{Var}(\tilde{\psi}_k) = \mathrm{Var}(\psi_k) \le 2k^{-r}$. Therefore, as in the previous case of Ψ_1, we can add new breakpoints and obtain a partition of $\mathbb{R}$ into at most $C_0 2^n$ intervals I such that $\tilde{\Psi}_2$ is linear on I and $\tilde{\Psi}_2(I)$ is contained in an interval of length $\le 2^{-n}$. If P_I denotes the Taylor polynomial of order r of b expanded about the center of $\tilde{\Psi}_2(I)$, then $P_I(\tilde{\Psi}_2)$ is a polynomial of order r on I. The piecewise-polynomial function S_2 is defined to be $P_I(\tilde{\Psi}_2)$ on each I. Then, for each of the intervals I, we have

$$\begin{aligned} \|b_2 - S_2\|_{L_\infty(I)} &\le \|b(\Psi_2) - b(\tilde{\Psi}_2)\|_{L_\infty(I)} + \|b(\tilde{\Psi}_2) - P_I(\tilde{\Psi}_2)\|_{L_\infty(I)} \\ &\le \|b'\|_{L_\infty[0,1]} 2^{-n\alpha} + \|b^{(r)}\|_{L_\infty[0,1]} 2^{-nr} \le C2^{-n\alpha} \end{aligned} \tag{7.3}$$

with C independent of n. Now, by (6.2), b_2 and S_2 vanish outside of a set of measure at most Cn^{-r+1} with C depending only on r. Hence, from (7.3),

$$\|b_2 - S_2\|_{L_p(\mathbb{R})} \leq Cn^{(-r+1)/p}2^{-n\alpha}.$$

It follows that

$$\sigma_m(b_2)_p \leq Cn^{(-r+1)/p}2^{-n\alpha}, \quad m \geq C_0 2^n. \tag{7.4}$$

Since $b(0) = 0$ and $\|\Psi_3\|_{L_\infty(\mathbb{R})} = \sup_{k>2^n} \|\psi_k\|_{L_\infty(\mathbb{R})} \leq 2^{-nr}$, we have

$$\|b_3\|_{L_\infty(\mathbb{R})} \leq \|b'\|_{L_\infty[0,1]}\|\Psi_3\|_{L_\infty(\mathbb{R})} \leq \|b'\|_{L_\infty[0,1]}2^{-nr}.$$

Because b_3 has compact support, we have

$$\sigma_m(b_3)_p \leq C2^{-nr}, \quad m \geq 1. \tag{7.5}$$

Now, $v_0 = b_1 + b_2 + b_3$, and therefore the estimates (7.2), (7.4), and (7.5) give

$$\sigma_m(v_0)_p \leq Cn^{(-r+1)/p}2^{-n\alpha}, \quad m \geq C_0 2^n.$$

From our assumption on r, we have $\tau(r-1)/p > 1$ and therefore

$$\sum_{n=1}^{\infty}[2^{n\alpha}\sigma_{2^n}(v_0)_p]^\tau < \infty.$$

From the characterization (3.5), we obtain that $v_0 \in B^{\alpha}_{\tau(\alpha,p)}$. □

We shall next show that $v_1 := v(\,\cdot\,,1)$ is not in any $B^{\beta}_{\tau(\beta,q)}$, for $1 \leq q < \infty$ and $\beta > (\alpha-1)/q + 1$. For this, we shall assume that

$$b'(x) \geq c, \quad x \in (0,1) \tag{7.6}$$

for some $c > 0$.

THEOREM 7.2. *Under the assumptions of Theorem* 7.1 *and the added assumption* (7.6), *we have* $v_1 \notin B^{\beta}_{\tau(\beta,q)}$ *for all* $1 \leq q < \infty$ *and* $\beta > (\alpha-1)/q + 1$.

Proof. The widths of the constant states in $b(\phi_k)$ are the same as for ϕ_k, and because of assumption (7.6), the heights of the jumps are $\geq c2^{-k}$. Therefore, the same argument as given in the proof of part (ii) of Theorem 6.1 shows that (6.7) holds with $v_1 = b(u_1)$ in place of u_1. Hence v_1 is not in $B^{\beta}_{\tau(\beta,q)}$. □

In summary, Theorems 7.1 and 7.2 give the following results.

THEOREM 7.3. *Given that* $\alpha > 1$ *and* $1 < p \leq \infty$ *and that the flux* f *to the univariate conservation law* (1.1) *has derivative* $a(u) = f'(u)$, *which is strictly increasing and whose inverse function* b *is in* C^r *with* $(r-1)\tau(\alpha,p)/p > 1$, *and also satisfies* (7.6), *the initial condition* $v_0 = b(u_0)$ *is in* $B^{\alpha}_{\tau(\alpha,p)}$ *but the solution* $v(\,\cdot\,,1)$ *to* (1.1) *at time* $t = 1$ *for this initial condition is not in any* $B^{\beta}_{\tau(\beta,q)}$, *for* $1 \leq q < \infty$ *and* $\beta > (\alpha-1)/q + 1$. *Consequently, none of the spaces* $B^{\alpha}_{\tau}(L_\tau)$, $\tau \neq 1/(\alpha+1)$ *are regularity spaces for* (1.1).

8. Regularity in several space dimensions.

We shall next consider the regularity of the solution to the conservation law (1.1) in several space dimensions. The proof that the spaces B^{α}_{τ}, $\tau = (\alpha+1)^{-1}$, are regularity spaces for conservation laws in one space dimension rests on the fact that they arise in the characterization of approximation classes for methods of nonlinear approximation in $L_1(\mathbb{R})$ such as wavelets and free-knot splines. The Besov spaces $B^{\alpha}_{\tau}(L_\tau(\mathbb{R}^d))$, $\tau = (\alpha/d+1)^{-1}$, play the analogous role in nonlinear approximation in several space dimensions. For example, they arise in the characterization of nonlinear approximation by wavelet sums (see [4]). One might expect therefore that they are regularity spaces for conservation laws in several space dimensions. We shall show that this is not the case when $\alpha > 1$.

We assume that $a(u) := f'(u)$ is a continuously differentiable mapping from $\mathbb{R} \to \mathbb{R}^d$ and

$$\text{(i)} \quad a(0) = 0,$$
$$\text{(ii)} \quad a'(0) \neq 0.$$

A slight change in the argument given below would allow the point 0 to be replaced by any other point $x \in \mathbb{R}$.

We write $a(u) =: (a_1(u), \dots, a_d(u))$. Without loss of generality, we can assume that $a_1'(0) \neq 0$ and $a_1'(u) > 0$, in a half-neighborhood $[0, \eta]$ of 0. In order to utilize our previous notation, we shall assume that $a_1(\eta) = 1$. However, a simple modification of the arguments given below would treat the general case of η. We denote by b_1 the inverse function (under composition of functions) to a_1 on $[0, \eta]$. Then, b_1 is defined on $[0, 1]$ and satisfies condition (7.6).

THEOREM 8.1. *Let $\alpha > 1$ and $0 < \tau \leq \infty$. If $b_1 \in C^r[0,1]$ for some sufficiently large integer $r > \max(\alpha, d)$ (described in part (ii) of the proof below), then the space $B^\alpha_\tau(L_\tau(\mathbb{R}^d))$ is not a regularity space for the conservation law (1.1).*

Proof. We shall consider the following three cases.

(i) $\alpha < d(1/\tau - 1)_+$.

In this case, the space $B^\alpha_\tau(L_\tau(\mathbb{R}^d))$ contains functions that are not locally integrable and hence this space cannot be a regularity space for (1.1).

(ii) $\alpha > 1$, $\alpha \geq d(1/\tau - 1)_+$, and $\alpha < 1/\tau$.

We use our previous univariate notation $\tau(\alpha, p) := (\alpha + 1/p)^{-1}$. In this case, we can write $\tau = \tau(\alpha, p)$ for some p with $1 < p < \infty$. We shall show that there is an initial condition w_0 of compact support that is in $B^\alpha_\tau(L_\tau(\mathbb{R}^d))$ but the solution $w(\cdot, 1) = E(1)w_0$ to (1.1) is not in $B^\alpha_\tau(L_\tau(\mathbb{R}^d))$.

We shall utilize the univariate construction of §7 with some modifications. For our fixed values of α and p, we assume that r is chosen as in §§6 and 7. Then, the construction of §6 applies and we let u_0 be the univariate function given in that section. Further, we let $v(\cdot, t)$ be the solution given in §7 to the univariate conservation law (1.1) with initial condition $v_0 := b_1(u_0)$ and transport velocity a_1. We recall that we have shown in §7 that $v_0 \in B^\alpha_\tau(L_\tau(\mathbb{R}))$ but $v_1 := v(\cdot, 1)$ is not in any of the spaces $B^\beta_{\tau(\beta,q)}$ for $1 \leq q < \infty$ and $\beta > 1 + (\alpha - 1)/q$. In particular, v_1 is not $B^\alpha_\tau(L_\tau(\mathbb{R}))$.

We consider now the multivariate conservation law (1.1) with the initial condition

$$w_0(x) := v_0(x_1)\varphi(x_2, \dots, x_d), \quad x = (x_1, \dots, x_d) \in \mathbb{R}^d, \tag{8.1}$$

with $\varphi(x_2, \dots, x_d) = \phi(x_2) \cdots \phi(x_d)$ and ϕ a compactly supported $C^\infty(\mathbb{R})$ function that is one on a sufficiently large (to be chosen momentarily) interval I centered at 0 and satisfying $\|\phi\|_{L_\infty(\mathbb{R})} = 1$. We denote by Q the cube I^d. We let $w = w(x, t)$ denote the solution to (1.1) at time t with initial condition w_0.

Let $\ell \geq 1$ be such that v_0 vanishes outside of $[-\ell, \ell]$. We claim that if the sidelength of Q is chosen sufficiently large, then

$$w(x, t) = v(x_1, t), \quad \text{a.e. } x \in [-2\ell, 2\ell]^d, \ 0 \leq t \leq 1. \tag{8.2}$$

We now prove this claim. Since v_0 (and hence w_0) has compact range, for any $x \in \mathbb{R}^d$, $|a(w_0(x))| \leq C_0$ with C_0 an absolute constant and $|\cdot|$ denoting Euclidean distance. Hence, the transport-velocity vector always has length bounded by C_0.

Now, given $x \in [-2\ell, 2\ell]^d$ and $0 < t < 1$, we consider all points y that can be transported to x, that is, y should satisfy

$$x = y + ta(w_0(y)). \tag{8.3}$$

If y and z are both solutions to (8.3), then when Q is large enough both points y and z would have to come from Q (because of our estimate for the size of the transport velocity). Then, $\varphi(y_2, \dots, y_d) = 1$ and similarly for z. Hence the first components of the vectors in (8.3) give $y_1 + tu_0(y_1) = x_1 = z_1 + tu_0(z_1)$. We have already noted in the univariate analysis of §6 that this implies $y_1 = z_1$. Hence, (8.1) gives $w_0(y) = w_0(z)$ and therefore (8.3) implies $y = z$. Thus, the function $\tilde{w}$ defined by

$$\tilde{w}(x,t) := w_0(y) = v_0(y_1) = v(x_1, t),$$

with y the solution to (8.3), is well defined for $x \in [-2\ell, 2\ell]^d$, and $\tilde{w} := \tilde{w}(x,t)$ satisfies the implicit equation

$$\tilde{w} = w_0(x - ta(\tilde{w})), \quad x \in [-2\ell, 2\ell]^d, \; 0 \le t < 1. \tag{8.4}$$

A direct calculation shows that $\tilde{w}$ is a weak solution to (1.1) on $[-2\ell, 2\ell]^d \times [0,1)$, and since $\tilde{w}$ is continuous and piecewise C^r on subdomains of $[-2\ell, 2\ell]^d \times [0,1)$ with smooth boundaries, $\tilde{w}$ is an entropy solution of (1.1) in this region.

If we let $t \to 1$, then $w(\,\cdot\,, t)$ converges in $L_1(\mathbb{R}^d)$ to $w_1 := w(\,\cdot\,, 1)$. On the other hand, as $t \to 1$, $w(x,t) = v(x_1, t)$ converges to $v_1(x_1)$ a.e. on $[-2\ell, 2\ell]^d$. This shows that $w_1(x) = v_1(x_1)$, a.e. $x \in [-2\ell, 2\ell]^d$, and verifies our claim for $t = 1$.

To complete the proof of the theorem in this case, we shall estimate the Besov norms of w_0 and w_1. We first show that w_1 is not in $B^\alpha_\tau(L_\tau(\mathbb{R}^d))$. It is enough to consider differences $h = h_1 e_1$, $e_1 := (1, 0, \dots, 0)$ in the first coordinate direction, with $0 < h_1 \le 1/r$. Then, $\Delta^r_h(w_1, x) = \Delta^r_{h_1}(v_1, x_1)$ whenever $\Delta^r_{h_1}(v_1, x_1) \ne 0$ and $x \in [-2\ell, 2\ell]^d$. Hence,

$$\|\Delta^r_h(w_1, \cdot)\|_{L_\tau(\mathbb{R}^d)} \ge (4\ell)^{\frac{d-1}{\tau}} \|\Delta^r_{h_1}(v_1, \cdot)\|_{L_\tau(\mathbb{R})}, \quad 0 < h_1 \le 1/r.$$

It follows that $\omega_r(w_1, s)_\tau \ge \omega_r(v_1, s)_\tau$, $0 < s \le 1/r$. Now, we know from Theorem 6.3 that v_1 is not in the Besov space $B^\alpha_\tau(L_\tau(\mathbb{R}))$ and therefore, since $r > \alpha$ (by an earlier remark of §3),

$$\int_0^{1/r} [s^{-\alpha}\omega_r(v_1, s)_\tau]^\tau \frac{ds}{s} = \infty. \tag{8.5}$$

We can replace v_1 by w_1 in (8.5) and conclude that w_1 is not in $B^\alpha_\tau(L_\tau(\mathbb{R}^d))$.

Next, we show that w_0 is in $B^\alpha_\tau(L_\tau(\mathbb{R}^d))$. Let $s > 0$ and let $h = (h_1, \dots, h_d) \in \mathbb{R}^d$ satisfy $|h| \le s$. We define the translation operator $T(h)$ by $T(h)g := g(\,\cdot + h)$, $h \in \mathbb{R}^d$. We define the difference operator D_k by $D_k g := g(\cdot + h_1 e_1 + \cdots + h_k e_k) - g(\cdot + h_1 e_1 + \cdots + h_{k-1} e_{k-1})$ for any function g on $\mathbb{R}^d$. Then, $\Delta_h = \sum_{k=1}^d D_k$. Therefore,

$$\Delta^r_h = \sum D_{k_1} \cdots D_{k_r} \tag{8.6}$$

with the sum taken over all distinct r-tuples $(k_1, \dots, k_r)$ with $k_j \in \{1, \dots, d\}$.

We consider the effect of a general term in (8.6) on w_0. Since all the operators D_k, $k = 1, \dots, d$, and $T(h)$ commute, we can write such a term as

$$D_{k_1} \cdots D_{k_r} = T(\xi) \Delta^j_{h_1 e_1} \Delta^{\lambda_2}_{h_2 e_2} \cdots \Delta^{\lambda_d}_{h_d e_d}, \tag{8.7}$$

with $0 \le \lambda_2, \dots, \lambda_d$ and $\lambda_2 + \cdots + \lambda_d = r - j$ and ξ some point in $\mathbb{R}^d$. The difference operator $\Delta^j_{h_1e_1}$ acts only with respect to x_1, and the remaining difference operators in (8.7) applies only to $x_2, \dots, x_d$. Hence,

$$D_{k_1} \cdots D_{k_r}(w_0) = T(\xi)\Delta^j_{h_1}(v_0, x_1)\Delta^{\lambda_2}_{h_2e_2}(\phi, x_2) \cdots \Delta^{\lambda_d}_{h_de_d}(\phi, x_d).$$

Since $\phi \in C^\infty$, we have for $\mu := \min(1, \tau)$ and for a constant C depending only on d, r, and μ,

$$\omega_r(w_0, s)^\mu_\tau \le C \sum_{j=0}^{r} [s^{r-j}\omega_j(v_0, s)_\tau]^\mu, \tag{8.8}$$

where for the purposes of this formula and the formulas below we define $\omega_0(v_0, s)_\tau := \|v_0\|_{L_\tau(\mathbb{R})}$ for all s. From Marchaud's inequality (see, for example, §8 of Chapter 2 in [5]), the jth term, $j \ne 0, r$, of the sum in (8.8) for $0 < s \le 1$ can be bounded by

$$Cs^{r\mu} \int_s^\infty [\sigma^{-j}\omega_r(v_0, \sigma)_\tau]^\mu \frac{d\sigma}{\sigma} \le Cs^{r\mu} \int_s^1 [\sigma^{-j}\omega_r(v_0, \sigma)_\tau]^\mu \frac{d\sigma}{\sigma} + Cs^{r\mu}\|v_0\|^\mu_{L_\tau(\mathbb{R})}.$$

Returning to (8.8), we obtain for $s \le 1$

$$\begin{aligned} \omega_r(w_0, s)^\mu_\tau &\le Cs^{r\mu}\|v_0\|^\mu_{L_\tau(\mathbb{R})} + \omega_r(v_0, s)^\mu_\tau \\ &\qquad + Cs^{r\mu} \sum_{j=1}^{r-1} \int_s^\infty [\sigma^{-j}\chi_{[0,1]}(\sigma)\omega_r(v_0, \sigma)_\tau]^\mu \frac{d\sigma}{\sigma}. \end{aligned} \tag{8.9}$$

It follows therefore from Hardy's inequality (see, e.g., §3 of Chapter 2 of [5]) that

$$\begin{aligned} &\int_0^\infty \left[s^{-\alpha}\chi_{[0,1]}(s)\omega_r(w_0, s)_\tau\right]^\tau \frac{ds}{s} \\ &\le C\left\{\|v_0\|^\tau_{L_\tau(\mathbb{R})} + \sum_{j=1}^{r} \int_0^\infty \left[s^{r-j-\alpha}\chi_{[0,1]}(s)\omega_r(v_0, s)_\tau\right]^\tau \frac{ds}{s}\right\}. \end{aligned} \tag{8.10}$$

Since $0 \le s \le 1$, the terms $s^{r-j-\alpha}$ can each be replaced by $s^{-\alpha}$. We have remarked earlier that in the definition of the Besov norm, the integral in (3.2) can be taken over $[0, 1]$. Since v_0 is in $B^\alpha_\tau(L_\tau(\mathbb{R}))$, we conclude that the right side of (8.10) is finite, and therefore w_0 is in $B^\alpha_\tau(L_\tau(\mathbb{R}^d))$.

(iii) $\tau \ge 1/\alpha$.

This case can be proved in a similar way to (ii). We let $\tilde{v}$ be the solution to the univariate problem (1.1) with transport velocity a_1 for a compactly supported univariate function $\tilde{v}_0$ in C^∞. We can choose $\tilde{v}_0$ so that no characteristics meet before time $t = 1$ and at time $t = 1$, $\tilde{v}_1 := \tilde{v}(\,\cdot\,, 1)$ has a single downward jump discontinuity at $x = 0$ of size 1. Moreover, we can require that $\tilde{v}_1$ vanishes on $(0, \infty)$ and is continuous on $(-\infty, 0)$.

As in part (ii), we consider the initial condition $w_0 = \tilde{v}_0\varphi$ with φ as in (ii). Then w_0 is in every space $B^\alpha_\tau(L_\tau(\mathbb{R}^d))$. At time $t = 1$, for $h > 0$ sufficiently small, we have

$$|\Delta^r_{he_1}(w_1, x)| \ge 1/2, \quad x = (x_1, \dots, x_d) \in [-1, 1]^d, \quad -h \le x_1 \le 0.$$

Therefore,

$$\omega_r(w_1, s)^\tau_\tau \ge C\,s, \quad s \in [0, 1]. \tag{8.11}$$

This gives that

$$|w_1|^{\tau}_{B^{\alpha}_{\tau}(L_\tau(\mathbb{R}^d))} \geq C \int_0^1 s^{-\alpha\tau}\, ds.$$

Since $\alpha\tau \geq 1$, the last integral diverges and shows that w_1 is not in $B^{\alpha}_{\tau}(L_\tau(\mathbb{R}^d))$. □

REFERENCES

[1] R. Sharpley and C. Bennett, *Interpolation of Operators*, Academic Press, New York, 1988.

[2] A. P. Calderón, *Spaces between L^1 and L^∞ and the theorem of Marcinkiewicz*, Studia Math., 26 (1966), pp. 273–299.

[3] M. Crandall and L. Tartar, *Some relations between nonexpansive and order preserving mappings*, Proc. Amer. Math. Soc., 78 (1980), pp. 1–21.

[4] R. DeVore, B. Jawerth, and V. Popov, *Compression of wavelet decompositions*, Amer. J. Math., 114 (1992), pp. 737–785.

[5] R. DeVore and G. G. Lorentz, *Constructive Approximation*, Grundlehren, Springer Verlag, New York, 1993.

[6] R. DeVore and B. Lucier, *High order regularity for conservation laws*, Indiana Univ. Math. J., 39 (1990), pp. 413–430.

[7] ———, *High order regularity for solutions of the inviscid Burgers equation*, in Nonlinear Hyperbolic Problems, Proceedings of the Advanced Workshop held in Bordeaux, France, June 13–17, 1988, Lecture Notes in Mathematics 1402, C. Carrasso, P. Charrier, B. Hanouzet, and J-L. Joly, eds., Springer-Verlag, New York, 1989, pp. 406–413.

[8] R. DeVore and V. Popov, *Interpolation of Besov spaces*, Trans. Amer. Math. Soc., 305 (1988), pp. 397–414.

[9] ———, *Interpolation spaces and non-linear approximation*, in Function Spaces and Applications, Lecture Notes in Mathematics 1302, M. Cwikel, J. Peetre, Y. Sagher, and H. Wallin, eds., Springer-Verlag, Berlin, 1988, pp. 191–205.

[10] S. N. Kružkov, *First order quasilinear equations in several independent variables*, Math. USSR Sb., 10 (1970), pp. 217–243.

[11] P. D. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock waves*, Regional Conference Series in Applied Mathematics 11, Society for Industrial and Applied Mathematics, Philadelphia, 1973.

[12] B. J. Lucier, *A moving mesh numerical method for hyperbolic conservation laws*, Math. Comp., 46 (1986), pp. 59–69.

[13] ———, *On Sobolev regularizations of hyperbolic conservation laws*, Comm. Part. Diff. Equations, 10 (1985), pp. 1–28.

[14] ———, *Regularity through approximation for scalar conservation laws*, SIAM J. Math. Anal., 19 (1988), pp. 763–773.

[15] P. Petrushev, *Direct and converse theorems for spline and rational approximation and Besov spaces*, in Function Spaces and Applications, Lecture Notes in Mathematics 1302, M. Cwikel, J. Peetre, Y. Sagher, and H. Wallin, eds., Springer-Verlag, Berlin, 1988, pp. 363–377.

[16] A. I. Vol'pert, *The spaces BV and quasilinear equations*, Math. USSR Sb., 2 (1967), pp. 225–267.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 708–724, May 1996

006

NONUNIQUENESS AND UNIQUENESS IN THE INITIAL-VALUE PROBLEM FOR BURGERS' EQUATION*

DANIEL B. DIX†

Abstract. A sharp local existence and uniqueness theory for the initial-value problem for Burgers' equation is given in the Sobolev spaces H^s, $-1/2 < s \leq 0$. It is proved that these results cannot be extended to any $s < -1/2$ because uniqueness fails. A particular nontrivial solution is found which converges to 0 in the H^s-norm as $t \to 0^+$.

Key words. uniqueness, nonuniqueness, initial-value problem, Burgers' equation, Sobolev spaces, distributional solutions

AMS subject classifications. 35A07, 35K55, 35Q53, 35R25

Introduction. It is now an established tradition in the mathematical study of pure initial-value problems for evolutionary partial differential equations to pose these problems for inital data in various Sobolev spaces X and to consider as solutions mappings from a time interval $[0,T]$ into X which may not have sufficient regularity to qualify as classical solutions. The concept of a solution in the distributional sense has been an especially fruitful one in the study of linear problems. And by default, it has become the environment within which nonlinear problems are studied. This is despite the obvious difficulty that there is no generally satisfactory way to multiply two distributions to obtain another distribution. This fact leads us to suspect that if we insist on considering nonlinear initial-value problems with distributional initial data and distributional solutions then eventually, as the distributions become progressively more singular, some departure from the linear pattern will be observed.

As an example of what we mean by the "linear pattern," let us consider how the pure initial-value problem for the linear heat equation (see §1 for an explanation of any unfamiliar notation)

$$\begin{aligned} u_t - u_{xx} &= 0 \qquad &&\text{on } (0,T)\times\mathbb{R}, \\ u(t) &\to u_0 \qquad &&\text{in } \mathcal{S}'(\mathbb{R}) \text{ as } t\to 0^+ \end{aligned}$$

behaves in the scale of Sobolev spaces H^s, $s \in \mathbb{R}$. If $u_0 \in H^s$, then it is not hard to see that there is a solution $u \in C([0,T],H^s)$ for all $T > 0$ ($u(t)\hat{}(k) = e^{-k^2 t}\hat{u}_0(k)$ for all $t \geq 0$). This solution satisfies the equation in the sense that both terms of the equation are in $\mathfrak{D}'((0,T),H^{s-2})$ and add up to zero in that space of distributions. Actually, u is determined by a smooth function on $(0,T)\times\mathbb{R}$ which is a classical solution of the heat equation. Thus local (in fact global) solutions of the heat equation exist for (almost) arbitrarily singular initial data. Furthermore, this solution is the unique one to this problem in a very general sense. In order to make sense out of the initial condition, we should have that the distribution $u \in \mathfrak{D}'((0,T),H^s)$ be given by a mapping in $L^1_{\text{loc}}((0,T),H^s)$. It can be shown (see Theorem 2.1 below) that any such mapping which satisfies the equation in the sense of $\mathfrak{D}'((0,T),H^{s-2})$ is uniquely determined by its initial data (assumed in the topology of $\mathcal{S}'(\mathbb{R})$ as $t\to 0^+$).

*Received by the editors March 12, 1994; accepted for publication September 22, 1994. The existence theory is from the author's Ph.D. thesis directed by Charles Amick at the University of Chicago, 1988. The example of nonuniqueness was found while the author was a Postdoctoral Member of the Institute for Mathematics and Its Applications, 1988–1989. The uniqueness theory was obtained at the University of South Carolina.

†Department of Mathematics, 400E LeConte, University of South Carolina, Columbia, SC 29208 (dix@math.sc.edu).

In this paper, we will explore, by the vehicle of a simple but canonical example, what seems to be emerging as the "nonlinear pattern." Our example will be Burgers' equation,

$$\begin{aligned} u_t + uu_x - u_{xx} &= 0 \qquad \text{on } (0,T) \times \mathbb{R}, \\ u(t) &\to u_0 \qquad \text{in } H^s \text{ as } t \to 0^+. \end{aligned}$$

We will show that in a certain well-defined sense this initial-value problem is locally well posed in H^s for $s > -1/2$ (local existence, uniqueness, and continuous dependence on intial conditions) and that it fails to be well posed (in the same sense) when $s < -1/2$. Although we do not know if local solutions must exist emanating from arbitrary initial data in H^s, for any $s < -1/2$, we do know that the initial-value problem fails to be well posed because such solutions can *fail to be unique*. We will prove this by exhibiting a particular nontrivial solution $v(t)$ of Burgers' equation which converges to zero in the H^s topology as $t \to 0^+$, $s < -1/2$. Obviously, the nonuniqueness is not a result of singular data but rather one of allowing a potential solution to converge to its initial data in a too weak of a topology.

In order to justify our usage of the phrase "nonlinear pattern," we will now compare our results with some obtained by Haraux and Weissler [7] for certain semilinear parabolic equations. One particular case of their results concerned the pure initial-value problem for the semilinear heat equation

$$\begin{aligned} u_t - |u|^4 u - u_{xx} &= 0 \qquad \text{on } (0,T) \times \mathbb{R}, \\ u(t) &\to u_0 \qquad \text{in } L^p(\mathbb{R}) \text{ as } t \to 0^+ \end{aligned}$$

in the scale of Banach spaces L^p, $p \geq 1$. They showed that this initial-value problem is locally well posed in L^p for $p \geq 2$ but that it is not well posed for $1 \leq p < 2$ because, again, uniqueness fails. Their counterexample was a nontrivial self-similar solution which converges to zero as $t \to 0^+$ in the L^p-norm when $1 \leq p < 2$. Our counterexample for Burgers' equation is not (nor could be) a similarity solution. Haraux and Weissler obtained no information about whether or not multiple solutions can arise from any nonzero initial datum. In our example, we find infinitely many solutions emanating (in a weak sense) from general initial data.

Our proof of the existence and uniqueness of solutions of the initial-value problem for Burgers' equation with initial data in H^s, $s > -1/2$, is a robust contraction-mapping argument. The critical Sobolev index $s = -1/2$ is where the argument fails to yield a local solution for data of arbitrary size. This is probably a happy consequence of the simplicity of Burgers' equation and the fact that we set up the argument in the correct norms. But our results do suggest that in other situations, if the proper norms are used, breakdown of the contraction-mapping argument could signal nonuniqueness. Other heuristics might be proposed to explain "why" the critical index is $s = -1/2$. For example, is it merely coincidence that if u solves Burgers' equation, then so does $u^\lambda(x,t) = \lambda u(\lambda x, \lambda^2 t)$ and $\|u^\lambda(\cdot,t)\|_{\dot{H}^s} = \lambda^{s+1/2}\|u(\cdot,\lambda^2 t)\|_{\dot{H}^s}$? Unfortunately, we do not know how to give a generally applicable heuristic to detect the critical index.

In contrast to the robustness of our existence and uniqueness theory, our construction of the counterexample depends on the detailed knowledge of the solutions of Burgers' equation which was furnished by the Hopf–Cole transformation [8]. Although this procedure will not apply to most other equations, the flip side is that we obtain a detailed picture of what happens in the important special case of Burgers' equation. One can only conjecture at this point how much of that detailed picture is generically true. We also remark that the Hopf–Cole transformation naturally yields a local existence theory in L^1 but not in H^s, $s > -1/2$. Our existence and uniqueness theory has nothing to do with the Hopf–Cole transformation.

Lest the reader imagine that consideration of initial-value problems in Sobolev spaces of negative indices is only sensible for parabolic equations, where one is saved, so to speak, by the strong smoothing effect, we would like to point out the recent work of Kenig, Ponce, and Vega [9] concerning the Korteweg–deVries equation

$$\begin{aligned} u_t + uu_x + u_{xxx} &= 0 \qquad \text{on } (0,T) \times \mathbb{R}, \\ u(t) &\to u_0 \qquad \text{in } H^s \text{ as } t \to 0^+. \end{aligned}$$

They showed, using contraction-mapping arguments, that this problem is well posed in H^s for $s > -5/8$. The evidence is not yet compelling enough to identify $s = -5/8$ as the critical Sobolev index for this problem. However, in light of our work, it is not unreasonable to conjecture that there will be a critical index $s_0 \leq -5/8$ and that uniqueness will fail for $s < s_0$.

One feature of any result asserting uniqueness of solutions of a nonlinear initial-value problem lying in $C([0,T], H^s)$ for $s < 0$ is the need to make rigorous sense of the equation for every potential solution. The usual procedure, followed for example by Kenig, Ponce, and Vega, is to prove that a solution exists in a proper subset of $C([0,T], H^s)$ and then to prove that it is unique only with respect to competitors lying in that subset. That is because in order for something to be a competitor, we must be able to decide if it satisfies the equation or not. Since we only know how to multiply functions, not general distributions, this proper subset invariably consists only of functions. So the appearance of generality which results from discussion of the space $C([0,T], H^s)$ of distributions is an illusion, since all the potential solutions are functions.

In this respect, the uniqueness theorem we present is different. We will give a well-defined sense in which every element of $C([0,T], H^s)$ either is or is not a solution of Burgers' equation. Thus we avoid making the a priori restriction to a proper subset of $C([0,T], H^s)$. However, from the standpoint of hindsight, our uniqueness theorem shows that only functions in the proper subset occur as solutions. The idea behind the way we make sense of the equation is derived from some fairly recent advances in the nonlinear theory of generalized functions; c f. Colombeau [3], Biagioni [1], Egorov [6], and Biagioni and Oberguggenburger [2]. The viewpoint of these works is to replace the distributional setting, which has some inadequacies for nonlinear problems, with the more general and more flexible setting of Colombeau generalized functions. In this work, we extract from this above mentioned body of work only the analytical ideas we need to address our problem, and hence we do not need to discuss Colombeau generalized functions directly. In a sequel to this paper [5], we will succeed in repairing the nonuniqueness herein descibed by explicitly adopting the formalism of Colombeau generalized functions.

1. Notation.

- $J = (0, \infty)$.
- $u_t, \partial_x u = u_x$ denote the partial derivatives of u with respect to t and x, respectively.
- $(\mathfrak{F}g)(k) = \hat{g}(k) = \int_{-\infty}^{\infty} e^{-ikx} g(x)\,dx$ is the Fourier transform of g. If $u(x,t)$ is a function of $x \in \mathbb{R}$ and $t \geq 0$, then $u(t)\hat{}(k)$ is the Fourier transform of u in the x variable alone.

- $m(D)$, where $m : \mathbb{R} \to \mathbb{C}$ is a function, is the Fourier multiplier operator defined formally by the rule $[m(D)g]\widehat{\ }(k) = m(k)\hat{g}(k)$. Thus $D = -i\partial_x$.
- $\mathfrak{D}((0,T))$ is the space of Schwartz test functions; $\mathfrak{D}'((0,T),X)$, where X is a topological vector space, is the space of continuous linear maps $\mathfrak{D}((0,T)) \to X$, i.e., the space of X-valued distributions.
- $\mathcal{S}(\mathbb{R})$ is the space of Schwartz tempered test functions; $\mathcal{S}'(\mathbb{R})$ is its dual as a topological vector space, i.e., the space of tempered distributions. $\mathcal{S}'(\mathbb{R},X)$, where X is a topological vector space, is the space of continuous linear maps $\mathcal{S}(\mathbb{R}) \to X$, i.e., the space of X-valued tempered distributions.
- $\|u(x)\|_{L^p(x)} = \|u\|_{L^p}$, where $L^p = L^p(\mathbb{R})$ is the ordinary Lebesgue space.
- $L^1_{\mathrm{loc}}(\Omega, X)$ is the set of measurable functions $\Omega \to X$ (where Ω is a subset of $\mathbb{R}^n$ contained in the closure of its interior and X is a Banach space) that are locally (Bochner) integrable, i.e., they are integrable on every subset of Ω which is compact in $\mathbb{R}^n$. When $X = \mathbb{R}$, we denote $L^1_{\mathrm{loc}}(\Omega,\mathbb{R}) = L^1_{\mathrm{loc}}(\Omega)$.
- $\beta(k) = (1+k^2)^{1/2}$ for all $k \in \mathbb{R}$.
- H^m, where $m \in \mathbb{R}$, is the usual Hilbert–Sobolev space consisting of all tempered distributions u such that $\beta(k)^m\hat{u}(k) \in L^2(k)$. $\|u\|_{H^m} = \|\beta(k)^m\hat{u}(k)\|_{L^2(k)}$.
- $\dot{H}^m$ is the homogeneous Hilbert–Sobolev space of all u such that $|k|^m\hat{u}(k) \in L^2(k)$; $\|u\|_{\dot{H}^m} = \|\,|k|^m\hat{u}(k)\|_{L^2(k)}$.
- L^p_s is the weighted Lebesgue space of all $v(k)$ such that $\beta(k)^s v(k) \in L^p(k)$. $\|u\|_{L^p_s} = \|\beta^s u\|_{L^p}$.
- $C(I,X)$, where $I \subset \mathbb{R}$ is an interval and X is a Banach space, denotes the space of all continuous mappings $I \to X$. $BC(I,X)$ is the Banach space of all bounded continuous mappings $I \to X$. $C^n(I,X)$ is the space of all n-times continuously differentiable mappings $I \to X$.
- $B(a,b) = \int_0^1 (1-x)^{a-1}x^{b-1}\,dx$ is the Beta function, $a>0, b>0$.
- $BC_s((0,T],H^r)$ is defined in §2 before Theorem 2.3. $BC_{s,0}((0,T],H^r)$ is defined at the beginning of §3.
- $F_{s,r}([0,T])$ is defined in §3.
- C_H is defined in Theorem 2.3. C_I is defined in Theorem 3.4.

2. The homogeneous linear equation.

In this section, we will consider the following initial-value problem:

$$u_t - u_{xx} = 0 \qquad \text{on } (0,T)\times\mathbb{R}, \tag{2.1}$$

$$u(t) \to u_0 \qquad \text{in } \mathcal{S}'(\mathbb{R}) \text{ as } t \to 0^+, \tag{2.2}$$

where $u_0 \in \mathcal{S}'(\mathbb{R})$. First we shall show that if a solution to this problem exists in the class $L^1_{\mathrm{loc}}((0,T),H^s)$, where $s \in \mathbb{R}$, then it is the only solution in that class. Then we will show that if $u_0 \in H^s$, then u defined by $u(t)\widehat{\ }(k) = e^{-k^2t}\hat{u}_0(k)$ not only lies in $L^1_{\mathrm{loc}}(J,H^s)$ and satisfies (2.1) and (2.2) but is actually much more regular and satisfies the equation in a classical sense. We will also introduce some natural function spaces which contain the solution.

THEOREM 2.1. *Suppose $s \in \mathbb{R}$, $0 < T \le \infty$, and $u \in L^1_{\mathrm{loc}}((0,T),H^s)$. Since $u \in L^1_{\mathrm{loc}}((0,T),H^{s-2}) \subset \mathfrak{D}'((0,T),H^{s-2})$ and $u_{xx} \in L^1_{\mathrm{loc}}((0,T),H^{s-2})$, we have both*

$$u_t, u_{xx} \in \mathfrak{D}'((0,T),H^{s-2}).$$

If u is such that (2.1) is satisfied in $\mathfrak{D}'((0,T),H^{s-2})$ and (2.2) is satisfied for $u_0 = 0$, then $u = 0$ in $\mathfrak{D}'((0,T),H^s)$.

Proof. Let $\iota : H^s \to H^{s-2}$ be the inclusion and let $[u]$ denote the distribution associated to the mapping u. Let $[\iota u]'$ denote the distributional derivative of $[\iota u] \in \mathfrak{D}'((0,T), H^{s-2})$. Since $\mathfrak{F} : H^{s-2} \to L^2_{s-2}$ is a continuous linear isomorphism, we can apply it to (2.1), which now reads

$$[\iota u]' - [\partial_x^2 u] = 0, \tag{2.3}$$

to obtain $[\mathfrak{F}\iota u]' - [\mathfrak{F}\partial_x^2 u] = 0$ in $\mathfrak{D}'((0,T), L^2_{s-2})$, where we have used the chain rule $\mathfrak{F}([\iota u]') = (\mathfrak{F}[\iota u])'$ and the property $\mathfrak{F}[\iota u] = [\mathfrak{F}\iota u]$, both of which are simple to prove. Since $\mathfrak{F}\iota u \in L^1_{\text{loc}}((0,T), L^2_{s-2})$, we have that $(\mathfrak{F}\iota u)\beta^{s-2} \in L^1_{\text{loc}}((0,T), L^2)$ and thus $(\mathfrak{F}\iota u)\beta^{s-2}$ has a measurable representative which we will denote by $u(t)\hat{}(k)\beta(k)^{s-2}$ for all $(t,k) \in (0,T) \times \mathbb{R}$. $u(t)\hat{}(k)$ is a measurable representative of the mapping $\mathfrak{F}\iota u$. Then $\mathfrak{F}\partial_x^2 u$ has a representative given by $-k^2 u(t)\hat{}(k)$. So the Fourier-transformed equation becomes $[u(t)\hat{}(k)]_t + [k^2 u(t)\hat{}(k)] = 0$ in $\mathfrak{D}'((0,T) \times \mathbb{R})$. Here we have used the isomorphism provided by the Schwartz kernels theorem [11], $\mathfrak{D}'((0,T), \mathfrak{D}'(\mathbb{R})) \cong \mathfrak{D}'((0,T) \times \mathbb{R})$, and the fact that under this isomorphism we have the correspondences

$$\begin{aligned}
[\mathfrak{F}\iota u] &\mapsto [u(t)\hat{}(k)], \\
[\mathfrak{F}\iota u]' &\mapsto [u(t)\hat{}(k)]_t, \\
[\mathfrak{F}\partial_x^2 u] &\mapsto [-k^2 u(t)\hat{}(k)].
\end{aligned}$$

Since $u(t)\hat{}(k)$ and $k^2 u(t)\hat{}(k)$ are in $L^1_{\text{loc}}((0,T) \times \mathbb{R})$, there exists a function $v(t)\hat{}(k) \in L^1_{\text{loc}}((0,T) \times \mathbb{R})$ which is absolutely continuous in $t \in (0,T)$ for every fixed $k \in \mathbb{R}$, $[v(t)\hat{}(k)] = [u(t)\hat{}(k)]$, and if $\partial_t(v(t)\hat{}(k))$ is defined to be the classical t-partial derivative of $v(t)\hat{}(k)$ whenever it exists (almost everywhere on $(0,T) \times \mathbb{R}$) and to be zero elsewhere, then $[\partial_t(v(t)\hat{}(k))] = [u(t)\hat{}(k)]_t$; see [10, Thm. 9.5, p. 24]. Thus we have the equation $\partial_t(v(t)\hat{}(k)) + k^2 v(t)\hat{}(k) = 0$ holding for almost every $(t,k) \in (0,T) \times \mathbb{R}$. So by Fubini's theorem, there exists a conull set $S \subset \mathbb{R}$ such that for every $k \in S$ we have that $\partial_t(v(t)\hat{}(k)) + k^2 v(t)\hat{}(k) = 0$ for almost every $t \in \mathbb{R}$. Since the product of two absolutely continuous functions is absolutely continuous and the product rule for differentiating the product holds, we have that if $k \in S$ then $\partial_t(e^{k^2 t} v(t)\hat{}(k)) = e^{k^2 t}\partial_t(v(t)\hat{}(k)) + k^2 e^{k^2 t} v(t)\hat{}(k) = 0$ for almost every $t \in (0,T)$. So $e^{k^2 t} v(t)\hat{}(k) = h(k)$ for all $(t,k) \in (0,T) \times S$, where $h(k)$ is some function defined for $k \in S$. From this equation, it follows that $h \in L^1_{\text{loc}}(\mathbb{R})$ and hence $[h] \in \mathfrak{D}'(\mathbb{R})$. For every $\theta \in \mathfrak{D}(\mathbb{R})$, we have $\int_{-\infty}^{\infty} v(t)\hat{}(k)\theta(k)\,dk = \int_{-\infty}^{\infty} e^{-k^2 t} h(k)\theta(k)\,dk \to \int_{-\infty}^{\infty} h(k)\theta(k)\,dk$ as $t \to 0^+$ by the dominated-convergence theorem. Let $v \in L^1_{\text{loc}}((0,T), L^2_s)$ be the mapping induced by $v(t)\hat{}(k)$. Then $v(t) \to [h]$ weak-$*$ in $\mathfrak{D}'(\mathbb{R})$ as $t \to 0^+$. On the other hand, $(\mathfrak{F}u)(t) \to 0$ in $\mathcal{S}'(\mathbb{R})$ as $t \to 0^+$. This implies $(\mathfrak{F}u)(t) \to 0$ weak-$*$ in $\mathcal{S}'(\mathbb{R})$ and hence weak-$*$ in $\mathfrak{D}'(\mathbb{R})$ as $t \to 0^+$. If $[h] \neq 0$, then there exist two disjoint weak-$*$ open neighborhoods of $[h]$ and 0, respectively, and hence there exists an $\epsilon > 0$ such that $v(t) \neq (\mathfrak{F}u)(t)$ for all $0 < t < \epsilon$. But $v(t) = (\mathfrak{F}u)(t)$ for almost every $t \in (0,T)$, which is a contradiction. Hence $[h] = 0$. So $v = 0$ and therefore $[u] = [\mathfrak{F}^{-1}v] = 0$ in $\mathfrak{D}'((0,T), H^s)$. □

One should compare our uniqueness theorem for (2.1) to the classical uniqueness theorems for the heat equation (see Widder [12]), which cover even the case where the initial data and the solution are allowed to grow exponentially as $|x| \to \infty$ but concern classical solutions.

THEOREM 2.2. *Suppose $s \in \mathbb{R}$, $u_0 \in H^s$, and $u(t)\hat{}(k) = e^{-k^2 t}\hat{u}_0(k)$ for all $t \in J$ and all $k \in S = \{\xi \in \mathbb{R} \mid |\hat{u}_0(\xi)| < \infty\}$. Then $u(t)\hat{}(k)$ induces a mapping v whose inverse Fourier transform $u = \mathfrak{F}^{-1}v$ satisfies $u \in C^n(J, H^r)$ for all $n \geq 0$ (an integer)* *and $r \geq s$ (a real number). Furthermore, $u(t) \to u_0$ in H^s as $t \to 0^+$ and the unique C^∞ representative of u is a classical solution of* (2.1).

Proof. The set S is conull in $\mathbb{R}$. First we will show that the mapping v induced by $u(t)\hat{\ }(k)$ satisfies $v \in C(J, L^2_r)$ for all real $r \geq s$. The following estimate shows that $v(t) \in L^2_r$ for all $t \in J$:

$$\begin{aligned} \|v(t)\|_{L^2_r} &= \|\beta(k)^{r-s}e^{-k^2t}\beta(k)^s\hat{u}_0(k)\|_{L^2(k)} \\ &\leq \sqrt{2\pi}\|\beta(k)^{r-s}e^{-k^2t}\|_{L^\infty(k)}\|u_0\|_{H^s} \\ &\leq \sqrt{2\pi}\|u_0\|_{H^s}\begin{cases} 1 & \text{if} \quad t \geq (r-s)/2, \\ \left(\frac{r-s}{2t}\right)^{(r-s)/2} e^{t-(r-s)/2} & \text{if} \quad t < (r-s)/2. \end{cases} \end{aligned} \tag{2.4}$$

Now let $t, s \geq \epsilon > 0$. A consequence of the mean-value theorem is the inequality

$$|e^{-k^2t} - e^{-k^2s}| \leq \sqrt{2}k^2e^{-\epsilon k^2}|t-s|.$$

This will imply $v \in C([\epsilon, \infty), L^2_r)$ since

$$\begin{aligned} \|v(t) - v(s)\|_{L^2_r} &= \|\beta(k)^{r-s}[e^{-k^2t} - e^{-k^2s}]\beta(k)^s\hat{u}_0(k)\|_{L^2(k)} \\ &\leq \sqrt{4\pi}\|k^2\beta(k)^{r-s}e^{-\epsilon k^2}\|_{L^\infty(k)}\|u_0\|_{H^s}|t-s|. \end{aligned}$$

So this first result follows since ϵ was arbitrary. This immediately implies that $u = \mathcal{F}^{-1}v \in C(J, H^r)$.

Now we will show that u satisfies (2.1). If $k \in S$, then $\partial_t(u(t)\hat{\ }(k)) = -k^2u(t)\hat{\ }(k)$ for every $t \in J$. $-k^2u(t)\hat{\ }(k)$ induces the mapping $\mathcal{F}\partial_x^2u \in C(J, L^2_{r-2})$. Therefore, $\partial_t(\beta(k)^{r-2}u(t)\hat{\ }(k))$ induces a mapping in $C(J, L^2)$ (namely $(\mathcal{F}\partial_x^2u)\beta^{r-2}$). Also since $u \in C(J, H^r)$, we have that $\iota u \in C(J, H^{r-2})$ and thus $\beta(k)^{r-2}u(t)\hat{\ }(k)$ induces a mapping in $C(J, L^2)$ (namely $(\mathcal{F}\iota u)\beta^{r-2}$). It follows that $[(\mathcal{F}\iota u)\beta^{r-2}]' = [(\mathcal{F}\partial_x^2u)\beta^{r-2}]$ in $\mathfrak{D}'(J, L^2)$. Now multiplication by β^{2-r} is a continuous linear map from L^2 to L^2_{r-2}. So by the chain rule, $[\mathcal{F}\iota u]' = [\mathcal{F}\partial_x^2u]$. Applying the chain rule again for the continuous linear map $\mathcal{F}^{-1} : L^2_{r-2} \to H^{r-2}$, we obtain $[\iota u]' - [\partial_x^2u] = 0$. So u satisfies (2.3). Since both ιu and ∂_x^2u are in $C(J, H^{r-2})$, we have that $\iota u \in C^1(J, H^{r-2})$. Since r is arbitrary, this implies that $u \in C^1(J, H^r)$ for all $r \in \mathbb{R}$. By the chain rule, $\partial_x^2u \in C^1(J, H^{r-2})$. Thus $\iota u \in C^2(J, H^{r-2})$ and this argument may be iterated to show that $u \in C^n(J, H^r)$ for every $n \geq 0$ and $r \in \mathbb{R}$. So there exits a unique C^∞ representative of u. So u satisfies (2.1) in a classical sense.

Finally, we will show that $u(t) \to u_0$ in H^s as $t \to 0^+$.

$$\begin{aligned} \|u(t) - u_0\|_{H^s} &= (2\pi)^{-1/2}\|\beta(k)^s(e^{-k^2t} - 1)\hat{u}_0(k)\|_{L^2} \\ &= \frac{1}{\sqrt{2\pi}}\left(\int_{-\infty}^{\infty} |e^{-k^2t} - 1|^2\beta(k)^{2s}|\hat{u}_0(k)|^2\,dk\right)^{1/2}. \end{aligned}$$

Since $|e^{-k^2t} - 1| \leq 2$ and $\beta(k)^{2s}|\hat{u}_0(k)|^2 \in L^1(k)$, we can apply the dominated-convergence theorem. □

Now we will introduce a family of specially weighted spaces which will be used in our discussion of the initial-value problem for the nonlinear Burgers' equation.

DEFINITION. *Suppose $T > 0, s \leq r$ are real numbers. Let $BC_s((0,T], H^r)$ denote the class of all mappings $u \in C([0,T], H^s) \cap C((0,T], H^r)$ that also satisfy the* *condition*

$$\text{(2.5)} \qquad \|u\|_{BC_s((0,T],H^r)} \overset{\text{def}}{=} \frac{1}{\sqrt{2\pi}} \sup_{t\in[0,T]} \|\beta(k)^s\beta(kt^{1/2})^{r-s}[\mathfrak{F}u(t)](k)\|_{L^2(k)} < \infty.$$

$BC_s((0,T],H^r)$ is a Banach space. This particular time-dependent weighting causes the $r-s$ derivatives higher than the first s derivatives of u to be increasingly deemphasized by the norm (2.5) as $t \to 0^+$. Hence the H^s-norm of $u(t)$ will be bounded absolutely for all $t \in (0,T)$ whereas the H^r-norm will be allowed to blow up as $t \to 0^+$ (assuming $r > s$). The fact that the solutions that we have found have this sort of behavior as $t \to 0^+$ can be seen from estimate (2.4).

THEOREM 2.3. *Suppose* $0 < T < \infty$, $s \in \mathbb{R}$, $u_0 \in H^s$, *and* $u(t)\hat{}(t) = e^{-k^2t}\hat{u}_0(k)$ *for all* $t \in (0,T)$ *and all* $k \in \mathbb{R}$ *such that* $|\hat{u}_0(k)| < \infty$. *Then for any real number* $r \geq s$, *we have* $u \in BC_s((0,T],H^r)$ *and* $\|u\|_{BC_s((0,T],H^r)} \leq C_H\|u_0\|_{H^s}$, *where* $C_H = \|\beta(\omega)^{r-s}e^{-\omega^2}\|_{L^\infty(\omega)}$.

Proof. The indicated continuity follows from Theorem 2.2. For the rest, we estimate

$$\begin{aligned}\|u\|_{BC_s((0,T],H^r)} &= \frac{1}{\sqrt{2\pi}} \sup_{t\in[0,T]} \|\beta(k)^s\beta(kt^{1/2})^{r-s}e^{-k^2t}\hat{u}_0(k)\|_{L^2(k)} \\ &\leq \sup_{t\in[0,T]} \|\beta(kt^{1/2})^{r-s}e^{-k^2t}\|_{L^\infty(k)} \frac{1}{\sqrt{2\pi}}\|\beta(k)^s\hat{u}_0(k)\|_{L^2(k)} \\ &= \|\beta(\omega)^{r-s}e^{-\omega^2}\|_{L^\infty(\omega)}\|u_0\|_{H^s}. \qquad \square\end{aligned}$$

3. The inhomogeneous linear equation. In preparation for consideration of the nonlinear problem (4.1)–(4.2) and to make precise the relation of (4.1) to its associated integral equation, we consider the following inhomogeneous heat equation:

$$\text{(3.1)} \qquad u_t - u_{xx} = f \qquad \text{in } \mathfrak{D}'((0,T),H^{s-2}),$$

$$\text{(3.2)} \qquad u(t) \to 0 \qquad \text{in } \mathcal{S}'(\mathbb{R}) \text{ as } t \to 0^+,$$

where $s \in \mathbb{R}$ and $f \in \mathfrak{D}'((0,T),H^{s-2})$ will be more precisely specified below. The usual procedure would be to specify some class of fs and then derive the properties of the solution u. However, first we will do the opposite: we will define our class of fs by the class where the solution u lives. This class of very general fs will be useful in our uniqueness proof. Then we will show that if g and h live in the same class as u does, then $f = \partial_x(gh)$ is contained in the class of fs. This result will, of course, have immediate applications in the next section.

DEFINITION. *Suppose* $0 < T < \infty$ *and that* $s \leq r$ *are real numbers. Define the class* $BC_{s,0}((0,T],H^r)$ *to consist of those* $u \in BC_s((0,T],H^r)$ *such that* $u(0) = 0$.

We embed $BC_{s,0}((0,T],H^r)$ into $\mathcal{S}'(\mathbb{R},H^s)$ in the following way. Suppose $u \in BC_{s,0}((0,T],H^r)$. Let $v \in BC([0,\infty),H^r)$ be the solution of the initial-value problem (2.1)–(2.2) for the heat equation with initial data given by $u(T)$. Define the mapping $\tilde{u} \in BC(\mathbb{R},H^s)$ by the rule

$$\tilde{u}(t) = \begin{cases} 0 & \text{if } t \leq 0, \\ u(t) & \text{if } t \in [0,T], \\ v(t-T) & \text{if } t \geq T. \end{cases}$$

This mapping $\tilde{u}$ determines an element of $\mathcal{S}'(\mathbb{R},H^s)$ in the usual manner.

DEFINITION. *Suppose $0 < T < \infty$ and that $s \le r$ are real numbers. Define $F_{s,r}([0,T])$ to consist of all $f \in \mathcal{S}'(\mathbb{R}, H^{s-2})$ such that $\tilde{u}_t - \tilde{u}_{xx} = f$ in $\mathcal{S}'(\mathbb{R}, H^{s-2})$ for some $u \in BC_{s,0}((0,T], H^r)$. Define*

$$\|f\|_{F_{s,r}([0,T])} \overset{\text{def}}{=} \|u\|_{BC_s((0,T],H^r)}.$$

If $v \in BC_{s,0}((0,T], H^r)$ also satisfies $\tilde{v}_t - \tilde{v}_{xx} = f$ in $\mathcal{S}'(\mathbb{R}, H^{s-2})$, then $(u-v)_t - (u-v)_{xx} = 0$ and $u(t) - v(t) \to 0$ in H^s as $t \to 0^+$. So by Theorem 2.1 we have $u = v$. Thus the norm in $F_{s,r}([0,T])$ is well defined. Also, the map $BC_{s,0}((0,T], H^r) \to F_{s,r}([0,T]) : u \mapsto \tilde{u}_t - \tilde{u}_{xx}$ is an isometric isomorphism.

Notice that every $f \in F_{s,r}([0,T])$ has compact support contained in $[0,T]$. If $\hat{f}$ denotes the Fourier transform of f in both the x and t variables, then $\hat{f}(k,\tau)$ can be thought of as the restriction to the real line of an entire L^2_{s-2}-valued function of τ. $\hat{u}$ can then be defined as the temperate L^2_s-valued distribution $\hat{u}(k,\tau) = \lim_{\epsilon \to 0^+} \hat{f}(k, \tau - i\epsilon)/(\epsilon + i\tau + k^2)$. u can then be recovered by taking the inverse Fourier transform. This procedure for recovering u from f can be made considerably more concrete when f is known to be in a more restricted class, which we will soon discuss.

First we will need a lemma which will enable us to estimate $f = \partial_x(gh)$ when g and h are in $BC_s((0,T], H^r)$.

LEMMA 3.1. *Suppose $a > 0$, $r \ge 0$ are real numbers, and $g, h \in H^r$. Then*

$$\|\beta(ka)^r [gh]\hat{}\,(k)\|_{L^\infty(k)} \le 2^{r/4} \|\beta(Da)^r g\|_{L^2} \|\beta(Da)^r h\|_{L^2}.$$

Proof. See proof of Lemma 2.3.1 in Dix [4]. □

LEMMA 3.2. *Suppose $0 < T \le 1$, $-1 < s \le 0$ and $r \ge 0$ are real numbers. Then*

$$\begin{aligned}\sup_{t \in (0,T]} t^{|s|/2} \|\beta(Dt^{1/2})^r u(t)\|_{L^2} \le \|u\|_{BC_s((0,T],H^r)} &\le \sup_{t \in (0,T]} t^{|s|/2} \|\beta(Dt^{1/2})^r u(t)\|_{L^2} \\ &\quad + \sup_{t \in [0,T]} \| |D|^s \beta(Dt^{1/2})^r u(t)\|_{L^2}.\end{aligned}$$

Proof. Since $0 \le t \le T \le 1$, we have the estimate $(1 + k^2 t)/(1 + k^2) \ge t$ for all $k \in \mathbb{R}$. Thus $t^{|s|/2} \le \beta(k)^s \beta(kt^{1/2})^{-s} \le |k|^s + t^{|s|/2}$. Now the desired estimates follow directly from the definition of $BC_s((0,T], H^r)$. □

LEMMA 3.3. *Suppose $0 < T < \infty$, $-1 < s \le 0$, and $r < 2s + 1/2$ are real numbers. Then there is a constant such that for every pair of measurable mappings $g, h : (0,T] \to L^2$, both satisfying the estimate*

$$M_g = \sup_{t \in (0,T]} t^{|s|/2} \|g(t)\|_{L^2} < \infty,$$

we have $u(t) \to 0$ in H^r as $t \to 0^+$, where $u : (0,T] \to L^2$ is defined by the rule

$$u(t) = \int_0^t e^{-D^2(t-\tau)} \partial_x [g(\tau) h(\tau)] \, d\tau$$

for all $t \in (0,T]$.

Proof. To see why the Fourier transform in x and the integral in τ commute, consult the proof of Theorem 2.3.3(1) in Dix [4]. Now estimate

$$\begin{aligned}
&\left\|\beta(D)^r \int_0^t \partial_x e^{-D^2(t-\tau)}[g(\tau)h(\tau)]\,d\tau\right\|_{L^2} \\
&= \frac{1}{\sqrt{2\pi}}\left\|\beta(k)^r \int_0^t ik e^{-k^2(t-\tau)}[g(\tau)h(\tau)]^\wedge(k)\,d\tau\right\|_{L^2(k)} \\
&\le \frac{1}{\sqrt{2\pi}}\int_0^t \|\,|k|\beta(k)^r e^{-k^2(t-\tau)}\|_{L^2(k)}\cdot\|[g(\tau)h(\tau)]^\wedge(k)\|_{L^\infty(k)}\,d\tau \\
&\le \frac{1}{\sqrt{2\pi}}M_gM_h\int_0^t \frac{\|\,|k|\beta(k)^r e^{-k^2(t-\tau)}\|_{L^2(k)}}{\tau^{|s|}}\,d\tau.
\end{aligned}\tag{3.3}$$

We have two cases. First of all, if $r \le 0$ then $\beta(k)^r \le |k|^r$. It will be convenient to assume $r > -3/2$. Once proved for such values of r, the result also follows for smaller values. So we see that the integral in (3.3) is bounded by

$$t^{(1/2+2s-r)/2}\|\,|\omega|^{1+r}e^{-\omega^2}\|_{L^2(\omega)}B((1/2+2s-r)/2, 1+s).$$

This tends to 0 as $t \to 0^+$ under our assumptions. The second case is if $r > 0$. This implies that $0 < r < 1/2 + 2s$ and thus $s > -1/4$. But we also have that $r < 2s + 1/2 \le 1/2$, and therefore $\beta(k)^r \le 1 + |k|^r$ for all $k \in \mathbb{R}$. Thus the integral in (3.3) is bounded by

$$\begin{aligned}
&t^{(1/2+2s-r)/2}\|\,|\omega|^{1+r}e^{-\omega^2}\|_{L^2(\omega)}B((1/2+2s-r)/2, 1+s) \\
&+t^{1/4+s}\|\omega e^{-\omega^2}\|_{L^2(\omega)}B(1/4, 1+s).
\end{aligned}$$

Clearly this tends to 0 as $t \to 0^+$. So the lemma is true. □

THEOREM 3.4. *Suppose* $0 < T \le 1$, $-1/2 < s \le 0$, $r \ge 0$, *and* $s \le q < r + 1/2$ *are real numbers. Let* $g, h \in X$, *where* $X = BC_s((0, T], H^r)$. *Then there exists a constant* $C_I > 0$, *depending only on* q, r, *and* s *such that* $\partial_x(gh) \in F_{s,q}([0, T])$ *and*

$$\|\partial_x(gh)\|_{F_{s,q}([0,T])} \le C_I T^{(s+1/2)/2} M_g M_h \le C_I T^{(s+1/2)/2}\|g\|_X\|h\|_X,$$

where $M_g = \sup_{t\in(0,T]} t^{|s|/2}\|\beta(D\sqrt{t})^r g\|_{L^2}$. *Furthermore, if* $u \in BC_{s,0}((0, T], H^r)$ *satisfies* (3.1)–(3.2) *with* $f = \partial_x(gh)$, *then* u *can be recovered from* f *via the formula*

$$u(t) = \int_0^t e^{-D^2(t-\tau)} f(\tau)\,d\tau$$

for all $t \in (0, T]$.

Proof. The plan of the proof is as follows: Let u be defined in terms of g and h by the above formula. First we will show that $u \in BC_{s,0}((0, T], H^q)$. Then we will show that u satisfies (3.1)–(3.2) with $f = \partial_x(gh)$. By Lemma 3.2, $gh \in L^1((0, T), L^1) \subset L^1((0, T), H^{s-1})$, and thus $\partial_x(gh) \in L^1((0, T), H^{s-2})$. So $f = \partial_x(gh)$ is a distribution in $\mathfrak{D}'((0, T), H^{s-2})$ or in $\mathcal{S}'((\mathbb{R}, H^{s-2})$ in the usual way (define f to vanish outside $(0, T)$). So if $u_t - u_{xx} = f$ in $\mathfrak{D}'((0, T), H^{s-2})$ and $u_t = f + u_{xx} \in L^1((0, T), H^{s-2})$, we have that u is absolutely continuous on $[0, T]$ with values in H^{s-2}. Therefore, $\tilde{u}_t - \tilde{u}_{xx} = f$ in $\mathcal{S}'((\mathbb{R}, H^{s-2})$. This shows that $f \in F_{s,q}([0, T])$.

First we will bound the $BC_s((0,T],H^q)$-norm of u. It suffices to do the case where $r \leq q < r+1/2$, since the norm is a nondecreasing function of q. We will use the second estimate of Lemma 3.2 to control u. Since $k^4(t-\tau)\tau \geq 0$, we have the inequality $(1+k^2t) \leq (1+(t-\tau)k^2)(1+\tau k^2)$ holding for all $k \in \mathbb{R}$ and $0 \leq \tau \leq t$. So

$$\beta(k\sqrt{t}) \leq \beta(k\sqrt{t-\tau})\beta(k\sqrt{\tau}).$$

Using this inequality and Lemma 3.1, we have that

$$\begin{aligned} t^{|s|/2}&\left\|\beta(D\sqrt{t})^q \int_0^t \partial_x e^{-D^2(t-\tau)}[g(\tau)h(\tau)]\,d\tau\right\|_{L^2} \\ &= \frac{t^{|s|/2}}{\sqrt{2\pi}}\left\|\int_0^t \beta(k\sqrt{t})^q ike^{-k^2(t-\tau)}[g(\tau)h(\tau)]\hat{}(k)\,d\tau\right\|_{L^2(k)} \\ &\leq \frac{t^{|s|/2}}{\sqrt{2\pi}}\int_0^t \|ik\beta(k\sqrt{t})^{q-r}\beta(k\sqrt{t-\tau})^r e^{-k^2(t-\tau)}\|_{L^2(k)} \\ &\qquad\qquad \cdot \|\beta(k\sqrt{\tau})^r[g(\tau)h(\tau)]\hat{}(k)\|_{L^\infty(k)}\,d\tau \\ &\leq \frac{2^{r/4}t^{|s|/2}}{\sqrt{2\pi}}M_gM_h\int_0^t \frac{\|ik\beta(k\sqrt{t})^{q-r}\beta(k\sqrt{\tau})^r e^{-k^2\tau}\|_{L^2(k)}}{(t-\tau)^{|s|}}\,d\tau \\ &\leq \frac{2^{r/4-1/2}}{\pi^{1/2}}t^{1/4+s/2}M_gM_h\int_0^1 \frac{\|\omega\beta(\omega\sigma^{-1/2})^{q-r}\beta(\omega)^r e^{-\omega^2}\|_{L^2(\omega)}}{\sigma^{3/4}(1-\sigma)^{|s|}}\,d\sigma. \end{aligned}$$

Estimating in a similar way, we get

$$\begin{aligned} &\left\||D|^s\beta(D\sqrt{t})^q \int_0^t \partial_x e^{-D^2(t-\tau)}[g(\tau)h(\tau)]\,d\tau\right\|_{L^2} \\ &= \frac{1}{\sqrt{2\pi}}\left\|\int_0^t |k|^s\beta(k\sqrt{t})^q ike^{-k^2(t-\tau)}[g(\tau)h(\tau)]\hat{}(k)\,d\tau\right\|_{L^2(k)} \\ &\leq \frac{1}{\sqrt{2\pi}}\int_0^t \|\,|k|^{1+s}\beta(k\sqrt{t})^{q-r}\beta(k\sqrt{t-\tau})^r e^{-k^2(t-\tau)}\|_{L^2(k)} \\ &\qquad\qquad \cdot \|\beta(k\sqrt{\tau})^r[g(\tau)h(\tau)]\hat{}(k)\|_{L^\infty(k)}\,d\tau \\ &\leq \frac{2^{r/4}}{\sqrt{2\pi}}M_gM_h\int_0^t \frac{\|\,|k|^{1+s}\beta(k\sqrt{t})^{q-r}\beta(k\sqrt{\tau})^e e^{-k^2\tau}\|_{L^2(k)}}{(t-\tau)^{|s|}}\,d\tau \\ &\leq \frac{2^{r/4-1/2}}{\pi^{1/2}}t^{(s+1/2)/2}M_gM_h\int_0^1 \frac{\|\omega\beta(\omega\sigma^{-1/2})^{q-r}\beta(\omega)^r e^{-\omega^2}\|_{L^2(\omega)}}{\sigma^{3/4+s/2}(1-\sigma)^{-s}}\,d\sigma. \end{aligned}$$

Using Lemma 3.2, we can bound M_g by $\|g\|_X$, and likewise for h. The integrals remaining in these estimates can be shown to be finite when $q-r < 1/2$ by a similar argument to that in the proof of Lemma 2.3.2 in Dix [4]. These two estimates together give the estimate stated in the theorem.

The proof that $u : (0,T] \to H^q$ is continuous is very similar to the proof of Theorem 2.3.3(1) in Dix [4]. The fact that $u(t) \to 0$ in H^s as $t \to 0^+$ follows from Lemma 3.3. Thus we see that $u \in BC_{s,0}((0,T],H^q)$.

The proof that u satisfies (3.1) is almost exactly the same as the proof of Theorem 2.3.3(3) in Dix [4]. □

4. Local existence in $C([0,T],H^s)$, $s>-1/2$. In this section, we prove the existence of a local solution to the initial-value problem

$$u_t+\tfrac{1}{2}\partial_x(u^2)-u_{xx}=f \qquad \text{on } (0,T)\times\mathbb{R}, \tag{4.1}$$

$$u(t)\to u_0 \qquad \text{in } H^s \text{ as } t\to 0^+, \tag{4.2}$$

where $-1/2<s\le 0$, $u_0\in H^s$, and $f\in F_{s,0}([0,T])$. At the same time, we will also consider the nearby problems and how the solution we construct depends on u_0 and f. This viewpoint will then prove to be useful in our discussion of uniqueness in the next section. This initial-value problem for $s>0$ can be done in a similar way to (in fact it is easier than) what we present here.

THEOREM 4.1. *Let $0<T<\infty$, $-1/2<s\le 0\le r$, and constants K and L satisfy the condition*

$$2[C_HK+L]C_IT^{(s+1/2)/2}<1.$$

Let U be the closed ball in H^s centered at 0 and of radius K. Let V be the closed ball in $F_{s,r}([0,T])$ centered at 0 and of radius L. Then for all $u_0\in U$ and all $f\in V$, there exists a unique $u=\mathcal{U}(u_0,f)\in BC_s((0,T],H^r)$ satisfying

$$\|u\|_{BC_s((0,T],H^r)}<\frac{1}{C_IT^{(s+1/2)/2}},$$

(4.1) *in $\mathfrak{D}'((0,T),H^{s-2})$, and* (4.2). *The mapping*

$$\mathcal{U}:U\times V\to BC_s((0,T],H^r):(u_0,f)\mapsto\mathcal{U}(u_0,f)$$

is Lipschitz on $U\times V$ with respect to the metric inherited from $H^s\times F_{s,r}([0,T])$.

Proof. Let X denote the Banach space $BC_s((0,T],H^r)$. Let Y denote the Banach space $F_{s,r}([0,T])$. If $R>0$, consider the complete metric space X_R defined to be the closed ball in X centered at 0 and of radius R. Define $\Lambda=U\times V$. Consider the operator $A:X_R\times\Lambda\to X$ defined for every $(u,u_0,f)\in X\times\Lambda$ by the rule

$$A(u,u_0,f)(t)=e^{-D^2t}u_0+\int_0^t e^{-D^2(t-\tau)}f(\tau)\,d\tau-\frac{1}{2}\int_0^t e^{-D^2(t-\tau)}\partial_x[u(\tau)^2]\,d\tau \tag{4.3}$$

for all $t\in[0,T]$. We are denoting by $\int_0^t e^{-D^2(t-\tau)}f(\tau)\,d\tau$ the solution w of the inhomogeneous heat equation with right-hand side f and zero initial data that is uniquely determined by f. We use this notation even though we have not demonstrated how to make sense of it except for fs in a subclass of V. This will not cause any trouble, though, since we will only be using the fact that the mapping from f to w is an isometric isomorphism from Y onto $BC_{s,0}((0,T],H^r)\subset X$.

If we use Theorems 2.3 and 3.4, we find that

$$\|A(u,u_0,f)\|_X\le C_HK+L+\frac{1}{2}C_IT^{(s+1/2)/2}\|u\|_X^2,$$

$$\|A(u,u_0,f)-A(\tilde{u},u_0,f)\|_X\le\frac{1}{2}C_IT^{(s+1/2)/2}\|u+\tilde{u}\|_X\|u-\tilde{u}\|_X.$$

So the conditions that $A(\cdot,u_0,f)$ map X_R into itself and be a contraction there uniformly for $(u_0,f)\in\Lambda$ are

$$C_HK+L+\frac{1}{2}C_IT^{(s+1/2)/2}R^2\le R,$$

$$\kappa=C_IT^{(s+1/2)/2}R<1.$$

A necessary and sufficient condition that a number $R > 0$ exists satisfying these two inequalities is that

$$2[C_H K + L]C_I T^{(s+1/2)/2} < 1.$$

We are assuming this condition holds, and hence we know that there is a fixed point $u = \mathcal{U}(u_0, f)$ satisfying $u = A(u, u_0, f)$. For all $(u_0, f) \in \Lambda$, the fixed point $u = \mathcal{U}(u_0, f)$ is unique in X_R, where R satisfies both of the above inequalities. In particular, the fixed point satisfies the estimate we stated. On the other hand, any solution u of (4.1)–(4.2) contained in the open ball we stated must first of all be contained in X_R for some R satisfying the necessary inequalities and must also be a fixed point $u = A(u, u_0, f)$ (see Theorem 4.2 below). Hence $u = \mathcal{U}(u_0, f)$. By Theorems 2.2 and 3.4, the fixed point satisfies (4.1) in $\mathfrak{D}'((0,T), H^{s-2})$ and (4.2).

In order to prove the asserted Lipschitz continuity of $\mathcal{U}$, let $u = \mathcal{U}(u_0, f)$ and $v = \mathcal{U}(v_0, g)$ for $(u_0, f), (v_0, g) \in \Lambda$. Then

$$\begin{aligned} \|u - v\|_X &= \|A(u, u_0, f) - A(v, v_0, g)\|_X \\ &\leq \|A(u, u_0, f) - A(v, u_0, f)\|_X + \|A(v, u_0, f) - A(v, v_0, g)\|_X \\ &\leq \kappa \|u - v\|_X + C_H \|u_0 - v_0\|_{H^s} + \|f - g\|_Y. \end{aligned}$$

Since $\kappa < 1$, we therefore have

$$\|\mathcal{U}(u_0, f) - \mathcal{U}(v_0, g)\|_X \leq \frac{1}{1-\kappa}[C_H \|u_0 - v_0\|_{H^s} + \|f - g\|_Y]. \qquad \square$$

The statement of this theorem shows that the local solution is as regular as f is. However, it leaves the false impression that the time interval of existence should shrink as the measured regularity of f increases. The largest time period T of existence which can be obtained via the contraction mapping argument corresponds to $r = 0$. However, the relation between the regularity of f and that of u does not depend on T at all, as the following result shows.

THEOREM 4.2. *Suppose* $0 < T < \infty$, $-1/2 < s \leq 0 \leq r$, $u_0 \in H^s$, $f \in F_{s,r}([0,T])$, *and* $u \in BC_s((0,T], L^2)$ *satisfies* (4.1) *in* $\mathfrak{D}'((0,T), H^{s-2})$ *and* (4.2). *Then u satisfies the integral equation $u = A(u, u_0, f)$, where the operator A is defined in* (4.3) *and* $u \in BC_s((0,T], H^r)$. *In particular, if $f = 0$, then $u \in BC_s((0,T], H^r)$ for all $r \geq 0$ and is represented by a smooth function on $(0,T) \times \mathbb{R}$ which is a classical solution of* (4.1).

Proof. First we will show that $u = A(u, u_0, f)$. This follows because $u - A(u, u_0, f)$ solves the homogeneous heat equation with initial data zero and hence by Theorem 2.1 must vanish. We have $u = A(u, 0, 0) + A(0, u_0, 0) + A(0, 0, f)$. By Theorem 2.3, $A(0, u_0, 0) \in BC_s((0,T], H^r)$. By definition, $A(0, 0, f) \in BC_s((0,T], H^r)$. By Theorem 3.4, $A(u, 0, 0)$ is almost 1/2 of an x-derivative smoother than u is (a priori). Using this smoothing effect, in a finite number of steps, we can infer that u actually lies in $BC_s((0,T], H^r)$.

When $f = 0$, we can iterate this same argument infinitely many times, showing that $u \in BC_s((0,T], H^r)$ for all $r \geq 0$. Using equation (4.1), we can then infer regularity of u in the t-variable (cf. the proof of Theorem 3.1(2) in Dix [4] for the details). Thus we have that u is in fact represented by a smooth function on $(0,T) \times \mathbb{R}$. Since u solves (4.1) in a distributional sense, its smooth representative is a classical solution of (4.1). $\square$

5. Uniqueness of solutions in $C([0, T], H^s)$, $s > -1/2$. Our goal in this section is to show that there is at most one solution u to (4.1)–(4.2) in $C([0,T],H^s)$. Here $0 < T < \infty$ is not necessarily small enough so that we know at least one solution exists via a contraction mapping argument as in the previous section. Our first result shows that solutions in $BC_s((0,T],L^2)$ of (4.1)–(4.2) are unique without restricting the size of the competing solution or the size of T.

THEOREM 5.1. *Suppose* $0 < T < \infty$, $-1/2 < s \leq 0$, $u_0 \in H^s$, $f \in F_{s,0}([0,T])$. *Suppose* $u, v \in BC_s((0,T],L^2)$ *are both solutions of* (4.1) *(in the sense that all three terms are in* $\mathfrak{D}'((0,T),H^{s-2})$ *and sum to zero in that space) and* (4.2). *Then* $u = v$.

Proof. Let T' denote the least upper bound of the set

$$\{T'' \in [0,T] |\ u(t) = v(t) \text{ in } H^s \text{ for all } t \in [0,T'']\}.$$

Since $u, v \in C([0,T],H^s)$, we have $u(T') = v(T')$. Suppose, by way of contradiction, that $T' < T$. Define $u_0' = u(T') = v(T')$. Define $f' \in F_{s,0}([0,T-T'])$ as follows. Let $w \in BC_{s,0}((0,T],L^2)$ be such that $w_t - w_{xx} = f$ (see §3 for the precise sense in which w is uniquely determined). Define $w'(t) = w(t+T') - e^{-D^2t}w(T')$ for all $t \in [0,T-T']$. Clearly, $w' \in BC_{s,0}((0,T-T'],L^2)$. Let this w' be considered as an element of $\mathcal{S}'(\mathbb{R},H^s)$ as in §3 and define $f' = w_t' - w_{xx}'$. We will also use, as a shorthand for this construction, the notation $f'(t) = f(t+T')$ for all $t \in [0,T-T']$. Define $u'(t) = u(t+T')$, $v'(t) = v(t+T')$ for all $t \in [0,T-T']$. Clearly, $u', v' \in BC_s((0,T-T'],L^2)$ are both solutions of the same initial-value problem and hence by Theorem 4.2 satisfy the integral equations $u' = A(u',u_0',f')$ and $v' = A(v',u_0',f')$ on $[0,T-T']$. Following the proof of Theorem 4.1, we can choose a number $T'' \in (0,T-T']$ and a number $R > 0$ such that $A(\cdot,u_0',f')$ maps the closed ball in $BC_s((0,T''],L^2)$ centered at 0 of radius R into itself and is a contraction there, and both u' and v' are in this ball. By the uniqueness of the fixed points of contraction mappings, we have that $u'(t) = v'(t)$ for all $t \in [0,T'']$. This implies that $u(t) = v(t)$ for all $t \in [0,T'+T'']$, which contradicts the definition of T'. Hence $T' = T$ and we are done. □

Our next order of business is to make sense of the equation (4.1) for an arbitrary function in $C([0,T],H^s)$. First we note that $X = BC_s((0,T],L^2)$ is a dense subset of $C([0,T],H^s)$; in fact, it contains the dense subset $C([0,T],L^2)$. Let $Y = F_{s,0}([0,T])$. Define the nonlinear mapping $B : X \to H^s \times Y : u \mapsto (u(0), \tilde{v}_t - \tilde{v}_{xx} + \frac{1}{2}\partial_x(u^2))$, where $v(t) = u(t) - e^{-D^2t}u(0)$ for all $t \in [0,T]$. Since $v \in BC_{s,0}((0,T],L^2)$, we have $\tilde{v}_t - \tilde{v}_{xx} \in Y$. (Note also that $v_t - v_{xx} = u_t - u_{xx}$ in $\mathfrak{D}'((0,T),H^{s-2})$, and so in this sense we henceforth will write $u_t - u_{xx} \in Y$.) Theorem 3.4 shows that $\partial_x(u^2)/2 \in Y$. Thus $B(u)$ is well defined on X with values in $H^s \times Y$. B can be considered to be a densely defined discontinuous nonlinear operator on $C([0,T],H^s)$. Let $\text{gra}(B) \subset X \times H^s \times Y \subset C([0,T],H^s) \times H^s \times Y$ denote the graph of the operator B.

DEFINITION. *Suppose* $(u_0, f) \in H^s \times Y$. *We say* $u \in C([0,T],H^s)$ *satisfies* (4.1)–(4.2) *if* (u, u_0, f) *is contained in the closure of* $\text{gra}(B)$ *with respect to the topology of the ambient space* $C([0,T],H^s) \times H^s \times Y$; *i.e., there exists a family* $\{(u^\epsilon, u_0^\epsilon, f^\epsilon)\}_{\epsilon>0}$ *contained in* $X \times H^s \times Y$ *satisfying*

$$u_t^\epsilon + \frac{1}{2}\partial_x[(u^\epsilon)^2] - u_{xx}^\epsilon = f^\epsilon \qquad \text{in } Y \subset \mathfrak{D}'((0,T),H^{s-2}),$$

$$u^\epsilon(0) = u_0^\epsilon \qquad \text{in } H^s$$

such that

$$\begin{aligned} u^\epsilon &\to u && \text{in } C([0,T],H^s),\\ u_0^\epsilon &\to u_0 && \text{in } H^s,\\ f^\epsilon &\to f && \text{in } Y \end{aligned}$$

as $\epsilon \to 0^+$.

Before we prove the uniqueness of solutions of (4.1)–(4.2) in this above-defined sense, we will comment on the origin of this definition. The primary source of this formulation comes from the theory of generalized functions; cf. Colombeau [3] and Egorov [6]. In those theories, one associates to each distribution $u \in C([0,T],H^s(\mathbb{R}))$ families $\{u^\epsilon\}_{\epsilon>0}$ of smooth approximations which converge to u as $\epsilon \to 0^+$. One then defines $u_0^\epsilon(x) = u^\epsilon(x,0)$ and $f^\epsilon(x,t) = u_t^\epsilon(x,t) + \frac{1}{2}\partial_x[u^\epsilon(x,t)^2] - u_{xx}^\epsilon(x,t)$. Then one has various senses in which one can impose the equation (4.1). The strongest sense is that of equality of Colombeau generalized functions, where one would require that some norm of the difference $f^\epsilon - \tilde{f}^\epsilon$, where $\{\tilde{f}^\epsilon\}_{\epsilon>0}$ is a family of smooth approximations "canonically" associated to the distribution f, tend rapidly to zero as $\epsilon \to 0^+$, where the rate of convergence could be made arbitrarily rapid by choosing a better mollifier to generate the "canonical" smooth approximations. The weakest sense is that f^ϵ should tend to f as distributions as $\epsilon \to 0^+$. The weaker the sense in which the equation is imposed, the more far reaching the uniqueness theorem. Our sense is intermediate between these two extremes in that we require $f^\epsilon \to f$ in the space of distributions Y as $\epsilon \to 0^+$. Since we do not say anything about the rate of convergence, e.g., $O(\epsilon^N)$, we also do not need to say anything about mollifiers or "canonical" smooth approximations. Also, since we do not need u^ϵ to be smooth in order to make sense of the terms in the equation, we only require $u^\epsilon \in X$.

The other source of inspiration for our formulation is the usual theory of densely defined unbounded *linear* operators B between Banach spaces $\mathcal{X}$ and $\mathcal{Y}$. In that theory, the pairs (x,y) in the closure of the graph, provided that closure is itself a graph, are the ones where the equation $Bx = y$ makes sense. The graph of B can be closed in $\mathcal{X} \times \mathcal{Y}$ even if the domain of B is a proper subset of $\mathcal{X}$. However, this domain will in general depend on the space $\mathcal{Y}$. In our application, we chose the target space $H^s \times F_{s,0}([0,T])$ as generally and as naturally as we could see how to. This choice leads to the nice property that the graph of our nonlinear operator B is closed (see Theorem 5.3 below).

THEOREM 5.2. *If* $0 < T \leq 1$, $-1/2 < s \leq 0$, $u_0 \in H^s$, *and* $f \in F_{s,0}([0,T])$, *then there exists only one solution* $u \in C([0,T],H^s)$ *to the initial-value problem* (4.1)–(4.2) *in the above sense.*

Proof. Let $u, v \in C([0,T],H^s)$ be solutions to the initial-value problem (4.1)–(4.2) in the above sense. Arguing as in the proof of Theorem 5.1, we see that it suffices to show that there exists a number $T' \in (0,T]$ such that $u(t) = v(t)$ for all $t \in [0,T']$. Let $Y = F_{s,0}([0,T])$. Define $K = \|u_0\|_{H^s} + 1$ and $L = \|f\|_Y + 1$. Choose $T' \in (0,T]$ such that the condition in Theorem 4.1 is satisfied for T'. Define $\tilde{u} = \mathcal{U}(u_0, f)$. By symmetry, it suffices to show that $u(t) = \tilde{u}(t)$ for all $t \in [0,T']$. Let the family $\{(u^\epsilon, u_0^\epsilon, f^\epsilon)\}_{\epsilon>0}$ be given as above for the solution u. Assume $\epsilon > 0$ is sufficiently small such that $\|u_0^\epsilon\|_{H^s(\mathbb{R})} < K$ and $\|f^\epsilon\|_Y < L$. Define $\tilde{u}^\epsilon = \mathcal{U}(u_0^\epsilon, f^\epsilon) \in BC_s((0,T'],L^2)$ to be the solution whose existence is asserted in Theorem 4.1. By Theorem 5.1, we have $u^\epsilon = \tilde{u}^\epsilon$ on $[0,T']$. By the continuity asserted in Theorem 4.1, we have that

u^ϵ converges to $\tilde{u}$ in $BC_s((0,T'],L^2)$ as $\epsilon \to 0^+$. Since this implies convergence in $C([0,T'],H^s)$, we have that $u = \tilde{u}$ on $[0,T']$. □

THEOREM 5.3. *If* $0 < T \leq 1$, $-1/2 < s \leq 0$, *and the densely defined operator* B *is as we have described above, then the graph* $\operatorname{gra}(B)$ *is closed in*

$$C([0,T],H^s) \times H^s \times F_{s,0}([0,T]).$$

Proof. Let the family $\{(u^\epsilon, u_0^\epsilon, f^\epsilon)\}_{\epsilon>0}$ be given in $\operatorname{gra}(B)$ converging in the above space to (u, u_0, f). By Theorem 5.2, we have that if $0 < T' \leq T$ is sufficiently small, then for all $T_0 \in [0, T-T']$ we have $u(t+T_0) = \mathcal{U}(u(T_0), f(\cdot + T_0))$ for all $t \in [0,T']$. ($f(\cdot + T_0)$ refers to the shorthand notation introduced in the proof of Theorem 5.1.) This implies in a straightforward manner that $u \in BC_s((0,T],L^2)$ and that $(u,u_0,f) \in \operatorname{gra}(B)$. □

6. Nonuniqueness of solutions in $C([0, T], H^s)$, $s < -1/2$.

Consider the well-known "N-wave" solution of Burgers' equation (see Whitham [13])

$$u(x,t) = \frac{x}{t}\frac{\sqrt{a/t}e^{-x^2/(4t)}}{1+\sqrt{a/t}e^{-x^2/(4t)}} = -2\partial_x \ln[1+\sqrt{a/t}e^{-x^2/(4t)}],$$

where $a > 0$. In the above reference, Whitham remarks in passing about the difficulty of interpreting this solution as a solution of an initial-value problem. We make the following assertions about this solution:

(1) for every $\psi \in \mathcal{S}(\mathbb{R})$, we have $\lim_{t\to 0^+} \int_{-\infty}^{\infty} u(x,t)\psi(x)\,dx = 0$, and thus $u(t) \to 0$ in $\mathcal{S}'(\mathbb{R})$ as $t \to 0^+$;

(2) for every $1/4 < s \leq 3/4$, we have $\sup_{t>0} t^s \|u(t)\|_{L^2} < \infty$;

(3) $\lim_{t\to 0^+} t^{1/4}\|u(t)\|_{L^2} = \infty$;

(4) if $s < -1/2$, then $\|u(t)\|_{H^s} \to 0$ as $t \to 0^+$.

To prove assertion (1), we first integrate by parts:

$$\int_{-\infty}^{\infty} u(x,t)\psi(x)\,dx = 2\int_{-\infty}^{\infty} \psi'(x)\ln[1+\sqrt{a/t}e^{-x^2/(4t)}]\,dx.$$

Now define the following function:

$$v(x,t) = \begin{cases} \ln[1+\sqrt{a/t}e^{-x^2/(4t)}] & \text{if } x^2 \geq 2t\ln(a/t) \\ 0 & \text{otherwise.} \end{cases}$$

Thus we have

$$\int_{-\infty}^{\infty} u(x,t)\psi(x)\,dx = 2\int_{-\infty}^{\infty} \psi'(x)v(x,t)\,dx + 2\int_{x^2<2t\ln(a/t)} \psi'(x)\ln[1+\sqrt{a/t}e^{-x^2/(4t)}]\,dx.$$

The first integral in tends to 0 as $t \to 0^+$ by the dominated-convergence theorem since $|v(x,t)| \leq \ln 2$ for all $(x,t) \in \mathbb{R} \times J$, and for every fixed $x \in \mathbb{R} \setminus \{0\}$, we have $v(x,t) \to 0$ as $t \to 0^+$. The second integral can be bounded in absolute value by

$$4\|\psi'\|_{L^\infty(\mathbb{R})} \ln(1+\sqrt{a/t})[2t\ln(a/t)]^{1/2},$$

which also tends to 0 as $t \to 0^+$. Thus (1) is true.

In order to prove (2), we introduce the new variable $\xi = x(4t)^{-1/2}$,

$$t^{2s}\|u(t)\|_{L^2}^2 = c\int_{-\infty}^{\infty}\left|\frac{\xi e^{-\xi^2}}{t^{3/4-s}+t^{1/4-s}a^{1/2}e^{-\xi^2}}\right|^2 d\xi. \tag{6.1}$$

It is an elementary calculation to show that for every $1/4 \le s \le 3/4$, $t > 0$, and $b > 0$, we have

$$t^{3/4-s}+t^{1/4-s}b \ge \frac{1}{2}\left[\frac{b^{3/4-s}}{(3/4-s)^{3/4-s}(s-1/4)^{s-1/4}}\right]^2.$$

Using this estimate in the above with $b = a^{1/2}e^{-\xi^2}$, we obtain

$$t^{2s}\|u(t)\|_{L^2}^2 \le c\int_{-\infty}^{\infty}\left|\frac{\xi e^{-\xi^2}}{[e^{-\xi^2}]^{2(3/4-s)}}\right|^2 dx = c\int_{-\infty}^{\infty}\left|\xi e^{-2\xi^2(s-1/4)}\right|^2 d\xi.$$

This is clearly bounded if $1/4 < s \le 3/4$, and so (2) is true.

To prove (3), we rewrite (6.1) and estimate as follows:

$$\begin{aligned} t^{1/2}\|u(t)\|_{L^2}^2 &= c\int_{-\infty}^{\infty}\left|\frac{\xi}{t^{1/2}e^{\xi^2}+a^{1/2}}\right|^2 d\xi \\ &\ge c\int_{\xi^2\le(1/2)\ln(a/t)}\left|\frac{\xi}{t^{1/2}e^{\xi^2}+a^{1/2}}\right|^2 d\xi \\ &\ge c\int_{\xi^2\le(1/2)\ln(a/t)}\frac{\xi^2}{4a}\,d\xi = c[\ln(a/t)]^{3/2}. \end{aligned}$$

Since this tends to ∞ as $t \to 0^+$, (3) is true.

To prove (4), notice that $u - A(u,0,0)$ solves the heat equation with zero initial data in a distributional sense and hence must be zero. Now use (2) and Lemma 3.3 to yield the result.

Since $a > 0$ can be chosen arbitrarily, we see that there are infinitely many solutions in $C([0,T],H^s)$, $s < -1/2$, to (4.1)–(4.2) with $u_0 = 0$ and $f = 0$. This phenomenon can be understood intuitively as follows. If v solves the heat equation, then $u = -2\partial_x \ln v$ satisfies Burgers' equation. The "N-wave" solution arises from $v = 1 + \sqrt{a/t}e^{-x^2/(4t)}$, which is a solution of the heat equation with initial data $v_0 = 1 + c\delta$, where $c > 0$ depends on a and δ is the Dirac delta distribution. When we apply the function ln, however, the part tending to $c\delta$ makes no contribution in a distributional sense. The number 1 is not special. We could consider the solution of the heat equation with initial data

$$v_0(x) = \exp\left(-\frac{1}{2}\int_{-\infty}^{x} u_0(y)\,dy\right) + c\delta(x)$$

since $u = -2\partial_x \ln v$ would then satisfy Burgers' equation with initial data $u_0 \in L^1$. The initial data would be assumed in the sense of $H^s(\mathbb{R})$, $s < -1/2$, as one can show using the same method as we used to prove (4) above. Thus there are infinitely many nonequivalent solutions emerging from every initial data in L^1. Solutions u of an inhomogeneous Burgers' equation with $f \in L^1([0,T] \times \mathbb{R})$ can also be expressed in terms of solutions v of the heat equation with a potential via the same transformation $u = -2\partial_x \ln v$ [5]. Solutions of this variant of the heat equation can also be written down explicitly using the Feynman–Kac formula. If we use the same initial data v_0 as displayed above in this formula, then we see that there is nonuniqueness of solutions of the inhomogeneous Burgers' equation of the same type. We will omit the details.

Acknowledgment. The author would like to thank Hebe Biagioni for making available various material on Colombeau generalized functions.

REFERENCES

[1] H. A. Biagioni, *A Nonlinear Theory of Generalized Functions*, Lecture Notes in Mathematics 1421, Springer-Verlag, Berlin, 1990.

[2] H. A. Biagioni and M. Oberguggenberger, *Generalized solutions to Burgers equation*, J. Differential Equations, 97 (1992), pp. 263–287.

[3] J. F. Colombeau, *Elementary Introduction to New Genralized Functions*, North–Holland Mathematics Studies 113, North–Holland, Amsterdam, 1985.

[4] D. B. Dix, *Temporal asymptotic behavior of solutions of the Benjamin–Ono–Burgers' equation*, J. Differential Equations, 90 (1991), pp. 238–287.

[5] ———, *Uniqueness of generalized function solutions of Burgers' equation*, in preparation.

[6] Y. V. Egorov, *A contribution to the theory of generalized functions*, Russian Math. Surveys, 45 (1990), pp. 1–49.

[7] A. Haraux and F. B. Weissler, *Non-uniqueness for a semilinear initial value problem*, Indiana Univ. Math. J. 31 (1982), pp. 167–189.

[8] E. Hopf, *The partial differential equation $u_t + uu_x = \mu u_{xx}$*, Comm. Pure Appl. Math., 3 (1950), pp. 201–230.

[9] C. E. Kenig, G. Ponce, and L. Vega, *The Cauchy problem for the Korteweg–de Vries equation in sobolev spaces of negative indices*, Duke Math. J., 71 (1993), pp. 1–21.

[10] B. E. Petersen, *Introduction to the Fourier Transform & Pseudo-differential Operators*, Pitman, Boston, 1983.

[11] F. Treves, *Topological Vector Spaces, Distributions, and Kernels*, Academic Press, New York, 1967.

[12] D. V. Widder, *Positive temperatures on an infinite rod*, Trans. Amer. Math. Soc., 55 (1944), pp. 85–95.

[13] G. B. Whitham, *Linear and Nonlinear Waves*, John Wiley & Sons, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 725–737, May 1996

ANALYTICITY OF SOLITARY-WAVE SOLUTIONS OF MODEL EQUATIONS FOR LONG WAVES*

YI A. LI[†] AND JERRY L. BONA[‡]

Abstract. It is shown that solitary-wave solutions of model equations for long waves have an analytic extension to a strip in the complex plane that is symmetric about the real axis. The classes of equations to which the analysis applies include equations of Korteweg–de Vries type, the regularized long-wave equations, and particular instances of nonlinear Schrödinger equations.

Key words. nonlinear dispersive wave equations, solitary waves, regularity, analyticity, Korteweg–de Vries-type equations, regularized long-wave-type equations, Schrödinger-type equations

AMS subject classifications. 30B40, 35B40, 35B60, 35B65, 35Q35, 35Q51, 35Q53, 35Q55, 35S30, 45E10, 45G10, 76B15, 76B25, 76C10

1. Introduction. This note is concerned with solitary-wave solutions of model equations for long waves and aims to cast light on their regularity properties. The prototypical example in view is the well-known travelling-wave solution

$$u(x,t) = \phi_c(x-(c+1)t) = 3c\,\mathrm{sech}^2\left(\frac{c^{1/2}}{2}(x-(c+1)t)\right) \tag{1.1}$$

of the classical Korteweg–de Vries equation

$$u_t + u_x + uu_x + u_{xxx} = 0. \tag{1.2}$$

For any positive value of c, the function of x and t defined in (1.1) via the function ϕ_c of one real variable is an exact solution of (1.2) which is infinitely differentiable and which decays rapidly to zero at $\pm\infty$. These properties are possessed by solitary-wave solutions of a considerable range of evolution equations that feature a balance between nonlinearity and dispersion. As these special travelling-wave solutions of nonlinear, dispersive wave equations are known in many cases to play a significant role in the long-term asymptotics of general classes of solutions, they have come in for detailed study in the last couple of decades. Existence and regularity theory for solitary waves has been developed recently by Benjamin et al. [3] and Weinstein [12]. Their results apply to a broad class of model equations to be introduced presently. The outcome of these theories is that the relevant profiles ϕ_c of the solitary-wave solutions are often positive C^∞-functions having a single maximum and which decay monotonically to zero at infinity, just as does the sech^2 solutions of the Korteweg–de Vries equation displayed above. Moreover, ϕ_c and all its derivatives lie in $L_1 \cap L_\infty$.

In fact, the sech^2-solitary-wave solution of (1.2) has further regularity than just C^∞-smoothness. The function ϕ_c in (1.1) defined on the real axis $\mathbb{R}$ is real analytic and admits an analytic extension to the complex strip $\{z = x+iy : |y| < \pi/c^{1/2}\}$. It is this latter property on which attention will be focused in the present study. While the theory developed here seems to apply to a considerable range of equations, the ideas are most transparently presented in the context of the following relatively concrete

*Received by the editors January 21, 1994; accepted for publication (in revised form) October 3, 1994.

[†] Department of Mathematics, University of Minnesota, Minneapolis, MN 55455.

[‡] Department of Mathematics and the Texas Institute for Computational and Applied Mathematics, University of Texas, Austin, TX 78712.

classes of model equations for waves in nonlinear dispersive media:

$$u_t + u_x + u^p u_x - (Mu)_x = 0 \qquad \text{(Korteweg–de Vries type)}, \tag{1.3}$$

$$u_t + u_x + u^p u_x + (Mu)_t = 0 \qquad \text{(regularized long-wave type)}, \tag{1.4}$$

$$iu_t - Mu + |u|^p u = 0 \qquad \text{(Schrödinger type)}. \tag{1.5}$$

In the first two models, p is a positive integer, while p is a positive even integer in (1.5). The linear operator M is a Fourier multiplier operator defined by

$$(\widehat{Mv})(\xi) = \alpha(\xi)\hat{v}(\xi) \tag{1.6}$$

whose nonnegative symbol α satisfies certain growth conditions to be spelled out presently. The linear transformation M is called the dispersion operator and its symbol α is related to the linear dispersion relation for the model in question (see Benjamin [2] or Whitham [13]).

We intend to show that as a rule, solitary-wave solutions of these model equations possess the property of being extensible to an analytic function defined on a strip in the complex plane $\mathbb{C}$, which lies symmetrically about the real axis $\mathbb{R}$ on which the wave profile is ostensibly defined. This fact is interesting in its own right, but in addition, it has implications regarding uniqueness [9] and appears to be useful in assessing whether or not a particular solitary wave is actually a soliton (cf. [5], [6], [8]).

The plan of the paper is as follows. In the next section, a few convenient notational conventions are introduced. In §3, the main result for travelling-wave solutions of Korteweg–de Vries type and regularized long-wave type is enunciated and proved. Section 4 is concerned with the analogous result for nonlinear Schrödinger equations. The paper concludes with a few comments about regularity issues related to those discussed here.

2. Notation. By $L_p = L_p(\mathbb{R})$ for p in the range $1 \le p \le \infty$, we mean the standard class of pth-power Lebesgue-integrable functions on the real line $\mathbb{R}$ with the usual modification if $p = \infty$. The standard norm on L_p will be denoted by $\| \cdot \|_p$. The Fourier transform of a Lebesgue-measurable function ϕ defined on $\mathbb{R}$ is denoted by $\hat{\phi}$ and is defined to be

$$\hat{\phi}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(x) e^{i\xi x} dx. \tag{2.1}$$

The convolution of two functions f and g defined on $\mathbb{R}$ is written $f * g$. Multiple convolution of a function with itself will appear frequently, and it is therefore convenient to introduce notation for this operation. If ϕ is a measurable function defined on $\mathbb{R}$ and n is a positive integer, define the function $\mathcal{V}_n\phi$ by the recipe

$$\mathcal{V}_1\phi = \phi,$$

and for $n > 1$,

$$\begin{aligned} \mathcal{V}_n\phi(x) &= (\phi * \mathcal{V}_{n-1}\phi)(x) \\ &= \int_{-\infty}^{\infty} \phi(x-y)\mathcal{V}_{n-1}\phi(y)dy. \end{aligned} \tag{2.2}$$

By a *solitary-wave solution* of (1.3) or (1.4) for a given positive integer p and dispersion symbol α, we shall mean a function $\phi : \mathbb{R} \to \mathbb{R}$ such that ϕ, ϕ', and $M\phi$ all lie in L_2 and such that for some positive constant c, $\phi(x-ct)$ defines an L_2-solution of (1.3) or (1.4). A similar definition will be adopted later for solutions of (1.5). As mentioned already, existence of such solutions for a wide range of symbols α has been dealt with in the recent works of Benjamin et al. [3] and Weinstein [12].

For any $x \in \mathbb{R}$, the greatest integer less than or equal to x is denoted by $\lfloor x \rfloor$.

3. Results for Korteweg–de Vries type and regularized long-wave models.

After a preparatory lemma, the principal result for equations of the types depicted in (1.3) and (1.4) is stated and proved.

LEMMA 1. *Let $c > 1$ be given. Suppose $\phi = \phi(x-ct)$ defines a solitary-wave solution of* (1.3) *or* (1.4) *for a given value of p and symbol α of the dispersion operator M. Suppose also that for some positive constants A and r, $\alpha(\xi) \geq A|\xi|^r$ for all $\xi \in \mathbb{R}$. Then the function*

$$\hat{\psi}(\xi) = \frac{\hat{\phi}(\xi)}{\sqrt{2\pi}[(p+1)(c-1)]^{\frac{1}{p}}} \tag{3.1}$$

lies in $L_1 \cap L_2$ and solves the equation

$$(1+\lambda\alpha(\xi))\,\hat{\psi}(\xi) = \mathcal{V}_{p+1}\hat{\psi}(\xi), \tag{3.2}$$

where $\lambda = 1/(c-1)$ if ϕ is a solution of (1.3) *and $\lambda = c/(c-1)$ if ϕ is a solution of* (1.4).

Proof. Suppose ϕ defines a solitary-wave solution of (1.3) as described in §2. Then

$$(-c+1)\phi' + \phi^p\phi' - M\phi' = 0,$$

from which it follows that, at least in the sense of tempered distributions,

$$[(c-1)+M]\phi - \frac{1}{p+1}\phi^{p+1} = \text{constant}. \tag{3.3}$$

Since each term on the left-hand side is an L_2-function by assumption, the constant on the right-hand side must be zero. Applying the Fourier transform to (3.3) and using (1.6) leads directly to the desired result (3.2) with $\lambda = 1/(c-1)$.

Because $c > 1$ and M has a nonnegative symbol, it follows from (3.3) with the constant equal to zero that

$$\phi = \frac{1}{p+1}[(c-1)+M]^{-1}\phi^{p+1}. \tag{3.3$'$}$$

Since $\phi \in H^1$ by assumption, the product ϕ^{p+1} is also in H^1. For any $s \in \mathbb{R}$, the linear operator $(c-1+M)^{-1}$ maps H^s into H^{s+r}. Hence it transpires from (3.3$'$) that $\phi \in H^{1+r}$. In consequence, $\phi^{p+1} \in H^{1+r}$, whence $\phi \in H^{1+2r}$, and so on. It is thus inferred that $\phi \in H^\infty$, from which it is adduced at once that

$$\int_{-\infty}^{\infty} (1+\xi^2)^m\,|\hat{\psi}(\xi)|^2 d\xi < \infty \tag{3.4}$$

for any m. An immediate consequence of (3.4) is that $\hat{\psi} \in L_1 \cap L_2$, as stated in the lemma.

The same considerations lead to the advertised result when ϕ defines instead a solitary-wave solution of (1.4). □

With this simple lemma in hand, the main issue may be confronted. The idea is to demonstrate that if ϕ defines a solitary-wave solution of (1.3) or (1.4), then its Fourier transform $\hat{\phi}$ has exponential decay at $\pm\infty$. In consequence, the Paley–Wiener theorem assures that ϕ itself is analytic in a complex strip centered about the real axis.

We begin with a special case of the main result, which will prove to be instructive and which contains the essence of the argument that applies to the more general situations.

THEOREM 2. *Let an integer $p \geq 1$ and a wave speed $c > 1$ be given. Suppose that ϕ as in Lemma 1 defines a solitary-wave solution of (1.3) or (1.4) corresponding to the dispersive symbol $\alpha(\xi) = |\xi|^m$ for some real number $m \geq 1$. Then there exists a constant $\sigma > 0$ such that for any μ with $0 < \mu < \sigma$,*

$$\sup_{\xi \in \mathbb{R}} e^{\mu|\xi|} |\hat{\phi}(\xi)| < \infty. \tag{3.5}$$

Proof. By Lemma 1, it suffices to prove (3.5) for the function $\hat{\psi}$ defined in (3.1) that satisfies equation (3.2).

For any k with $0 \leq k \leq m$ and $\lambda > 0$, define the nonnegative function f_k for $\xi \geq 0$ by

$$f_k(\xi) = \frac{\xi^k}{1 + \lambda \xi^m}.$$

It is straightforward to determine that for all $\xi \geq 0$,

$$f_k(\xi) \leq \frac{\delta_k}{\lambda^{\frac{k}{m}}}, \tag{3.6}$$

where $\delta_k = \left(\frac{k}{m}\right)^{\frac{k}{m}} \left(1 - \frac{k}{m}\right)^{1-\frac{k}{m}}$ if $0 < k < m$, and $\delta_m = \delta_0 = 1$.

Case I. $m \geq 1$ is an integer.

Suppose that $\hat{\psi}$ satisfies (3.2) and (3.4). When $0 \leq k \leq m-1$, (3.6) may be used to conclude that

$$\begin{aligned} \left|\xi^k \hat{\psi}(\xi)\right| &= \frac{|\xi|^k}{1 + \lambda|\xi|^m} \left|\mathcal{V}_{p+1} \hat{\psi}(\xi)\right| \\ &\leq \frac{\delta_k}{\lambda^{\frac{k}{m}}} \left|\mathcal{V}_{p+1} \hat{\psi}(\xi)\right| \\ &\leq \frac{1}{\lambda^{\frac{k}{m}}} \mathcal{V}_{p+1} |\hat{\psi}|(\xi) \\ &\leq \frac{1}{\lambda^{\frac{k}{m}}} \left(\frac{kp}{m} + 1\right)^{k-1} \mathcal{V}_{p+1} |\hat{\psi}|(\xi) \end{aligned} \tag{3.7}$$

for any $\xi \in \mathbb{R}$. On the other hand, for any $n \geq 0$ and any $\xi_1 \in \mathbb{R}$, we have

$$
\begin{aligned}
|\xi_1|^{m+n}|\hat{\psi}(\xi_1)| &= \frac{|\xi_1|^{m+n}}{1+\lambda|\xi_1|^m}\left|\mathcal{V}_{p+1}\hat{\psi}(\xi_1)\right| \le \frac{|\xi_1|^n}{\lambda}\mathcal{V}_{p+1}|\hat{\psi}|(\xi_1) \\
&= \frac{1}{\lambda}\int_{-\infty}^{\infty}\left|\hat{\psi}(\xi_1-\xi_2)\right|\int_{-\infty}^{\infty}\left|\hat{\psi}(\xi_2-\xi_3)\right|\int_{-\infty}^{\infty} \\
&\quad \cdots\int_{-\infty}^{\infty}\left|\left(\sum_{i=1}^{p}(\xi_i-\xi_{i+1})+\xi_{p+1}\right)^n\hat{\psi}(\xi_p-\xi_{p+1})\hat{\psi}(\xi_{p+1})\right| d\xi_{p+1}d\xi_p\cdots d\xi_2 \\
&\le \frac{1}{\lambda}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\sum_{\sum_{i=1}^{p+1}r_i=n}\frac{n!}{r_1!\cdots r_{p+1}!} \\
&\quad \cdot\left|\xi_{p+1}^{r_{p+1}}\hat{\psi}(\xi_{p+1})\prod_{i=1}^{p}(\xi_i-\xi_{i+1})^{r_i}\hat{\psi}(\xi_i-\xi_{i+1})\right| d\xi_{p+1}\cdots d\xi_2 \\
&\le \frac{1}{\lambda}\sum_{|r|=n}\binom{n}{r}\left|(\cdot)^{r_1}\hat{\psi}(\cdot)\right| * \left|(\cdot)^{r_2}\hat{\psi}(\cdot)\right| * \cdots * \left|(\cdot)^{r_{p+1}}\hat{\psi}(\cdot)\right|(\xi_1),
\end{aligned}
\tag{3.8}
$$

where $|(\cdot)^{r_i}\hat{\psi}(\cdot)|(\xi) = |\xi^{r_i}\hat{\psi}(\xi)|$ and we have introduced the standard multiindex notation $r = (r_1, \dots, r_{p+1})$, $|r| = r_1 + \cdots + r_{p+1}$, and $\binom{n}{r} = \frac{n!}{r_1!\cdots r_{p+1}!}$. If $0 \le l \le m-1$, then using (3.7) in (3.8) leads to the inequality

$$
\begin{aligned}
|\xi_1|^{m+l}|\hat{\psi}(\xi_1)| &\le \frac{1}{\lambda}\sum_{|r|=l}\binom{l}{r}\left(\prod_{i=1}^{p+1}\frac{1}{\lambda^{\frac{r_i}{m}}}\right)\overbrace{\mathcal{V}_{p+1}|\hat{\psi}| * \cdots * \mathcal{V}_{p+1}|\hat{\psi}|}^{p+1 \text{ copies of } \mathcal{V}_{p+1}|\hat{\psi}|}(\xi_1) \\
&= \frac{1}{\lambda^{1+\frac{l}{m}}}\sum_{|r|=l}\binom{l}{r}\mathcal{V}_{(p+1)^2}|\hat{\psi}|(\xi_1) = \frac{1}{\lambda^{\frac{m+l}{m}}}(p+1)^l\mathcal{V}_{(p+1)^2}|\hat{\psi}|(\xi_1).
\end{aligned}
\tag{3.9}
$$

It follows from (3.7) and (3.9) that for any $\xi \in \mathbb{R}$ and any k with $0 \le k \le 2m-1$, one has

$$
\left|\xi^k\hat{\psi}(\xi)\right| \le \frac{1}{\lambda^{\frac{k}{m}}}\left(\frac{kp}{m}+1\right)^{k-1}\mathcal{V}_{(p+1)(\lfloor\frac{k}{m}\rfloor p+1)}|\hat{\psi}|(\xi),
\tag{3.10}
$$

where, as mentioned previously, $\lfloor\frac{k}{m}\rfloor$ denotes the greatest integer less than or equal to $\frac{k}{m}$.

It is intended to establish (3.10) for all values of k, and to this end we argue by induction, supposing that the inequality (3.10) is true for all k with $0 \le k \le nm-1$ for a fixed integer $n \ge 2$. Let $k = nm + l$ for some integer l in $[0, m-1]$. Then (3.8) and the induction hypothesis allow one to infer the following inequality:

(3.11)

$$\begin{aligned}\left|\xi^k\hat{\psi}(\xi)\right| &\le \frac{1}{\lambda}\sum_{|r|=(n-1)m+l}\binom{(n-1)m+l}{r}\left|(\cdot)^{r_1}\hat{\psi}\right| * \left|(\cdot)^{r_2}\hat{\psi}\right| * \cdots * \left|(\cdot)^{r_{p+1}}\hat{\psi}\right|(\xi)\\ &\le \frac{1}{\lambda}\sum_{|r|=(n-1)m+l}\binom{(n-1)m+l}{r}\left(\prod_{i=1}^{p+1}\frac{\left(\frac{r_ip}{m}+1\right)^{r_i-1}}{\lambda^{\frac{r_i}{m}}}\right)\mathcal{V}_{(p+1)\left(\left\lfloor\frac{r_1}{m}\right\rfloor p+1\right)}|\hat{\psi}| *\\ &\quad * \mathcal{V}_{(p+1)\left(\left\lfloor\frac{r_2}{m}\right\rfloor p+1\right)}|\hat{\psi}| * \cdots * \mathcal{V}_{(p+1)\left(\left\lfloor\frac{r_{p+1}}{m}\right\rfloor p+1\right)}|\hat{\psi}|(\xi)\\ &= \frac{1}{\lambda^{\frac{nm+l}{m}}}\sum_{|r|=(n-1)m+l}\binom{(n-1)m+l}{r}\left(\prod_{i=1}^{p+1}\left(\frac{r_ip}{m}+1\right)^{r_i-1}\right)\mathcal{V}_{\sum_{j=1}^{p+1}(p+1)\left(\left\lfloor\frac{r_j}{m}\right\rfloor p+1\right)}|\hat{\psi}|(\xi).\end{aligned}$$

If inequality (3.10) is specialized to the case $k = 0$, one infers that $|\hat{\psi}| \le \mathcal{V}_{p+1}|\hat{\psi}|$. Using this fact and the elementary formula

$$\begin{aligned}&\sum_{i=1}^{p+1}(p+1)\left(\left\lfloor\frac{r_i}{m}\right\rfloor p+1\right)\\ &= (p+1)\left(p\sum_1^{p+1}\left\lfloor\frac{r_i}{m}\right\rfloor + \sum_1^{p+1}1\right) \le (p+1)\left(p\left\lfloor\sum_1^{p+1}\frac{r_i}{m}\right\rfloor + p + 1\right)\\ &= (p+1)\left(p\left\lfloor\frac{(n-1)m+l}{m}\right\rfloor + p + 1\right) = (p+1)\left(p\left\lfloor\frac{nm+l}{m}\right\rfloor + 1\right)\\ &= (p+1)\left(\left\lfloor\frac{k}{m}\right\rfloor p + 1\right),\end{aligned}$$

one obtains

$$\mathcal{V}_{\sum_1^{p+1}(p+1)(\lfloor r_i/m\rfloor p+1)}|\hat{\psi}| \le \mathcal{V}_{(p+1)\left(\left\lfloor\frac{k}{m}\right\rfloor p+1\right)}|\hat{\psi}|. \tag{3.12}$$

Using a specialization of the multinomial Abel identity (see [11, p. 26]), namely

$$\begin{aligned}&A_N(x_1, x_2, \ldots, x_M)\\ &= \sum_{|k|=N}\binom{N}{k}\prod_{i=1}^{M}(x_i + k_i)^{k_i-1}\\ &= (x_1x_2\cdots x_M)^{-1}\left(\sum_1^M x_i\right)\left(\sum_1^M x_i + N\right)^{N-1},\end{aligned}$$

and the simple relation $\sum_1^{p+1}(r_i - 1) = (n-1)m + l - p - 1$, one obtains

$$\begin{aligned}&\sum_{|r|=(n-1)m+l}\binom{(n-1)m+l}{r}\prod_{i=1}^{p+1}\left(\frac{r_ip}{m}+1\right)^{r_i-1}\\ &= \sum_{|r|=(n-1)m+l}\binom{(n-1)m+l}{r}\left(\frac{p}{m}\right)^{\sum_1^{p+1}(r_i-1)}\prod_1^{p+1}\left(\frac{m}{p}+r_i\right)^{r_i-1}\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{p}{m}\right)^{(n-1)m+l-p-1} A_{(n-1)m+l}\left(\frac{m}{p}, \frac{m}{p}, \dots, \frac{m}{p}\right) \\
&= \left(\frac{p}{m}\right)^{(n-1)m+l-p-1} \left(\frac{m}{p}\right)^{-(p+1)} \frac{(p+1)m}{p}\left(\frac{(p+1)m}{p} + (n-1)m + l\right)^{(n-1)m+l-1} \\
&= (p+1)\left(\frac{nm+l}{m}p + 1\right)^{(n-1)m+l-1} \\
&= (p+1)\left(\frac{kp}{m} + 1\right)^{k-1-m} \\
&\le \left(\frac{kp}{m} + 1\right)^{k-1}.
\end{aligned}
$$

The latter inequality, when combined with (3.11), (3.12), and the induction hypothesis, yields

$$\left|\xi^k \hat{\psi}(\xi)\right| \le \frac{1}{\lambda^{\frac{k}{m}}}\left(\frac{kp}{m}+1\right)^{k-1} \mathcal{V}_{(p+1)(\lfloor\frac{k}{m}\rfloor p+1)}|\hat{\psi}|(\xi)$$

for any integer $k \ge 0$. It follows that (3.10) holds for any $\xi \in \mathbb{R}$ and any integer $k \ge 0$.

Using the fact that $\hat{\psi} \in L_1(\mathbb{R})$, we infer that

$$
\begin{aligned}
\mathcal{V}_{(p+1)(\lfloor\frac{k}{m}\rfloor p+1)}|\hat{\psi}| &\le \|\hat{\psi}\|_2 \left\|\mathcal{V}_{(p+1)(\lfloor\frac{k}{m}\rfloor p+1)-1}|\hat{\psi}|\right\|_2 \\
&\le \|\hat{\psi}\|_2 \|\hat{\psi}\|_1 \left\|\mathcal{V}_{(p+1)(\lfloor\frac{k}{m}\rfloor p+1)-2}|\hat{\psi}|\right\|_2 \\
&\le \cdots \le \|\hat{\psi}\|_2^2 \|\hat{\psi}\|_1^{(p+1)(\lfloor\frac{k}{m}\rfloor p+1)-2},
\end{aligned}
$$

and so

$$|\xi^k \hat{\psi}(\xi)| \le \frac{\|\hat{\psi}\|_2^2}{\lambda^{k/m}}\left(\frac{kp}{m}+1\right)^{k-1} \|\hat{\psi}\|_1^{(p+1)(\lfloor k/m\rfloor p+1)-2} \tag{3.13}$$

for any $\xi \in \mathbb{R}$.

To complete the proof for Case I, consider the sequence

$$a_k = \frac{1}{k!\,\lambda^{\frac{k}{m}}}\left(\frac{kp}{m}+1\right)^{k-1} \|\hat{\psi}\|_1^{(p+1)(\frac{kp}{m}+1)} \tag{3.14}$$

for $k = 0, 1, 2, \dots$. Because the ratio $\frac{a_{k+1}}{a_k}$ takes the form

$$
\begin{aligned}
\frac{a_{k+1}}{a_k} &= \frac{k!\,\lambda^{\frac{k}{m}}\left(\frac{(k+1)p}{m}+1\right)^{k+1-1} \|\hat{\psi}\|_1^{(p+1)\left(\frac{(k+1)p}{m}+1\right)}}{(k+1)!\,\lambda^{\frac{k+1}{m}}\left(\frac{kp}{m}+1\right)^{k-1} \|\hat{\psi}\|_1^{(p+1)(\frac{kp}{m}+1)}} \\
&= \frac{1}{\lambda^{1/m}} \|\hat{\psi}\|_1^{\frac{(p+1)p}{m}} \left(\frac{p}{m} + \frac{1}{k+1}\right) \left[\left(1 + \frac{p}{m+kp}\right)^{\frac{m+kp}{p}}\right]^{\frac{p(k-1)}{pk+m}},
\end{aligned}
$$

it is readily seen that

$$\lim_{k\to\infty}\frac{a_{k+1}}{a_k}=\frac{ep}{m\ \lambda^{1/m}}\|\hat{\psi}\|_1^{\frac{(p+1)p}{m}}.$$

Hence the power series $\sum_{k=0}^{\infty}a_k\mu^k$ converges for $|\mu|<\frac{m\lambda^{1/m}}{ep}\|\hat{\psi}\|_1^{\frac{-(p+1)p}{m}}$. In consequence of (3.13) and (3.14), it is seen that for any $\xi\in\mathbb{R}$,

$$\begin{aligned}
e^{\mu|\xi|}|\hat{\psi}(\xi)| &= \sum_{k=0}^{\infty}\frac{\mu^k|\xi|^k}{k!}|\hat{\psi}(\xi)| \\
&\le \frac{\|\hat{\psi}\|_2^2}{\|\hat{\psi}\|_1^2}\sum_{k=0}^{\infty}\frac{\mu^k}{k!\ \lambda^{\frac{k}{m}}}\left(\frac{kp}{m}+1\right)^{k-1}\|\hat{\psi}\|_1^{(p+1)(\lfloor k/m\rfloor p+1)} \\
&\le \frac{\|\hat{\psi}\|_2^2}{\|\hat{\psi}\|_1^2}\sum_{k=0}^{\infty}a_k\mu^k<\infty,
\end{aligned}$$

provided $|\mu|<\frac{m\lambda^{1/m}}{ep}\|\hat{\psi}\|_1^{\frac{-(p+1)p}{m}}$. Thus the function $e^{\mu|\xi|}|\hat{\psi}(\xi)|$ appears to be uniformly bounded for such choices of μ, and this is the desired conclusion in case m is a positive integer.

Case II. $m>1$ is not an integer.

If $m_0=\lfloor m\rfloor$ and $\varrho=\max_{0\le\xi<\infty}\frac{1+\lambda\xi^{m_0}}{1+\lambda\xi^{m}}$, then it follows from (3.2) that

$$|\hat{\psi}(\xi)|\le\frac{\varrho}{1+\lambda|\xi|^{m_0}}\mathcal{V}_{p+1}|\hat{\psi}|(\xi). \tag{3.15}$$

Now one may use (3.15) and induction as in the proof of Case I to prove that

$$\left|\xi^k\hat{\psi}(\xi)\right|\le\frac{\varrho^{(p+1)\lfloor k/m_o\rfloor+1}}{\lambda^{k/m_o}}\left(\frac{kp}{m_0}+1\right)^{k-1}\mathcal{V}_{(p+1)\left(\left\lfloor\frac{k}{m_0}\right\rfloor p+1\right)}|\hat{\psi}|(\xi) \tag{3.16}$$

holds for any integer $k\ge0$ and all $\xi\in\mathbb{R}$. In consequence, the following inequality is obtained for integers k and $\xi\in\mathbb{R}$:

$$\left|\xi^k\hat{\psi}(\xi)\right|\le\frac{\varrho^{(p+1)\lfloor k/m_0\rfloor+1}}{\lambda^{k/m_0}}\left(\frac{kp}{m_0}+1\right)^{k-1}\|\hat{\psi}\|_2^2\|\hat{\psi}\|_1^{(p+1)(\lfloor k/m_0\rfloor p+1)-2}.$$

Thus it appears that for all $\xi\in\mathbb{R}$,

$$\begin{aligned}
e^{\mu|\xi|}|\hat{\psi}(\xi)| &= \sum_{k=0}^{\infty}\frac{\mu^k|\xi|^k}{k!}|\hat{\psi}(\xi)| \\
&\le \frac{\|\hat{\psi}\|_2^2}{\|\hat{\psi}\|_1^2}\sum_{k=0}^{\infty}\frac{\mu^k\varrho^{(p+1)\lfloor k/m_0\rfloor+1}}{k!\ \lambda^{k/m_0}}\left(\frac{kp}{m_0}+1\right)^{k-1}\|\hat{\psi}\|_1^{(p+1)(\lfloor k/m_0\rfloor p+1)}<\infty
\end{aligned}$$

for any μ satisfying $0<\mu<\frac{m_0\lambda^{1/m_0}}{ep\varrho^{(p+1)/m_0}}\|\hat{\psi}\|_1^{-(p+1)p/m_0}$.

If the results just obtained for $\hat{\psi}$ are translated into results about $\hat{\phi}$, it appears that if $\phi(x-ct)$ defines a solitary-wave solution of (1.3), then

$$\sup_{\xi\in\mathbb{R}}e^{\mu|\xi|}|\hat{\phi}(\xi)|<\infty \tag{3.17}$$

for any μ satisfying

$$0 < \mu < \frac{m(c-1)^{p/m}(p+1)^{(p+1)/m}(2\pi)^{(p+1)p/2m}}{ep} \|\hat{\phi}\|_1^{\frac{-(p+1)p}{m}} = \rho_1(m,c,\phi)$$

when $1 \le \lfloor m \rfloor = m$, or for any μ satisfying

$$0 < \mu < \frac{m_0(c-1)^{p/m_0}(p+1)^{(p+1)/m_0}(2\pi)^{(p+1)p/2m_0}}{ep\varrho^{(p+1)/m_0}} \|\hat{\phi}\|_1^{\frac{-(p+1)p}{m}} = \rho_1(m,c,\phi)$$

when $1 \le m_0 = \lfloor m \rfloor < m$.

On the other hand, if $\phi(x-ct)$ defines a solitary-wave solution of (1.4), then (3.17) holds for this ϕ for any μ with

$$0 < \mu < \frac{mc^{1/m}(c-1)^{p/m}(p+1)^{(p+1)/m}(2\pi)^{(p+1)p/2m}}{ep} \|\hat{\phi}\|_1^{\frac{-(p+1)p}{m}} = \rho_2(m,c,\phi)$$

when $1 \le \lfloor m \rfloor = m$, or for any μ with

$$0 < \mu < \frac{m_0c^{1/m_0}(c-1)^{p/m_0}(p+1)^{(p+1)/m_0}(2\pi)^{(p+1)p/2m_0}}{ep\varrho^{(p+1)/m_0}} \|\hat{\phi}\|_1^{\frac{-(p+1)p}{m_0}} = \rho_2(m,c,\phi)$$

when $1 \le m_0 = \lfloor m \rfloor < m$.

The theorem is thus seen to be valid if one chooses $\sigma = \rho_1$ for solutions of (1.3) and $\sigma = \rho_2$ for solutions of (1.4). □

An inspection of the proof presented above shows that the specific assumption $\alpha(\xi) = |\xi|^m$ is not needed. Indeed, the presumption that there are positive constants $A > 0$ and $m \ge 1$ such that

$$A|\xi|^m \le \alpha(\xi) \tag{3.18}$$

for all $\xi \in \mathbb{R}$ suffices for our theory. The lower bound in (3.18) implies that the normalized Fourier transform $\hat{\psi}$ satisfies

$$|\hat{\psi}(\xi)| = \frac{1}{1+\lambda\alpha(\xi)}\left|\mathcal{V}_{p+1}\hat{\psi}(\xi)\right| \le \frac{1}{1+\lambda A|\xi|^m}\left|\mathcal{V}_{p+1}\hat{\psi}(\xi)\right|, \tag{3.19}$$

and it is this inequality that is the basis for the estimates appearing in the proof of Theorem 2. In consequence of these remarks, we can assert the following corollary to the proof of Theorem 2.

COROLLARY 3. *Let $u(x,t) = \phi(x-ct)$ be a solitary-wave solution of the equation*

$$u_t + u_x + u^p u_x - (Mu)_x = 0$$

or the equation

$$u_t + u_x + u^p u_x + (Mu)_t = 0,$$

where $p \ge 1$ is an integer and $\widehat{M\phi}(\xi) = \alpha(\xi)\hat{\phi}(\xi)$ with $\alpha(\xi)$ satisfying (3.18) *for some $m \ge 1$ and $A > 0$. Then there exists a constant $\sigma > 0$ such that*

$$\sup_{\xi\in\mathbb{R}} e^{\mu|\xi|}|\hat{\phi}(\xi)| < \infty$$

for any μ with $0 < \mu < \sigma$.

The result concerning analyticity of ϕ now follows immediately from Theorem 2 or Corollary 3 together with the Paley–Wiener theorem.

THEOREM 4. *Let ϕ satisfy the assumptions of Corollary* 3 *and let $\sigma > 0$ be as in the conclusion of this corollary. Then there is a function $\Phi(z)$ defined and holomorphic on the open strip $\{z \in \mathbb{C} : |\Im z| < \sigma\}$ such that $\Phi(x) = \phi(x)$ for all $x \in \mathbb{R}$.*

Proof. Let μ lie in the open interval $(0, \sigma)$. Choose a $\mu_1 > 0$ satisfying $0 < \mu < \mu_1 < \sigma$. Then it follows that

$$\begin{aligned}\int_{-\infty}^{\infty} e^{2\mu|\xi|}|\hat{\phi}(\xi)|^2 d\xi &= \int_{-\infty}^{\infty} e^{-2(\mu_1-\mu)|\xi|}e^{2\mu_1|\xi|}|\hat{\phi}(\xi)|^2 d\xi \\ &\leq \sup_{\xi\in\mathbb{R}} \left(e^{\mu_1|\xi|}|\hat{\phi}(\xi)|\right)^2 \int_{-\infty}^{\infty} e^{-2(\mu_1-\mu)|\xi|} d\xi < \infty.\end{aligned} \tag{3.20}$$

Define the function

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{\phi}(\xi)e^{-iz\xi} d\xi$$

for any $z = x + iy \in \Omega = \{z \in \mathbb{C};\ |\Im z| < \sigma\}$. Using (3.20) and the Paley–Wiener theorem [10], one may conclude that $\Phi(z)$ is a well-defined, analytic function on Ω. Of course, Plancherel's theorem implies that $\Phi(x) = \phi(x)$ for any $x \in \mathbb{R}$. □

An immediate consequence of the analyticity expressed in Theorem 4 is the following interesting result.

COROLLARY 5. *Suppose the hypotheses of Corollary* 3 *to hold and let ϕ be a solitary-wave solution of* (1.3) *or* (1.4). *Then ϕ cannot have compact support, nor can it be the case that in any bounded set $S \subset \mathbb{R}$, there are more than a finite number of points $x_\nu \in S$ such that $\phi(x_\nu) = \nu$. In particular, ϕ has at most finitely many zeros in any bounded subset of* $\mathbb{R}$.

Remarks. It is worth contrasting the last result with that obtained for the evolution equation

$$u_t + u^p u_x + (u^q)_{xxx} = 0, \tag{3.21}$$

where $q > 1$ is an integer. In equation (3.21), the dispersive term is singular, and this fact accounts for the compactly supported travelling-wave solutions (compactons) discovered recently by Rosenau (see Hyman and Rosenau [7]). As Corollary 5 shows, such solutions are not possible when the dispersion is nonsingular.

In case the symbol $\alpha(\xi) = |\xi|^m$, where m is an even integer, one may establish the analyticity of ϕ by recourse to the local theory of ordinary differential equations. It is not immediately transparent even in this case that the real analyticity thereby established extends to analyticity in a complex strip. However, a little work in this context reveals the truth of this assertion. These methods make no impression in case the symbol α does not generate a local operator.

4. Further extensions. It is the purpose of this short section to expand the range of the discussion to include equations of Schrödinger type as depicted in (1.5). In (1.5), it is supposed that M is a dispersion operator with the symbol α as in (1.6) and that $p = 2r$ is an even natural number.

The travelling-wave solutions of (1.5) of interest here have the general form $e^{i\omega t}\psi_{\omega,\theta}(x - \theta t)$, where ω and θ are real numbers with $0 \leq \omega \leq 2\pi$, say, and with $\psi : \mathbb{R} \to \mathbb{C}$ a smooth function lying in $L_1 \cap L_\infty$. Of special interest are the so-called *bound states.* These are standing-wave solutions of (1.5) for which $\theta = 0$, $\omega = \Omega > 0$, and $|\psi|$ tends rapidly to zero at infinity. The function $\psi_{\omega,0} = \phi_\Omega$ defining a bound state $e^{i\Omega t}\phi_\Omega(x)$ satisfies the equation

$$\Omega\phi + M\phi - |\phi|^{2r}\phi = 0. \tag{4.1}$$

In the applications associated to Schrödinger equations, particular importance is attached to *ground states,* which are bound states that minimize energy subject to fixed charge. The associated waveforms ϕ_Ω are analogous to solitary-wave solutions of (1.3) and (1.4) in that they are real valued, even, and rapidly decreasing to zero at infinity. Such solutions fall under the auspices of our previous theory.

THEOREM 6. *Let $\Omega > 0$ and let ϕ_Ω be a ground-state solution of* (1.5) *that lies in $L_1 \cap L_2$. Suppose $p = 2r$, where r is a positive integer, and suppose the symbol α of M to satisfy* (3.18). *Then there exist a $\sigma > 0$ and a function Φ_Ω defined and analytic on the strip $\{z = x + iy : |y| < \sigma\}$ such that $\Phi_\Omega(x) = \phi_\Omega(x)$ for all $x \in \mathbb{R}$.*

The range of applicability of this result may be considerably broadened if the dispersion operator is suitably specialized. Consider, for example, the special case where $\alpha(k) = k^2$, corresponding to the one-dimensional equation

$$iu_t + u_{xx} + |u|^{2r}u = 0, \tag{4.2}$$

with $r = 1$ corresponding to the classical cubic Schrödinger equation. In this case, we have the following simple lemma (cf. Bona and Soyeur [4]) relating bound states to more general travelling-wave solutions. Define $T_\theta : H^1 \to H^1$ by

$$(T_\theta u)(x) = e^{i\frac{1}{2}\theta x}u(x). \tag{4.3}$$

LEMMA 7. *Let ϕ be an H^1-function and let $\psi = T_\theta\phi$ for some $\theta \in \mathbb{R}$. Then ϕ defines a bound state of* (4.2) *corresponding to the parameter $\Omega = \omega - \frac{1}{4}\theta^2 > 0$ if and only if $\psi = \psi_{\omega,\theta}$ defines a travelling-wave solution of* (4.2).

Suppose that $e^{i\omega t}\psi_{\omega,\theta}(x - \theta t)$ is a travelling-wave solution of (4.1) corresponding to a bound state $e^{i\Omega t}\phi_\Omega(x)$ under the transformation in (4.3). Suppose also that ϕ_Ω is actually a ground state. Then according to Theorem 6, ϕ_Ω is the restriction to the real axis of a function Φ_Ω that is analytic in a strip $\{z : |\Im(z)| < \sigma\}$. It follows that $\psi_{\omega,\theta}$ is likewise the restriction to the real axis of a function $\Psi_{\omega,\theta}$ analytic in the same strip, namely the function

$$\Psi_{\omega,\theta}(z) = e^{i\frac{1}{2}\theta z}\phi_\Omega(z).$$

While this result is a consequence of the general theory, such considerations are not required in this special case. Equation (4.1) for ϕ_Ω can be solved explicitly in case $M = -\partial_x^2$, and one readily finds that

$$\phi_\Omega(x) = A\,\mathrm{sech}^{1/r}(Bx),$$

where $A = \sqrt[2r]{(r+1)\Omega}$ and $B = r\sqrt{\Omega}$.

A more challenging situation arises when the symbol $\alpha(k)$ is a perturbation of the Laplacean. Suppose that $\psi_{\omega,\theta}$ defines a travelling-wave solution of (1.5) by the formula

$$u(x,t) = e^{i\omega t}\psi_{\omega,\theta}(x - \theta t).$$

Then $\psi_{\omega,\theta}$ satisfies the equation

$$-\omega\psi - i\theta\psi' - M\psi + |\psi|^{2r}\psi = 0.$$

Guided by the considerations that arose when $M = -\partial_x^2$ in (4.2) and (4.3), we write $\psi_{\omega,\theta}(x) = e^{i\frac{1}{2}\theta x}\phi(x)$. A computation shows ϕ to satisfy the equation

$$(-\omega + \theta^2/2)\phi - i\theta\phi' - \widetilde{M}\phi + |\phi|^{2r}\phi = 0, \tag{4.4}$$

where the symbol $\tilde{\alpha}$ of the operator $\widetilde{M}$ is given by

$$\tilde{\alpha}(\xi) = \alpha(\xi - \theta/2).$$

Assuming that ϕ is real valued, equation (4.4) takes the form

$$\widetilde{M}\phi + i\theta\phi' + (\omega - \theta^2/2)\phi = \phi^{2r+1}, \tag{4.5}$$

or, in Fourier-transformed variables,

$$\left[\alpha(\xi - \theta/2) + \xi\theta + \omega - \theta^2/2\right] \hat{\phi}(\xi) = \widehat{\phi^{2r+1}}(\xi). \tag{4.6}$$

Write $\alpha(\xi) = \xi^2 + \beta(\xi)$, where $\beta(\xi) \geq c|\xi|^m$ for some constants $m \geq 0$ and $c \geq 0$. Then the symbol on the left-hand side of (4.6) may be written as

$$\xi^2 + \beta(\xi - \theta/2) + \omega - \frac{1}{4}\theta^2 \geq \xi^2 + c|\xi - \theta/2|^m + \Omega,$$

where $\Omega = \omega - \frac{1}{4}\theta^2$ as before. If $\Omega > 0$, then obviously we have

$$\xi^2 + \beta(\xi - \theta/2) + \Omega \geq A_1 + A_2|\xi|^2 \tag{4.7}$$

for suitably chosen positive constants A_1 and A_2. Because of (4.7), the theory developed in §3 may be brought to bear, and we ascertain immediately that ϕ has an analytic extension Φ to a strip in $\mathbb{C}$ centered about the real axis. In consequence of the relationship between ϕ and ψ, the same conclusion is drawn about ψ. This result is summarized in our last proposition.

PROPOSITION 8. *Suppose the symbol α of the dispersion operator M to have the form $\alpha(\xi) = \xi^2 + \beta(\xi)$, where $\beta(\xi) \geq c|\xi|^m$ for some $c \geq 0$ and $m \geq 0$. Let $u(x,t) = e^{i\omega t}\psi_{\omega,\theta}(x - \theta t)$ be a travelling-wave solution of* (1.5), *where $\omega - \frac{1}{4}\theta^2 > 0$. Suppose $\psi_{\omega,\theta}(y) = e^{i\frac{1}{2}\theta y}\phi(y)$, where ϕ is real-valued. Then there is a $\sigma > 0$ and a function $\Psi_{\omega,\theta}$ analytic in the strip $\{z : |\Im(z)| < \sigma\}$ such that $\Psi_{\omega,\theta}(x) = \psi_{\omega,\theta}(x)$ for all $x \in \mathbb{R}$.*

Remark. Equation (4.1) arises in more than one space dimension in the form

$$iu_t - Mu + |u|^{2r}u = 0, \tag{4.8}$$

where $u = u(x_1, x_2, \ldots, x_n, t)$ and M is a Fourier multiplier operator defined by

$$\widehat{Mv}(\xi_1, \xi_2, \ldots, \xi_n) = \alpha(\xi_1, \xi_2, \ldots, \xi_n)\, \hat{v}(\xi_1, \xi_2, \ldots, \xi_n).$$

Travelling-wave solutions analogous to those considered in one dimension have the form $e^{i\omega t}\psi_{\omega,\theta}(x - \theta t)$ where $t, \omega \in \mathbb{R}$ and $x, \theta \in \mathbb{R}^n$. Bound states correspond to $\theta = 0$.

It follows readily from the techniques developed in §3 that a ground-state solution ϕ of (4.8) is the restriction to $\mathbb{R}^n$ of a function Φ which is defined in a "strip" $\{(z_1, z_2, \ldots, z_n) \in \mathbb{C}^n : |\Im(z_j)| < \sigma, \text{ for } 1 \le j \le n\}$ and comprises an analytic function of n complex variables there.

In case $M = -\Delta$, then the n-dimensional analog of Lemma 7 allows bound states to be related to general travelling waves via the operator T_θ given by $T_\theta w(x) = e^{i\frac{1}{2}\theta\cdot x}w(x)$ and thereby to extend the results on analyticity to more general travelling-wave solutions.

5. Conclusion. Solitary-wave solutions of the classes (1.3), (1.4), and (1.5) of nonlinear, dispersive wave equations have been shown to possess an analytic extension into a complex strip around their original domain of definition. This further regularity property of such travelling-wave solutions lays the groundwork for a broader use of complex-variable methods in the study of these equations. Such techniques have already proven to be useful in discussing a number of thorny problems connected with uniqueness and soliton behavior (cf. [1], [5], [6], [8], [9]). Perhaps the door now stands ajar to further developments along these lines.

An interesting project for further study would be to determine the type of singularities that arise when a solitary wave is extended into the complex plane. The examples in hand indicate that these extensions will be meromorphic or fractional powers of meromorphic functions. We have conjectured this to be the case under fairly general conditions, but a proof has remained elusive.

REFERENCES

[1] C. J. Amick and J. F. Toland, *Uniqueness of Benjamin's solitary-wave solutions of the Benjamin–Ono equation*, IMA J. Appl. Math., 46 (1991), pp. 21–28.

[2] T. B. Benjamin, *Lectures on Nonlinear Wave Motion*, in Lectures in Applied Mathematics 15, American Mathematical Society, Providence, RI, 1974, pp. 3–47.

[3] T. B. Benjamin, J. L. Bona, and D. K. Bose, *Solitary-wave solutions of nonlinear problems*, Philos. Trans. Roy. Soc. London Ser. A, 331 (1990), pp. 195–244.

[4] J. L. Bona and A. Soyeur, *On the stability of solitary-wave solutions of model equations for long waves*, J. Nonlinear Sci., 4 (1994), pp. 449–470.

[5] G. Bowtell and A. E. G. Stuart, *A particle representation for Korteweg–de Vries solitons*, J. Math. Phys., 24 (1983), pp. 969–981.

[6] A. C. Bryan and A. E. G. Stuart, *On the nonexistence of soliton solutions of the regularized long-wave equation*, preprint.

[7] J. M. Hyman and P. Rosenau, *Compactons: Solitons with finite wavelength*, Phys. Rev. Lett., 70 (1993), pp. 564–567.

[8] M. D. Kruskal, *The Korteweg–de Vries Equation and Related Evolution Equations*, in Lectures in Applied Mathematics 15, American Mathematical Society, Providence, RI, 1974, pp. 61–83.

[9] Y. Li, *Uniqueness and analyticity of solitary waves*, Ph.D. thesis, Pennsylvania State University, University Park, PA, 1994.

[10] R. E. A. C. Paley and N. Wiener, *Fourier Transforms in the Complex Domain*, American Mathematical Society, Providence, RI, 1934.

[11] J. Riordan, *Combinatorial Identities*, John Wiley & Sons, New York, 1968.

[12] M. Weinstein, *Existence and dynamic stability of solitary wave solutions of equations arising in long wave propagation*, Comm. Partial Differential Equations, 12 (1987), pp. 1133–1173.

[13] G. B. Whitham, *Linear and Nonlinear Waves*, John Wiley & Sons, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 738–764, May 1996

THE LINEARIZATION OF THE INITIAL-BOUNDARY VALUE PROBLEM OF THE NONLINEAR SCHRÖDINGER EQUATION*

A. S. FOKAS† AND A. R. ITS‡

Abstract. We consider the nonlinear Schrödinger (NLS) equation in the variable $q(x,t)$ with both x and t in $[0,\infty)$. We assume that $q(x,0)=u(x)$ and $q(0,t)=v(t)$ are given, that $u(0)=v(0)$, and that $u(x)$ *and* $v(t)$ as well as their first two derivatives belong to $L_1 \cap L_2(\mathbb{R}^+)$. We show that the solution of this initial-boundary value problem can be reduced to solving a Riemann–Hilbert (RH) problem in the complex k-plane with jumps on $Im(k^2)=0$. This RH problem is equivalent to a linear integral equation which has a unique global solution. This linear integral equation is uniquely defined in terms of certain functions (scattering data) $b(k)$ and $c(k)$. The function $b(k)$ can be effectively computed in terms of $u(x)$. However, although the analytic properties of $c(k)$ are completely determined, the relationship between $c(k)$, $u(x)$ and $v(t)$ is highly nonlinear. In spite of this difficulty, we can give an effective description of the asymptotic behavior of $q(x,t)$ for large t. In particular, we show that as $t \to \infty$, solitons are generated moving away from the boundary. In addition, our formalism can be used to generate effectively pairs of functions $q(0,t)$ and $q_x(0,t)$ compatible with a given $q(x,0)$ as well as to determine the associated $q(x,t)$. It is important to emphasize that the analysis of this problem, in addition to techniques of exact integrability, requires the essential use of general partial differential equations (PDE) techniques.

Key words. nonlinear Schrödinger equation

AMS subject classifications. Primary 35Q55; Secondary 35Q15

1. Introduction.

For integrable equations, a method exists for solving the initial-value problem on the infinite line for decaying initial data. For evolution equations in one spatial variable, this method reduces the solution of the Cauchy problem to the formulation of a certain classical mathematical problem called the Riemann–Hilbert (RH) problem. A RH problem can be solved via a linear integral equation. A distinguished property of integrable equations is that they can be written as the compatibility condition of a pair of linear eigenvalue equations, called the Lax pair [1]. The associated RH problem is essentially determined by the x-part of the Lax pair; the t-part of the Lax pair plays only an auxiliary role. In the case of the nonlinear Schrödinger (NLS) equation, the relevant RH problem is formulated in the complex k-plane with a jump on $Im(k)=0$.

Many physical problems are formulated as initial-boundary value problems. For example, such a problem arises in the modeling of certain ionospheric experiments when one directs a radio frequency wave at the ionosphere. At the reflection point of the wave, a sufficient level of electron plasma waves is excited and nonlinearity becomes important. This problem gives rise to the NLS equation with $x, t\epsilon[0,\infty)$ [2]. Furthermore, several other physical problems can be reduced to initial-boundary value problems. For example, such a problem arises in connection with optical switches [3] and can also be modeled by a NLS. The occurrence of the Korteweg–de Vries (KdV) equation on the quarter-plane is discussed in [21]–[26].

*Received by the editors May 24, 1993, accepted for publication (in revised form) September 1, 1994. This research was partially supported by National Science Foundation grants DMS-9111611 and DMS-9204075 and Air Force Office of Scientific Research grant F49620-93-1-008-DEP.

†Department of Mathematics and Computer Science and Institute for Nonlinear Studies, Clarkson University, Potsdam, NY 13699-5815 and Department of Mathematical Sciences, Loughborough University, Loughborough LE11 STU, United Kingdom.

‡Department of Mathematical Sciences, Indiana University–Purdue University Indiana, 402 North Blackford Street, Indianapolis, IN 46202-3216.

Recently, a new formalism has been developed [4], [5] for studying initial-boundary value problems on the half-infinite line for decaying initial and boundary data. This formalism also reduces the solution of the initial-boundary value problem to the solution of a single RH problem. However, for the formulation of this RH problem, both the x- and the t-parts of the Lax pair play an important role. Actually, it is the t-part which determines where, in the complex k-plane, the jumps occur. In the case of the NLS, the jumps occur on $Im(k^2) = 0$, which is a reflection of the fact that the t-part of the Lax pair contains k^2, which in turn is a consequence of the fact that the NLS involves a second derivative in x. Here we study the initial-boundary value problem for the NLS equation in detail. We show that the analysis of this problem also requires, in addition to techniques from exact integrability, the essential use of more general PDE techniques. In the cases studied so far, the exact methods could be used to establish existence of global solutions as well as to study the properties of these solutions. In contrast, in the problem studied here, exact methods are used only to study the properties of solutions. It shows that a hybrid between exact methods and general PDE techniques can provide a powerful approach for analyzing problems of mathematical and physical significance. We expect that a wide class of problems can be analyzed in a similar manner.

We consider the NLS equation

$$iq_t + q_{xx} - 2\lambda|q|^2 q = 0, \quad x, t \in [0, \infty), \quad \lambda = \pm 1, \tag{1.1}$$

where $q(x,0) = u(x)$ and $q(0,t) = v(t)$ are given. We assume that

$$\begin{gathered} u(x) \in H_2(\mathbb{R}^+), \qquad v(t) \in C_2(\mathbb{R}^+), \qquad u(0) = v(0), \\ xu(x) \;\text{ and }\; x^2u(x) \in L_2(\mathbb{R}^+), v(t) \in L_1 \cap L_2(\mathbb{R}^+), \\ v(t) \in L_1 \cap L_2(\mathbb{R}^+), \qquad v'(t),\;\; tv(t),\;\; tv'(t),\;\; tv''(t) \in L_1(\mathbb{R}^+), \end{gathered} \tag{1.2}$$

where H_2 denotes that a function and its first two derivatives belong to L_2, C_2 denotes that a function is twice differentiable, and prime denotes differentiation.

The cases $\lambda = 1$ and $\lambda = -1$ are usually referred to as the defocusing and focusing cases, respectively. Equation (1.1) is the compatibility condition of the following Lax pair for the 2×2 matrix $w(x,t,k)$ [6]:

$$w_x + ik\sigma_3 w = Qw, \tag{1.3a}$$

$$w_t + Uw = wC(t), \qquad U(x,t,k) \div 2ik^2\sigma_3 + i\lambda|q|^2\sigma_3 - 2kQ + iQ_x\sigma_3, \tag{1.3b}$$

where $\sigma_3 = \text{diag}(1,-1)$, the 2×2 matrix $C(t)$ is an arbitrary function of t, and $Q(x,t)$ is an off-diagonal matrix with 12 and 21 entries given by q and $\lambda\bar{q}$, respectively.

We have developed the following linearization scheme for the solution of the initial-boundary value problem of the NLS. Given $q(x,0)$, construct s_1^+ and s_2^+ by $s_1^+(k) \div \psi_1(0,k)$, $s_2^+(k) \div \psi_2(0,k)$, where $(\psi_1(x,k), \psi_2(x,k))^T$ is the solution of (1.3a) with $q(x,t)$ replaced by $q(x,0)$, satisfying the boundary condition

$$\lim_{x \to \infty} [(\psi_1, \psi_2)^T \exp(-ikx)] = (0,1)^T.$$

Define $b(k)$ by $b = s_1^+/\overline{s_2^+}$. Let $c(k)$, $k \in \mathbb{R}^- \cup i\mathbb{R}^+$ be the boundary value of a function meromorphic for $k \in$ II (I, II, III, and IV denote the first, second, third, and fourth

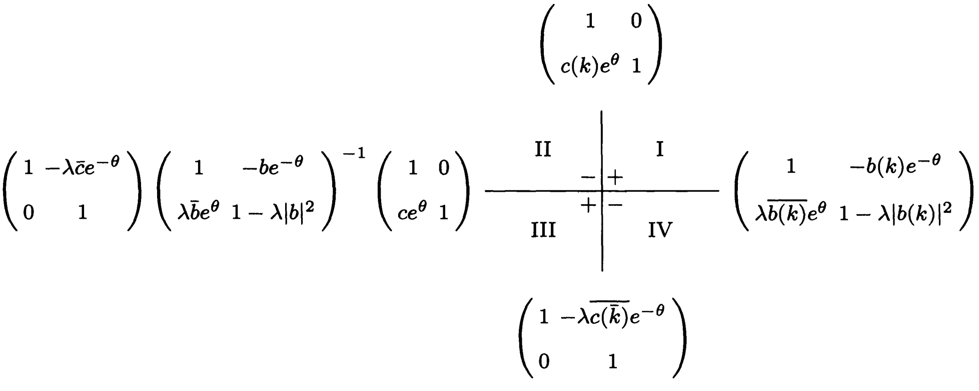

$\begin{pmatrix} 1 & 0 \\ c(k)e^{\theta} & 1 \end{pmatrix}$
II
I
$-$
$+$
$+$
$-$
III
IV
$\begin{pmatrix} 1 & -\lambda\bar{c}e^{-\theta} \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & -be^{-\theta} \\ \lambda\bar{b}e^{\theta} & 1-\lambda|b|^2 \end{pmatrix}^{-1}\begin{pmatrix} 1 & 0 \\ ce^{\theta} & 1 \end{pmatrix}$
$\begin{pmatrix} 1 & -b(k)e^{-\theta} \\ \lambda\overline{b(\bar{k})}e^{\theta} & 1-\lambda|b(k)|^2 \end{pmatrix}$
$\begin{pmatrix} 1 & -\lambda\overline{c(\bar{k})}e^{-\theta} \\ 0 & 1 \end{pmatrix}$

FIG. 1.1. *The RH problem associated with the initial-boundary value problem of NLS. The x, t dependence enters only through $\theta(x,t) = 2i(kx + 2k^2t)$.*

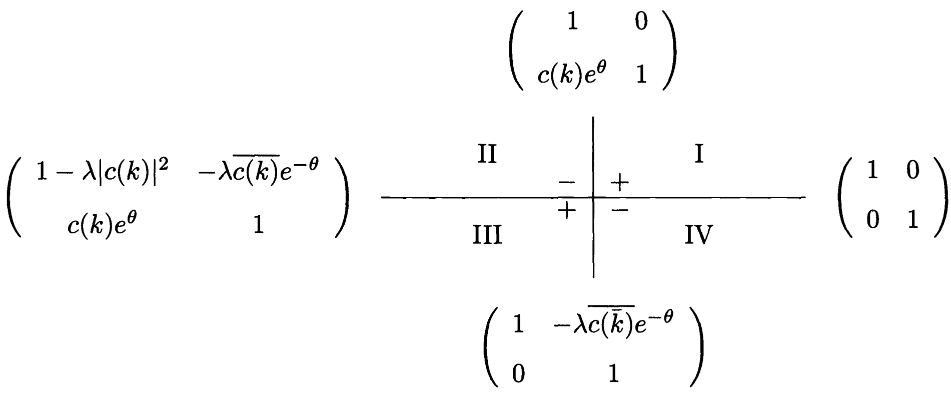

$\begin{pmatrix} 1 & 0 \\ c(k)e^{\theta} & 1 \end{pmatrix}$
II
I
$-$
$+$
$+$
$-$
III
IV
$\begin{pmatrix} 1-\lambda|c(k)|^2 & -\lambda\overline{c(\bar{k})}e^{-\theta} \\ c(k)e^{\theta} & 1 \end{pmatrix}$
$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
$\begin{pmatrix} 1 & -\lambda\overline{c(\bar{k})}e^{-\theta} \\ 0 & 1 \end{pmatrix}$

FIG. 1.2. *The RH problem for the focusing NLS with of $q(x,0) = 0$. The x, t dependence enters only through $\theta(x,t) = 2i(kx + 2k^2t)$.*

quadrants of the complex k-plane) with poles at the zeros of $s_2^+(k)$ and at the points $\{k_j\}_1^N$, $k_j \in$ II which are assumed to be different than the zeros of $s_2^+(k)$ (generic case); let c_j denote the residues of $c(k)$ at k_j; also, $c(k) \to 0$ as $k \to \infty$. Have $s_1^+(k)$, $s_2^+(k)$, and $c(k)$ solve an RH problem for a 2×2 meromorphic function $\hat{Z}_p$ with possible poles only at $\{k_j\}_1^N$. This RH problem is depicted in Figure 1.1. Finally, determine $q(x,t)$ by $q(x,t) = 2i\lim_{k\to\infty}(k\hat{Z}_p(x,t,k))_{12}$, $k \in$ I, where the subscript 12 denotes the 12 components of the matrix $\hat{Z}_p$. The points $\{k_j\}_1^N$ (which have the meaning of the discrete spectrum of equation (1.3b) evaluated at $x = 0$ and supplemented with the boundary condition $w_1 s_2^+(k) - w_2 s_1^+(k) = 0$ at $t = 0$) give rise to solitons which always move away from the boundary.

In the case of $q(x,0) = 0$, the RH problem reduces to the one depicted in Figure 1.2; it is specified by the boundary values of $c(k)$ on $k \in \mathbb{R}^- \cup i\mathbb{R}^+$, by $\{k_j\}_1^N$, and by $\{c_j\}_1^N$.

Unfortunately, although we have a complete characterization of the analytic properties of $c(k)$, we have not found an effective way of computing $c(k)$ in terms of $q(0,t)$ and $q(x,0)$. (For more details see the discussion below.) In spite of this fact, we can

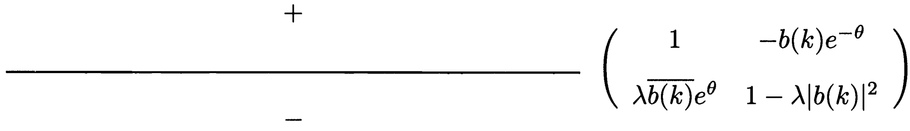

+
−
$\begin{pmatrix} 1 & -b(k)e^{-\theta} \\ \lambda\overline{b(k)}e^{\theta} & 1-\lambda|b(k)|^2 \end{pmatrix}.$

FIG. 1.3.

give an effective description of the long-time behavior ($\lambda = -1$):

$$q(x,t) = -2\eta_j \frac{\exp[-2i\xi_j x - 4i(\xi_j^2 - \eta_j^2)t - i\varphi_j]}{\cosh[2\eta_j(x + 4\xi_j t) - \Delta_j]} + 0(t^{-\frac{1}{2}}),$$

(1.4)

$$t \to \infty, -\frac{x}{4t} = \xi_j + 0\left(\frac{1}{t}\right), \quad j = 1, \ldots, N,$$

where

(1.5a) $$\eta_j = Im(k_j), \quad \xi_j = Re(k_j),$$

(1.5b)
$$\begin{aligned} \varphi_j = &-\frac{\pi}{2} + \arg c_j \\ &+ \sum_{l=1_{l\neq j}}^{N} [\text{sign}(\xi_l - \xi_j) - 1] \arg\left(\frac{k_j - k_l}{k_j - \bar{k}_l}\right) \\ &+ \frac{1}{\pi}\int_{-\infty}^{-x/4t} \frac{\log[1 + |b(\mu) + \overline{c(\mu)}|^2]}{(\mu - \xi_j)^2 + \eta_j^2}(\mu - \xi_j)d\mu, \end{aligned}$$

(1.5c)
$$\begin{aligned} \Delta_j = &-\log 2\eta_j + \log|c_j| \\ &+ \sum_{l=1_{l\neq j}}^{N} [\text{sign}(\xi_l - \xi_j) - 1] \log\left|\frac{k_j - k_l}{k_j - \bar{k}_l}\right| - \frac{\eta_j}{\pi}\int_{-\infty}^{-x/4t} \frac{\log[1 + |b(\mu) + \overline{c(\mu)}|^2]}{(\mu - \xi_j)^2 + \eta_j^2}d\mu. \end{aligned}$$

All $k_j \in$ II, thus all $\xi_j < 0$ and the solitons move away from the boundary. The summation terms in the above equations describe the interaction among solitons, while the integration terms describe the interaction between solitons and the dispersive part.

Discussion. The linearization scheme developed in this paper can be summarized as follows: Given $q(x, 0)$, construct $b(k)$. Then, if $c(k)$ is *any* suitably decaying function meromorphic for $k \in$ II, the solution of the RH problem of Figure 1.1 generates the solution $q(x,t)$ corresponding to initial data $q(x, 0)$ and *some* boundary data $q(0, t)$. The main limitation of our result is that for given $q(x, 0)$ and $q(0, t)$, we cannot construct $c(k)$ by solving a linear problem. Nevertheless, we claim that for *any* given $q(x, 0)$ and $q(0, t)$ satisfying (1.2), the corresponding function $c(k)$ exists; in other words, the RH problem of Figure 1.1 solves the initial-boundary value problem (1.1) for *general* initial-boundary data.

The above RH problem is quite natural. To appreciate this we first recall the RH problem that corresponds to the NLS with $x \in (-\infty, \infty)$. This RH problem is depicted in Figure 1.3. Comparing the RH problems of Figure 1.1 and Figure 1.3, we see that the jumps for $k \in \mathbb{R}^+$ are identical. The jump for $k \in i\mathbb{R}^+$ cannot have a nonzero entry in the 12 position since $e^{-\theta}$ is unbounded for $k \in i\mathbb{R}^+$. The jump for $k \in i\mathbb{R}^-$ follows by symmetry considerations. Finally, the jump for $k \in \mathbb{R}^-$ follows from the cyclic condition that the product of the jump matrices equals unity (this is a reflection of continuity at $k = 0$). The fact that $c(k)$ has analytic continuation for $k \in \mathrm{II}$ can also be easily understood. At $t = 0$, the RH problem of Figure 1.1 must be reduced to the one that defines $q(x, 0)$. At $t = 0$, the term e^{θ} has analytic continuation in II. Thus the jumps along the imaginary axis can be mapped to a jump on the negative real axis. In this way, at $t = 0$ one finds the RH problem of Figure 1.3 with θ replaced by $2ikx$. This RH problem corresponds precisely to $q(x, 0)$.

The fact that $q(x, 0)$, $q(0, t)$ and $c(k)$ are related in a nonlinear way is a reflection of the fact that $q_x(0, t)$ depends nonlinearly on $q(x, 0)$ and $q(0, t)$. To appreciate this, we first recall the solution of the linearized problem

$$iq_t + q_{xx} = 0, \quad x, t \in [0, \infty), \tag{1.6}$$

where $q(0, t)$ and $q(x, 0)$ are given and decaying for large t and large x. This problem can be solved by the sine transform. However, in order to draw comparisons with the nonlinear problem, we shall use a Fourier transform,

$$\hat{q}(k, t) = \int_0^\infty dx e^{ikx} q(x, t). \tag{1.7}$$

The evolution of the Fourier data $\hat{q}(k, t)$ is given by

$$\hat{q}_t + ik^2 \hat{q} = iq_x(0, t) + kq(0, t). \tag{1.8}$$

In equation (1.8), $q(0, t)$ is known but $q_x(0, t)$ is unknown (the sine transform is precisely used in order to eliminate $q_x(0, t)$). This apparently ominous situation can be bypassed by using the fact that the solution $\hat{q}(k, t)$ of (1.8) is analytic in the upper half of the k complex plane. It turns out that this requirement implies

$$\hat{q}(k, 0) = -\int_0^\infty dt e^{ik^2 t}(iq_x(0, t) + kq(0, t)). \tag{1.9}$$

Given $q(x, 0)$ and $q(0, t)$ and using the substitution $k = e^{i\pi/4}\sqrt{\rho}$, $\rho > 0$, equation (1.9) yields $q_x(0, t)$. It is important to notice that if $q_x(0, t)$ and $q(0, t)$ are arbitrary functions, then the rhs of equation (1.9) will be analytic for $k \in \mathrm{I} \cup \mathrm{III}$. However, in order for $q_x(0, t)$ and $q(0, t)$ to be the boundary values of the solution of equation (1.6), it is necessary and sufficient that the rhs of equation (1.9) has analytic continuation across the positive imaginary k axis ($\hat{q}(k, 0)$ is analytic for $k \in \mathrm{I} \cup \mathrm{II}$).

Before discussing the nonlinear problem, we emphasize that using these type of analyticity arguments, it is possible to solve linear equations for which the standard spectral theory fails. An example of such an equation is the linearized KdV $u_t + u_{xxx} = 0$, for which there does not exist a proper generalization of the sine transform.

We now discuss the nonlinear problem. Let $(\hat{\psi}_1(t, k), \hat{\psi}_2(t, k))^T$ be the solution of the vector equation

$$\hat{\psi}_t + (2ik^2\sigma_3 + i\lambda|q(0, t)|^2\sigma_3 - 2kQ(0, t) + iQ_x(0, t)\sigma_3)\hat{\psi} = 0 \tag{1.10}$$

satisfying the boundary condition $\lim_{t\to\infty}[(\hat{\psi}_1, \hat{\psi}_2)\exp(-2ik^2 t)] = (0, 1)^T$. Let $r(k) \doteqdot \hat{\psi}_1(0, k)/\hat{\psi}_2(0, k)$. It turns out that for arbitrary decaying functions $q(0, t)$ and $q_x(0, t)$,

$r(k)$ is a meromorphic function for $k \in \mathrm{I} \cup \mathrm{III}$ and $r(k) \to 0$ as $k \to \infty$. However, if $q(0,t)$ and $q_x(0,t)$ are the boundary values of the NLS, then in addition $r(k)$ satisfies

$$r(k) = \frac{s_1^+(k)}{s_2^+(k)}, \quad k \in \mathrm{I}, \tag{1.11}$$

where s_1^+ and s_2^+ are determined from $q(x,0)$ and are analytic for $k \in \mathrm{I} \cup \mathrm{II}$. Equation (1.11) is the analogue of equation (1.9). It shows that although the relationship between $q(x,0)$, $q(0,t)$, and $q_x(0,t)$ is highly nonlinear, its reflection in the k-plane (scattering space) is rather simple: $r(k)$ has analytic continuation across the positive imaginary k axis.

There exists an invertible correspondence between the "potential" $\{q(0,t), q_x(0,t)\}$ and the scattering data $r(k)$: Given $q(0,t)$ and $q_x(0,t)$, equation (1.10) implies $r(k)$. Conversely given a meromorphic function $r(k)$, one can find $q(0,t)$ and $q_x(0,t)$ by solving a RH problem with the jump $r(k)$ for $k^2 \in \mathbb{R}$. This provides an effective way for deriving pairs of functions $q(0,t)$ and $q_x(0,t)$ *compatible* with a given $q(x,0)$. Indeed, given $q(x,0)$, one first computes $r(k)$ for $k \in \mathrm{I}$ from equation (1.11). Let $r(k)$, $k \in \mathrm{III}$ be any suitably decaying meromorphic function. Then the solution of the above RH problem yields $q(0,t)$ and $q_x(0,t)$.

Unfortunately, given $q(0,t)$ and $q(x,0)$, we cannot compute $r(k)$ for $k \in \mathrm{III}$ by solving a linear problem. This is a consequence of the fact that now we have a "mixed" problem where one gives "half" the potential, i.e., $q(0,t)$ and "half" the scattering data, i.e., $r(k)$, for $k \in \mathrm{I}$. It turns out that this problem can be formulated as a *nonlinear* RH problem and will be discussed elsewhere.

The study of the large-t behavior of $q(x,t)$ reduces to the study of the large-t behavior of the RH problem of Figure 1.1. Because the x, t dependence of this RH problem is rather simple, it is possible to give an effective asymptotic description of $q(x,t)$ as $t \to \infty$.

It was mentioned earlier that the analysis of equation (1.1) requires an essential use of general PDE techniques. This follows from the fact that in order to study the map between $\{q(0,t), q_x(0,t)\}$ and $r(k)$, one needs a priori estimates for $q_x(0,t)$. The uniqueness and existence of a global solution for the NLS on the quarter-plane is established in [13]. This result makes fundamental use of certain equations which are the analogues of the first three conserved quantities. Therefore, this theory uses L_2 estimates. However, the methods of exact integrability are based on L_1 estimates. It is therefore crucial for us to extend the results of [13] from L_2 to L_1. This poses significant technical difficulties. In this paper, following the ideas of [14], we show how to get L_1 estimates for $q(x,t)$ for all $t \geq 0$. The more difficult problem of obtaining L_1 estimates for $q_x(0,t)$ is discussed in [14].

There exist certain particular boundary conditions for which the difficulties discussed above disappear and the problem can be solved by exact methods. Such cases are discussed in [15]–[20].

It is also worth mentioning that the condition (1.11) is similar to the restrictions on the scattering data that appear in the boundary problem for the elliptic version of the sine-Gordon equation [27].

Notation. We will use the following notation: an overbar denotes complex conjugation; $a^+(k)$ denotes a function analytic for $k \in \mathbb{C}^+$ (upper half of the complex k-plane); $a^-(k)$ denotes a function analytic for $k \in \mathbb{C}^-$; $\hat{a}^+(k)$ denotes a function analytic for $k \in \mathrm{I}$ (first quadrant) $\cup$ III (third quadrant); $\hat{a}^-(k)$ denotes a function analytic for $k \in \mathrm{II}$ (second quadrant) $\cup$ IV (fourth quadrant); the subscript

p denotes meromorphicity instead of analyticity, for example $\hat{a}_p^+(k)$ is meromorphic for $k \in \mathrm{I} \cup \mathrm{III}$; if the matrix A is denoted by $A = (A^+, A^-)$, then $A^+ = (A_1^+, A_2^+)^T$ and $A^- = (A_1^-, A_2^-)^T$, denote the first and the second columns on the matrix A; $\mathrm{I} \doteq \mathrm{diag}(1,1)$.

Summary of results. In §2, we present the relevant formalism. (a) In §2.1, we analyze the x-part of the Lax pair, i.e., equation (1.3a). We define the 2×2 matrices $\varphi(x,t,k)$ and $\psi(x,t,k)$ as the solutions of equation (1.3a) specified by the boundary conditions $\varphi(0,t,k) = \mathrm{I}$, and $\lim_{x\to\infty} \psi(x,t,k)\exp(ikx\sigma_3) = \mathrm{I}$, respectively. We define the matrices $\Phi(x,t,k)$ and $\Psi(x,t,k)$ by the equations $\Phi(x,t,k) = \varphi(x,t,k)\exp(ikx\sigma_3)$ and $\Psi(x,t,k) = \psi(x,t,k)\exp(ikx\sigma_3)$. These functions are analytic in the complex k-plane cut along $Im(k) = 0$, they tend to the unit matrix as $k \to \infty$, and they have unit determinant. Actually, $\Psi = (\Psi^-, \Psi^+)$, and the matrix $\Phi = (\Phi^+, \Phi^-)$ is an entire function with respect to k. The eigenfunctions Φ and Ψ satisfy certain symmetry conditions involving complex conjugation. We define the 2×2 matrix χ_p by

$$\text{(1.12)} \qquad \begin{aligned} \chi_p^-(x,t,k) &= (\Psi^-(x,t,k), \Phi^-(x,t,k)/\overline{\psi_2^+(0,t,\bar{k})}),\\ \chi_p^+(x,t,k) &= (\Phi^+(x,t,k)/\psi_2^+(0,t,k), \Psi^+(x,t,k)), \end{aligned}$$

for $k \in \mathbb{C}^-$ and $k \in \mathbb{C}^+$, respectively. The matrix χ_p has unit determinant and satisfies a certain jump condition

$$\text{(1.13)} \qquad \chi_p^-(x,t,k) = \chi_p^+(x,t,k)F(x,t,k), \quad k \in \mathbb{R},$$

where $F(x,t,k)$ is a 2×2 matrix involving $\psi_1^+(0,t,k)$ and $\psi_2^+(0,t,k)$.

(b) In §2.2, we use the t-part of the Lax pair to analyze $\psi(0,t,k)$. This function satisfies equation (1.3b) evaluated at $x = 0$, where $C(t) = 2ik^2\sigma_3$. We define the 2×2 matrices $\hat{\Phi}(t,k)$ and $\hat{\Psi}(t,k)$ by the requirement that they satisfy the same equation as $\psi(0,t,k)$, and that $\hat{\Phi}(0,k) = \mathrm{I}$, $\lim_{t\to\infty} \hat{\Psi}(t,k) = \mathrm{I}$. These functions are analytic in the complex k-plane cut along $Im(k^2) = 0$, they tend to the unit matrix as $k \to \infty$, and they have unit determinant. Actually $\hat{\Psi} = (\hat{\Psi}^-, \hat{\Psi}^+)$, and the matrix $\hat{\Phi} = (\hat{\Phi}^+, \hat{\Phi}^-)$ is an entire function with respect to k. The eigenfunctions $\hat{\Phi}$ and $\hat{\Psi}$ satisfy the same symmetry condition as those satisfied by Φ and Ψ. We define the matrix $\hat{Y}_p(t,k)$ by

$$\text{(1.14)} \qquad \begin{gathered} \left(\frac{\hat{\Phi}^+(t,k)}{s_2^+(k)}, \psi^+(0,t,k)\right), \left(\frac{\hat{\Psi}^-(t,k)}{\rho(k)}, \psi^+(0,t,k)\right),\\ \left(\psi^-(0,t,k), \frac{\hat{\Psi}^+(t,k)}{\overline{\rho(\bar{k})}}\right), \left(\psi^-(0,t,k), \frac{\hat{\Phi}^-(t,k)}{\overline{s_2^+(\bar{k})}}\right) \end{gathered}$$

for $k \in \mathrm{I}, \ldots, \mathrm{IV}$, where the scalars $s_2^+(k)$ and $\rho(k)$ are defined by

$$\text{(1.15)} \qquad \begin{gathered} s_2^+(k) \doteq \psi_2^+(0,0,k), \quad k \in \mathbb{C}^+; \qquad \rho(k) \doteq \hat{\Psi}_1^-(0,k)s_2^+(k) - \hat{\Psi}_2^-(0,k)s_1^+(k), \quad k \in \mathrm{II};\\ s_1^+(k) \doteq \psi_1^+(0,0,k), \quad k \in \mathbb{C}^+. \end{gathered}$$

The matrix $\hat{Y}_p$ has unit determinant and satisfies a certain jump condition

$$\text{(1.16)} \qquad \hat{Y}_p^-(t,k) = \hat{Y}_p^+(t,k)e^{-2ik^2t\sigma_3}G(k)e^{2ik^2t\sigma_3}, \quad k^2 \in \mathbb{R},$$

where the 2 × 2 matrix $G(k)$ involves the functions $b(k)$ and $c(k)$, which are defined by

$$(1.17) \qquad b(k) \doteqdot \frac{s_1^+(k)}{s_2^+(k)}, \quad k \in \mathbb{R}; \qquad c(k) \doteqdot \frac{\hat{\Psi}_2^-(0,k)}{s_2^+(k)\rho(k)}, \quad k \in i\mathbb{R}^+ \cup \mathbb{R}^-.$$

(c) In §2.3, we formulate an RH problem whose solution gives the solution of the initial-boudnary value problem of the NLS equation. Let the 2 × 2 matrix $\hat{Z}_p(x,t,k)$ be defined by

(1.18)

$$\chi_p(x,t,k)\begin{pmatrix} 1 & 0 \\ \frac{(\hat{Y}_p)_{21}(t,k)}{(\hat{Y}_p)_{22}(t,k)}e^{2ikx} & 1 \end{pmatrix}, \ k \in \mathbb{C}^+; \ \chi_p(x,t,k)\begin{pmatrix} 1 & \frac{(\hat{Y}_p)_{12}(t,k)}{(\hat{Y}_p)_{11}(t,k)}e^{-2ikx} \\ 0 & 1 \end{pmatrix}, \ k \in \mathbb{C}^-.$$

Then the meromorphic function $\hat{Z}_p(x,t,k)$, which has unit determinant, satisfies the RH problem

(1.19)

$$\hat{Z}_p^-(x,t,k) = \hat{Z}_p^+(x,t,k)e^{-i(kx+2k^2t)\sigma_3}G(k)e^{i(kx+2k^2t)\sigma_3}, \quad k^2 \in \mathbb{R}, \ \hat{Z} \to I \text{ as } k \to \infty.$$

The jumps of this RH problem are given in Figure 1.1. The possible poles of $\hat{Z}_p(x,t,k)$ can occur only at the zeros of $\rho(k)$ for $k \in$ II and at the complex conjugates of these zeros. The additional conditions on the residues of $\hat{Z}_p$ have the usual "solitonic" form (equation (2.4.1) in §2.4).

In §3, we discuss rigorous aspects of the formalism developed in §2. This includes finding conditions on $q(x,0)$ and $q(0,t)$ which guarantee the following:

(i) The linear integral equations for $\Phi(x,t,k)$ and $\Psi(x,t,k)$ are uniquely solvable for fixed t and k.

(ii) The linear integral equations for $\hat{\Phi}(t,k)$ and $\hat{\Psi}(t,k)$ are uniquely solvable for fixed k.

(iii) The RH problem of Figure 1.1 is uniquely solvable.

The integral equations for $\Phi(x,t,k)$, $\Psi(x,t,k)$ and for $\hat{\Phi}(t,k)$, $\hat{\Psi}(t,k)$ are of the Volterra type; therefore, they possess global, bounded, continuous solutions provided that $q(x,t)$, with t fixed, and $\{q(0,t), q_x(0,t)\}$ belong in L_1. Furthermore, under certain additional L_2 conditions, $\Phi(0,t,k) - \mathrm{I}$, $\Psi(0,t,k) - \mathrm{I}$, $\hat{\Phi}(t,k) - \mathrm{I}$, and $\hat{\Psi}(t,k) - \mathrm{I} \in H_1$ for fixed t. This implies that the jump matrices of the RH problem of Figure 1.1 belong to H_1. Also, this RH problem satisfies a Schwartz-reflection symmetry. These two facts imply its unique solvability [7].

The most difficult part of the rigorous theory involves deriving L_1 estimates using PDE techniques. In §3, we show how this can be achieved for $q(x,t)$ and $xq(x,t)$, where t is fixed. The analogous but more difficult result for $q_x(0,t)$ is established in [14].

In §4, we use the elegant approach of [11] to study the asymptotic behavior of $q(x,t)$ as $t \to \infty$.

2. The formal analysis. We first study the x-RH problem.

2.1. The x-problem. Let the 2 × 2 matrices $\varphi(x,t,k)$ and $\psi(x,t,k)$ be the solutions of equation (1.3a) specified by the boundary conditions $\varphi(0,t,k) = \mathrm{I}$ and $\lim_{x\to\infty} \exp(ikx\sigma_3)\psi(x,t,k) = \mathrm{I}$, respectively. The matrix $\varphi(x,t,k)$ satisfies the

Volterra integral equation

$$\varphi(x,t,k) = e^{-ikx\sigma_3} + \int_0^x d\xi e^{-ik(x-\xi)\sigma_3} Q(\xi,t)\varphi(\xi,t,k). \tag{2.1.1}$$

The eigenfunction $\psi(x,t,k)$ satisfies a similar equation with $\int_0^x$ replaced by $-\int_x^\infty$. Letting

$$\Phi(x,t,k) = \varphi(x,t,k)e^{ikx\sigma_3}, \qquad \Psi(x,t,k) = \psi(x,t,k)e^{ikx\sigma_3}, \tag{2.1.2}$$

it follows that Φ satisfies the Volterra integral equation

$$\Phi(x,t,k) = I + \int_0^x d\xi e^{-ik(x-\xi)\sigma_3} Q(\xi,t)\Phi(\xi,t,k)e^{ik(x-\xi)\sigma_3}. \tag{2.1.3}$$

This equation, and the analogous equation satisfied by $\Psi(x,t,k)$, imply that

$$\begin{gathered}\Phi = (\Phi^+,\Phi^-) = \begin{pmatrix} \Phi_1^+ & \Phi_1^- \\ \Phi_2^+ & \Phi_2^- \end{pmatrix}, \qquad \Psi = (\Psi^-,\Psi^+) = \begin{pmatrix} \Psi_1^- & \Psi_1^+ \\ \Psi_2^- & \Psi_2^+ \end{pmatrix}, \\ \Phi,\Psi \to \mathrm{I} \text{ as } k\to\infty. \end{gathered} \tag{2.1.4}$$

This notation is explained in the introduction. It is important to notice that, since Φ satisfies a Volterra integral equation with compact support, Φ is an entire function with respect to k.

The eigenfunctions Φ and Ψ have unit determinant. Furthermore, they satisfy certain symmetry conditions. Suppressing for convenience of writing the x,t dependence, these symmetries are

(2.1.5)

$$\overline{\Phi_1^+(k)} = \Phi_2^-(\bar{k}), \qquad \overline{\Psi_1^-(k)} = \Psi_2^+(\bar{k}), \qquad \overline{\Phi_2^+(k)} = \lambda\Phi_1^-(\bar{k}), \qquad \overline{\Psi_2^-(k)} = \lambda\Psi_1^+(\bar{k}).$$

Since the functions φ and ψ are both solutions of equations (1.3), they are related through an x-independent matrix. Hence $\Phi = \Psi\exp(-ikx\sigma_3)(\psi(0,t,k))^{-1}\cdot\exp(ikx\sigma_3)$. Using this equation and defining χ_p by equation (1.12), one finds the jump condition (1.13), where the 2×2 matrix $F(x,t,k)$ is given by

$$\begin{gathered} F_{11} = 1, \qquad F_{22} = \frac{1}{|\psi_2^+(0,t,k)|^2}, \qquad F_{12} = \frac{-\psi_1^+(0,t,k)}{\overline{\psi_2^+(0,t,k)}}\exp(-2ikx), \\ F_{21} = \frac{\lambda\overline{\psi_1^+(0,t,k)}}{\psi_2^+(0,t,k)}\exp(2ikx), \end{gathered} \tag{2.1.6}$$

and $\psi(0,t,k) = \Psi(0,t,k)$.

2.2. The *t*-problem. The eigenfunction $\psi(x,t,k)$ satisfies equation (1.3b). Evaluating this equation as $x\to\infty$, it follows that $C(t) = 2ik^2\sigma_3$. Hence $\psi(0,t,k)$ solves

$$\psi_t(0,t,k) + 2ik^2[\sigma_3,\psi(0,t,k)] = \hat{Q}(t,k)\psi(0,t,k), \tag{2.2.1}$$

where

$$\hat{Q}(t,k) = 2kQ(0,t) - i\lambda|q|^2(0,t)\sigma_3 - iQ_x(0,t)\sigma_3. \tag{2.2.2}$$

The matrix $\hat{Q}(t,k)$ involves $Q(0,t)$, which is known, but it also involves $Q_x(0,t)$, which is unknown. The main idea of [4] and [5] is to solve equation (2.2.1) by formulating an inverse scattering problem. The analysis is similar to that of §2.1. The eigenfunctions $\hat{\Phi}(t,k)$ and $\hat{\Psi}(t,k)$ are the solutions of equation (2.2.1) specified by the boundary conditions $\hat{\Phi}(0,k) = \mathrm{I}$ and $\lim_{t\to\infty}\hat{\Psi}(t,k) = \mathrm{I}$, respectively. In analogy with equation (2.1.3), $\hat{\Phi}$ satisfies the Volterra integral equation

$$\hat{\Phi}(t,k) = \mathrm{I} + \int_0^t d\tau e^{-2ik^2(t-\tau)\sigma_3}\hat{Q}(\tau,k)\hat{\Phi}(\tau,k)e^{2ik^2(t-\tau)\sigma_3}. \tag{2.2.3}$$

This equation, and the analogous equation satisfied by $\hat{\Psi}(x,t,k)$ (where $\int_0^t$ is replaced by $-\int_t^\infty$) imply that

$$\begin{gathered}\hat{\Phi} = (\hat{\Phi}^+,\hat{\Phi}^-) = \begin{pmatrix} \hat{\Phi}_1^+ & \hat{\Phi}_1^- \\ \hat{\Phi}_2^+ & \hat{\Phi}_2^- \end{pmatrix}, \quad \hat{\Psi} = (\hat{\Psi}^-,\hat{\Psi}^+) = \begin{pmatrix} \hat{\Psi}_1^- & \hat{\Psi}_1^+ \\ \hat{\Psi}_2^- & \hat{\Psi}_2^+ \end{pmatrix}, \\ \hat{\Phi},\hat{\Psi} \to \mathrm{I} \quad \text{as} \quad k \to \infty. \end{gathered} \tag{2.2.4}$$

This notation is explained in the introduction. The matrices $\hat{\Phi}$ and $\hat{\Psi}$ have unit determinant, and they satisfy the same symmetry relations as those satisfied by Φ and Ψ (equations (2.1.5)).

It turns out that the following important relationships are valid between $\psi(0,t,k)$, $\hat{\Psi}(t,k)$ and $\hat{\Phi}(t,k)$:

$$\begin{aligned} \psi^+(0,t,k) &= \frac{\psi_2^+(0,0,k)}{\hat{\Psi}_2^+(0,k)}\hat{\Psi}^+(t,k), \quad k \in \mathrm{I}; \\ \psi^-(0,t,k) &= \frac{\psi_1^-(0,0,k)}{\hat{\Psi}_1^-(0,k)}\hat{\Psi}^-(t,k), \quad k \in \mathrm{IV}, \end{aligned} \tag{2.2.5}$$

$$\psi^+(0,t,k) = \psi_2^+(0,0,k)\hat{\Phi}^-(t,k) + \psi_1^+(0,0,k)\hat{\Phi}^+(t,k)\exp(-4ik^2t), \quad k \in \mathbb{C}^+, \tag{2.2.6}$$

$$\psi^-(0,t,k) = \psi_1^-(0,0,k)\hat{\Phi}^+(t,k) + \psi_2^-(0,0,k)\hat{\Phi}^-(t,k)\exp(4ik^2t), \quad k \in \mathbb{C}^-, \tag{2.2.7}$$

where I and IV denote the first and the fourth quadrants of the complex k-plane. To derive these equations, we use the fact that ψ^-E^{-1}, ψ^+E, $\hat{\Psi}^-E^{-1}$, $\hat{\Phi}^-E$, $\hat{\Psi}^+E$, and $\hat{\Phi}^+E^{-1}$, where $E = \exp(2ik^2t)$, are related through equations involving constant coefficients. For example,

$$\psi^+(0,t,k) = \alpha\hat{\Psi}^+(t,k) + \beta\hat{\Psi}^-(t,k)\exp(-4ik^2t), \quad k > 0,$$

where α and β are t-independent scalars. Using $\beta = \det(\psi^+(0,t,k), \hat{\Psi}^+(t,k))\exp(4ik^2t)$ and the fact that the rhs of this equation is a + function going to zero as $t \to \infty$, it follows that $\beta = 0$. Then, evaluating this equation at $t = 0$, equation (2.2.5a) follows. Equations (2.2.5b), (2.2.6), and (2.2.7) are derived in a similar way, where one also uses the fact that $\hat{\Phi}$ is an entire function in the k-complex plane. We also note that equations (2.2.5b) and (2.2.7) follow from equations (2.2.5a) and (2.2.6) using the underlying symmetry conditions.

Let us introduce the notation

$$\psi(0) \doteqdot \psi(0,0,k), \quad \hat{\Psi}(0) \doteqdot \hat{\Psi}(0,k), \quad e = \exp(4ik^2t). \tag{2.2.8}$$

We define the 2×2 matrix $\hat{Y}_p$ by

$$\text{(2.2.9)} \qquad \begin{gathered} \left(\frac{\hat{\Phi}^+(t,k)}{\psi_2^+(0)}, \psi^+(0,t,k)\right), \quad \left(\frac{\hat{\Psi}^-(t,k)}{\rho(k)}, \psi^+(0,t,k)\right), \\ \left(\psi^-(0,t,k), \frac{\hat{\Psi}^+(t,k)}{\nu(k)}\right), \quad \left(\psi^-(0,t,k), \frac{\hat{\Phi}^-(t,k)}{\psi_1^-(0)}\right) \end{gathered}$$

for $k \in \mathrm{I}, \dots, \mathrm{IV}$, respectively, where the scalar functions $\rho(k)$ and $\nu(k)$ are given by

$$\text{(2.2.10)} \qquad \rho(k) \doteqdot \hat{\Psi}_1^-(0)\psi_2^+(0) - \hat{\Psi}_2^-(0)\psi_1^+(0), \quad \nu(k) \doteqdot \hat{\Psi}_2^+(0)\psi_1^-(0) - \hat{\Psi}_1^+(0)\psi_2^-(0).$$

In defining $\hat{Y}_p$, we use when possible $\psi(0,t,k)$ instead of $\hat{\Phi}(t,k)$ and $\hat{\Psi}(t,k)$; the scalars appearing in equation (2.2.9), are chosen by the requirement that $\det \hat{Y}_p = 1$. Since the $\det \hat{Y}_p$ is t-independent, it follows that $\det \hat{Y}_p = 1$ for $k \in \mathrm{II} \cup \mathrm{III}$. For $k \in \mathrm{I}$, the $\det \hat{Y}_p$ is given by

$$\hat{\Phi}_1^+ \frac{\psi_2^+}{\psi_2^+(0)} - \hat{\Phi}_2^+ \frac{\psi_1^+}{\psi_2^+(0)} = \hat{\Phi}_1^+ \frac{\hat{\Psi}_2^+}{\hat{\Psi}_2^+(0)} - \hat{\Phi}_2^+ \frac{\hat{\Psi}_1^+}{\hat{\Psi}_2^+(0)},$$

where we have used equation (2.2.5a). The rhs of this equation simplifies to 1 using the 22 component of the equation $(\hat{\Phi})^{-1}\hat{\Psi} = \exp(-2ik^2t\sigma_3)\hat{\Psi}(0)\exp(2ik^2t\sigma_3)$, i.e., equation (2.2.11b),

$$\text{(2.2.11)} \qquad \begin{aligned} \hat{\Phi}_2^-\hat{\Psi}_1^- - \hat{\Phi}_1^-\hat{\Psi}_2^- &= \hat{\Psi}_1^-(0), & \qquad \hat{\Phi}_1^+\hat{\Psi}_2^+ - \hat{\Phi}_2^+\hat{\Psi}_1^+ &= \hat{\Psi}_2^+(0), \\ \hat{\Phi}_2^-\hat{\Psi}_1^+ - \hat{\Phi}_1^-\hat{\Psi}_2^+ &= \hat{\Psi}_1^+(0)e^{-1}, & \qquad \Phi_1^+\hat{\Psi}_2^- - \hat{\Phi}_2^+\hat{\Psi}_1^- &= \hat{\Psi}_2^-(0)e. \end{aligned}$$

A similar analysis implies that $\det \hat{Y}_p = 1$ for $k \in \mathrm{IV}$.

The matrix $\hat{Y}_p$ satisfies the following jump conditions:

$$\text{(2.2.12)} \qquad \hat{Y}_p^-(t,k) = \hat{Y}_p^+(t,k)\hat{G}(t,k), \quad k^2 \in \mathbb{R};$$

$$\text{(2.2.13)} \qquad k \in i\mathbb{R}^+ : \quad \hat{G}_{11} = \hat{G}_{22} = 1, \qquad \hat{G}_{12} = 0, \qquad \hat{G}_{21} = \frac{\hat{\Psi}_2^-(0)}{\psi_2^+(0)\rho}e;$$

$$\text{(2.2.14)} \qquad k \in i\mathbb{R}^- : \quad \hat{G}_{11} = \hat{G}_{12} = 1, \qquad \hat{G}_{21} = 0, \qquad \hat{G}_{12} = -\frac{\hat{\Psi}_1^+(0))}{(\psi_1^-(0))\overline{\rho(\bar{k})}}e^{-1};$$

(2.2.15)

$$k \in \mathbb{R}^+ : \quad \hat{G}_{11} = 1, \quad \hat{G}_{22} = \frac{1}{\psi_2^+(0)\psi_1^-(0)}, \quad \hat{G}_{12} = -\frac{\psi_1^+(0)}{\psi_1^-(0)}e^{-1}, \quad \hat{G}_{21} = \frac{\psi_2^-(0)}{\psi_2^+(0)}e;$$

(2.2.16)

$$k \in \mathbb{R}^- : \quad \hat{G}_{11} = 1 - R_1R_2, \qquad \hat{G}_{22} = 1, \qquad \hat{G}_{12} = -R_1e^{-1}, \qquad G_{21} = R_2e,$$

where the scalar functions $R_1(k)$ and $R_2(k)$ are defined by

$$\text{(2.2.17)} \quad R_1 = \frac{\psi_1^+(0)\hat{\Psi}_2^+(0) - \psi_2^+(0)\hat{\Psi}_1^+(0)}{\psi_2^-(0)\hat{\Psi}_1^+(0) - \psi_1^-(0)\hat{\Psi}_2^+(0)}, \qquad R_2 \doteqdot \frac{\psi_1^-(0)\hat{\Psi}_2^-(0) - \psi_2^-(0)\hat{\Psi}_1^-(0)}{\hat{\Psi}_1^-(0)\psi_2^+(0) - \hat{\Psi}_2^-(0)\psi_1^+(0)}.$$

We now indicate how these jump conditions can be derived.

$k \in R^+$. Substituting the expression for $\hat{Y}_p$ in equation (2.2.12) and using $\det \psi = \det \hat{\Phi} = 1$, one finds $\hat{G}_{11} = 1$, $\hat{G}_{22} = 1/\psi_2^+(0)\psi_1^-(0)$, and

$$\hat{G}_{12} = \hat{\Phi}_1^- \frac{\psi_2^+}{\psi_1^-(0)} - \hat{\Phi}_2^- \frac{\psi_1^+}{\psi_1^-(0)}, \qquad \hat{G}_{21} = \hat{\Phi}_1^+ \frac{\psi_2^-}{\psi_2^+(0)} - \hat{\Phi}_2^+ \frac{\psi_1^-}{\psi_2^+(0)}.$$

Using equations (2.2.5) to eliminate $\psi_2^+, \psi_1^+, \psi_2^-$, and ψ_1^- and then using (2.2.11c) and (2.2.11d), these equations become

$$\hat{G}_{12} = -\frac{\psi_2^+(0)\hat{\Psi}_1^+(0)}{\psi_1^-(0)\hat{\Psi}_2^+(0)} e^{-1}, \qquad \hat{G}_{21} = \frac{\psi_1^-(0)\hat{\Psi}_2^-(0)}{\psi_2^+(0)\hat{\Psi}_1^-(0)} e.$$

Using equation (2.2.5) evaluated at $t = 0$, these expressions yield the expression for $\hat{G}_{12}$ and $\hat{G}_{21}$ given in equation (2.2.15).

$k \in iR^+$. Substituting the expressions for $\hat{Y}_p$ in equation (2.2.12), one finds $\hat{G}_{11} = 1$, $\hat{G}_{12} = 0$, and

$$\hat{G}_{22} = \hat{\Phi}_1^+ \frac{\psi_2^+}{\psi_2^+(0)} - \hat{\Phi}_2^+ \frac{\psi_1^+}{\psi_2^+(0)}, \qquad \hat{G}_{21} = \frac{\hat{\Phi}_1^+ \hat{\Psi}_2^- - \hat{\Phi}_2^+ \hat{\Psi}_1^-}{\rho \psi_2^+(0)}.$$

Using equation (2.2.5a) to eliminate ψ_2^+ and ψ_1^+ and then using equation (2.2.11b), it follows that $\hat{G}_{22} = 1$. Using equation (2.2.11d), it follows that $\hat{G}_{21} = \hat{\Psi}_2^-(0)e/\rho\psi_2^+(0)$.

$k \in iR^-, k \in R^-$. The derivation of the jump matrices defined in equations (2.2.14) and (2.2.16) is similar to the above and hence is omitted.

The number of independent functions appearing in the above jumps can be reduced by using the underlying symmetry conditions. Some of these functions can be computed effectively in terms of initial data, while the rest, although dependent in a nonlinear way on the boundary data, can be given an effective characterization in the k-complex plane.

Let us summarize the main formal results of this subsection.

PROPOSITION 2.1. *Let the* 2×2 *matrix value function* $\hat{Y}_p(t, k)$ *be defined by equation* (1.14). *Then* $\hat{Y}_p$ *satisfies equation* (1.16), *where the jump matrices are depicted in Figure* 1.1 *with* θ *replaced by* $4ik^2t$.

The functions $s_1^+(k)$ *and* $s_2^+(k)$ *defined in* (1.15) *can be computed in terms of the initial data* $q(x, 0)$. *The function* $c(k)$ *defined in* (1.17) *depends in a nonlinear way on the boundary data* $q(0, t)$. *These functions, which are called the scattering data, have the following properties:*

1. $s_1^+(k)$ *and* $s_2^+(k)$ *are analytic for* $k \in \mathbb{C}^+$; $s_1^+(k) \to 0$ *and* $s_2^+(k) \to 1$ *as* $k \to \infty$.
2.

(2.2.18)
$$|s_2^+(k)|^2 - \lambda |s_1^+(k)|^2 = 1, \quad k \in \mathbb{R}.$$

3. $c(k)$ *is analytic in* II *except for possible poles, which can occur at the zeros of* $s_2^+(k)$ *and at the points* $\{k_j\}_1^N$, $k_j \in$ II, *which are the zeros of the function* $\rho(k)$ *defined in* (1.15); $c(k) \to 0$ *as* $k \to \infty$.

Proof. It is straightforward to derive the jump matrices appearing in Figure 1.1 from equations (2.2.13)–(2.2.16). The jumps on R^+, $i\mathbb{R}^+$ follow from the notations introduced in (1.15), (1.17), the underlying symmetry, and the relation $|s_2^+|^{-2} =$

$1-\lambda|b|^2$, which in turn follows from $\det\psi(0)=1$ and the underlying symmetry. The jump on iR^- follows from the underlying symmetry. Taking again into account the underlying symmetry and the relation $\det\psi(0)=1$, it follows that for $k\in\mathbb{R}^-$,

$$b(k)-\lambda\overline{c(k)}=\frac{\psi_1^+(0)}{\psi_1^-(0)}-\frac{\hat{\Psi}_1^+(0)}{\psi_1^-(0)[\hat{\Psi}_2^+(0)\psi_1^-(0)-\hat{\Psi}_1^+(0)\psi_2^-(0)]}=$$

$$\frac{\psi_1^+(0)\hat{\Psi}_2^+(0)\psi_1^-(0)-\hat{\Psi}_1^+(0)(1+\psi_1^+(0)\psi_2^-(0))}{\psi_1^-(0)[\hat{\Psi}_2^+(0)\psi_1^-(0)-\hat{\Psi}_1^+(0)\psi_2^-(0)]}=\frac{\psi_1^+(0)\hat{\Psi}_2^+(0)-\hat{\Psi}_1^+(0)\psi_2^+(0)}{\hat{\Psi}_2^+(0)\psi_1^-(0)-\hat{\Psi}_1^+(0)\psi_2^-(0)}=-R_1.$$

Since $R_2=\lambda\bar{R}_1$, this implies the jump on $\mathbb{R}^-$ indicated in Figure 1.1.

The initial data $q(x,0)$ can be used to compute $\psi(x,0,k)$, which in turn implies $\psi(0,0,k)$. Hence all of the components of $\psi(0,0,k)$ can be effectively calculated in terms of $q(x,0)$. Equation (2.2.18) follows from $\det\psi(0)=1$. Although the scattering data $c(k)$ depends in a nonlinear way on the boundary data $q(0,t)$, its analytic structure can be characterized explicitly. The large-k behavior of s_1^+, s_2^+, and $c(k)$ follows from the large-k behavior of $\psi_1^+(0,0,k)$, $\psi_2^+(0,0,k)$, $\hat{\Psi}_2^-$, and $\hat{\Psi}_1^-$. The possible poles for $c(k)$ appear at the possible zeros of $s_2^+(k)$ and of $\rho(k)$. This completes the proof of Proposition 2.1.

Remark. If one investigates the direct and inverse problems associated with equation (1.3b) without assuming a relationship between $q(0,t)$ and $q_x(0,t)$, one finds that the basic scattering data is the reflection coefficient $r(k)$ defined by

$$r(k)=\frac{\hat{\Psi}_1^+(0,k)}{\hat{\Psi}_2^+(0,k)}. \tag{2.2.19}$$

The only analytic restriction on $r(k)$ is that it is meromorphic in $\mathrm{I}\cup\mathrm{III}$. However, in our case, since $q(0,t)$ and $q_x(0,t)$ are the boundary values of the NLS equation, there exists a relationship between $q(0,t)$ and $q_x(0,t)$. It is quite interesting that although this relationship is highly nonlinear, its representation in the scattering space is very simple: $r(k)$ has analytic continuation for $k\in\mathrm{II}$. Indeed, equation (2.2.5a) evaluated at $t=0$ becomes

$$\frac{\psi_1^+(0,0,k)}{\psi_2^+(0,0,k)}=\frac{\hat{\Psi}_1^+(0,k)}{\hat{\Psi}_2^+(0,k)},\quad k\in\mathrm{I}.$$

But $\psi^+(0,0,k)$ is analytic in $\mathrm{I}\cup\mathrm{II}$, hence $r(k)$ has an analytic continuation across $i\mathbb{R}^+$. The function $c(k)$ is related to $r(k)$ through the equation

$$c(k)=\frac{\lambda\overline{r(\bar{k})}}{(s_2^+(k))^2(1-\lambda\frac{s_1^+(k)}{s_2^+(k)}\overline{r(\bar{k})})}. \tag{2.2.20}$$

2.3. The RH problem. In this section, we formulate a RH problem, whose solution gives the solution of the initial boundary value problem of the NLS equation.

PROPOSITION 2.2. *Let the* 2×2 *matrix* $\hat{Z}_p(x,t,k)$ *be defined by equation* (1.18). *Then the following hold:*

(i) $\hat{Z}_p(x,t,k)$ *has unit determinant and it satisfies the RH problem* (1.19).

(ii) $\hat{Z}_p(x,t,k)$ *is a meromorphic function in the complex* k*-plane cut along* $Im(k^2)=0$. *Its possible poles can occur only at the zeros of* $\rho(k)$ *for* $k\in\mathrm{II}$ *and at the complex conjugate of these zeros.*

(iii) *The eigenfunction* $\Psi^+(x,t,k)$ *and the potential* $q(x,t)$ *can be obtained from* $\hat{Z}_p(x,t,k)$ *by*

(2.3.1)
$$\Psi^+(x,t,k) = \hat{Z}_p(x,t,k)(0,1)^T, \quad k \in \mathbb{C}^+; \quad q(x,t) = 2i \lim_{k\to\infty} (k\hat{Z}_p(x,t,k))_{12}, \quad k \in \mathrm{I},$$

where the subscript 12 *denotes the* 12 *components of* $\hat{Z}_p$.

Proof. For convenience of writing, we drop the subscript p in $\hat{Y}$.

(i). We first note that the jump matrix $F(x,t,k)$ defined in equation (2.1.6) can be written as

$$F(x,t,k) = \begin{pmatrix} 1 & 0 \\ \frac{\hat{Y}_{21}^{\mp}(t,k)}{\hat{Y}_{22}^{\pm}(t,k)} e^{2ikx} & 1 \end{pmatrix} \begin{pmatrix} 1 & -\frac{\hat{Y}_{12}^{\pm}(t,k)}{\hat{Y}_{11}^{\mp}(t,k)} e^{-2ikx} \\ 0 & 1 \end{pmatrix}, \quad k \gtrless 0. \tag{2.3.2}$$

Now consider the jump conditions for $\hat{Z}_p(x,t,k)$.

$k \in \mathbb{R}^+$. Equations (1.18), (1.13), and (2.3.2) yield

$$\begin{aligned}[\hat{Z}_p^+(x,t,k)]^{-1}\hat{Z}_p^-(x,t,k) &= \begin{pmatrix} 1 & 0 \\ -\frac{\hat{Y}_{21}^+}{\hat{Y}_{22}^+} e^{2ikx} & 1 \end{pmatrix} F(x,t,k) \begin{pmatrix} 1 & \frac{\hat{Y}_{12}^-}{\hat{Y}_{11}^-} e^{-2ikx} \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ \frac{\hat{Y}_{21}^- - \hat{Y}_{21}^+}{\hat{Y}_{22}^+} e^{2ikx} & 1 \end{pmatrix} \begin{pmatrix} 1 & \frac{\hat{Y}_{12}^- - \hat{Y}_{12}^+}{\hat{Y}_{11}^-} e^{-2ikx} \\ 0 & 1 \end{pmatrix},\end{aligned}$$

where $\hat{Y}_{jl}^{\pm} \equiv \hat{Y}_{jl}^{\pm}(t,k)$. Equation (1.16) implies

$$\hat{Y}_{21}^- - \hat{Y}_{21}^+ = \hat{Y}_{22}^+ \hat{G}_{21}(t,k), \qquad \hat{Y}_{12}^- - \hat{Y}_{12}^+ = \hat{Y}_{11}^- \hat{G}_{12}(t,k).$$

Using these equations it follows that

$$\begin{aligned}[\hat{Z}_p^+(x,t,k)]^{-1}\hat{Z}_p^-(x,t,k) &= \begin{pmatrix} 1 & 0 \\ \hat{G}_{21}(t,k)e^{2ikx} & 1 \end{pmatrix} \begin{pmatrix} 1 & \hat{G}_{12}(t,k)e^{-2ikx} \\ 0 & 1 \end{pmatrix} \\ &= e^{-i(kx+2k^2t)\sigma_3} G(k) e^{i(kx+2k^2t)\sigma_3}.\end{aligned}$$

$k \in i\mathbb{R}^+$. Equation (1.18) yields

$$[\hat{Z}_p^+((x,t,k)]^{-1}\hat{Z}_p^-(x,t,k) = \begin{pmatrix} 1 & 0 \\ \left(\frac{\hat{Y}_{21}^-}{\hat{Y}_{22}^-} - \frac{\hat{Y}_{21}^+}{\hat{Y}_{22}^+}\right) e^{2ikx} & 1 \end{pmatrix}.$$

Equation (1.16) implies

$$\hat{Y}_{22}^- = \hat{Y}_{22}^+, \quad \hat{Y}_{21}^- - \hat{Y}_{21}^+ = \hat{Y}_{22}^+ \hat{G}_{21}(t,k).$$

Using these equations, it follows that

$$[\hat{Z}_p^+(x,t,k)]^{-1}\hat{Z}_p^-(x,t,k) = \begin{pmatrix} 1 & 0 \\ \hat{G}_{21}(t,k)e^{2ikx} & 1 \end{pmatrix} = e^{-i(kx+2k^2t)\sigma_3} G(k) e^{i(kx+2k^2t)\sigma_3}.$$

The derivation of the jump conditions for $k \in \mathbb{R}^-$ and $k \in i\mathbb{R}^-$ is similar to the above and hence is omitted.

(ii). Let $k \in \mathrm{I}$. Using the definitions of $\chi_p(x,t,k)$ and $\hat{Y}_p(t,k)$ for $k \in \mathrm{I}$, i.e.,

$$\chi_p(x,t,k) = \left(\frac{\Phi^+(x,t,k)}{\psi_2^+(0,t,k)}, \Psi^+(x,t,k)\right), \ \hat{Y}_{21}(t,k) = \frac{\hat{\Phi}_2^+(t,k)}{s_2^+(k)}, \ \hat{Y}_{22}(t,k) = \psi_2^+(0,t,k),$$

equation (1.18) yields

$$\hat{Z}_p(x,t,k) = \left(\tfrac{1}{\psi_2^+(0,t,k)}\left(\Phi^+(x,t,k) + \tfrac{\hat{\Phi}_2^+(t,k)}{s_2^+(k)}\Psi^+(x,t,k)e^{2ikx}\right), \Psi^+(x,t,k)\right), \quad k \in \mathrm{I}. \tag{2.3.3}$$

We will show that at the possible zeros of $\psi_2^+(0,t,k)$, denoted by k_0,

$$\Phi^+(x,t,k_0) + \frac{\hat{\Phi}_2^+(t,k_0)}{s_2^+(k_0)}\Psi^+(x,t,k_0)e^{2ik_0x} = 0, \quad k_0 \in \mathrm{I}, \tag{2.3.4}$$

hence the possible poles of $\hat{Z}_p(x,t,k)$ for $k \in \mathrm{I}$ can occur only at the fixed possible zeros of $s_2^+(k)$.

To derive equation (2.3.4), we first relate $\Phi^+(x,t,k)$ to $\Psi^+(x,t,k)e^{2ikx}$, and to $\Phi^-(x,t,k)e^{2ikx}$ through an equation with x-independent coefficients. Evaluating this equation at $x = 0$ it follows that

$$\Phi^+(x,t,k) = \frac{\Psi^+(x,t,k)}{\psi_1^+(0,t,k)}e^{2ikx} - \frac{\psi_2^+(0,t,k)}{\psi_1^+(0,t,k)}\Phi^-(x,t,k)e^{2ikx}, \quad k \in \mathbb{C}^+. \tag{2.3.5}$$

Equation (2.3.4) follows from equation (2.3.5) evaluated at $k = k_0$. Indeed, when $k = k_0$, $\psi_2^+(0,t,k_0) = 0$, hence equation (2.2.6) yields

$$\psi_1^+(0,t,k_0) = s_2^+(k_0)\hat{\Phi}_1^-(t,k_0) + s_1^+(k_0)\hat{\Phi}_1^+(t,k_0)e^{-4ik_0^2t},$$

$$0 = s_2^+(k_0)\hat{\Phi}_2^-(t,k_0) + s_1^+(k_0)\hat{\Phi}_2^+(t,k_0)e^{-4ik_0^2t}, \quad k_0 \in \mathrm{I}. \tag{2.3.6}$$

These equations together with $\det\hat{\Phi} = 1$ imply $\psi_1^+(0,t,k_0)\hat{\Phi}_2^+(t,k_0) = -s_2^+(k_0)$.

Let $k \in \mathrm{II}$. Instead of equation (2.3.3) we now find

$$\hat{Z}_p(x,t,k) = \left(\tfrac{1}{\psi_2^+(0,t,k)}\left(\Phi^+(x,t,k) + \tfrac{\hat{\Psi}_2^-(t,k)}{\rho(k)}\Psi^+(x,t,k)e^{2ikx}\right), \Psi^+(x,t,k)\right), \quad k \in \mathrm{II}. \tag{2.3.7}$$

Again it can be shown that if $\psi_2^+(0,t,k_0) = 0$, $k_0 \in \mathrm{II}$, then

$$\Phi^+(x,t,k_0) + \frac{\hat{\Psi}_2^-(t,k_0)}{\rho(k_0)}\Psi^+(x,t,k_0)e^{2ik_0x} = 0, \quad k_0 \in \mathrm{II}; \tag{2.3.8}$$

hence the possible poles of $\hat{Z}_p(x,t,k)$ for $k \in \mathrm{II}$ can occur only at the fixed possible zeros of $\rho(k)$. Equation (2.3.8) follows from equation (2.3.5) using $\det\hat{Y}_p = 1$, $k \in \mathrm{II}$, i.e.,

$$\rho(k) = \hat{\Psi}_1^-(t,k)\psi_2^+(0,t,k) - \hat{\Psi}_2^-(t,k)\psi_1^+(0,t,k).$$

The absence of singularities of $\hat{Z}_p(x,t,k)$ at the possible zeros of $\psi_1^-(0,t,k)$ can be proven in a similar manner; alternatively, one can use the underlying symmetry.

To conclude the proof of statement (ii) it remains to notice that $s_2^+(k)$ cannot have zeros in I. In fact, if $s_2^+(k_0) = 0$ for $k_0 \in \mathrm{I}$, equation (2.2.6) implies

$$e^{4ik_0^2t}\psi^+(0,t,k_0) = s_1^+(k_0)\hat{\Phi}^+(t,k_0), \quad k_0 \in \mathrm{I}.$$

This means

$$\hat{\Phi}^+(t,k_0) \to 0, \quad t \to +\infty,$$

which contradicts $\det\hat{\Phi} = 1$ for all $t > 0$. Taking into account that $s_2^+(k)$ participates in the definition of $\hat{Z}_p$ only in the first quadrant I, we conclude that the function $\hat{Z}_p(x,t,k)$ can have poles only in the quadrants II and III and that these poles occur at the zeros of $\rho(k)$ and $\overline{\rho(\bar{k})}$, respectively. As we will see in §4, this fact implies that the possible solitons of the initial-boundary value problem (1.1) move away from the boundary.

(iii). Equation (2.3.1b) follows from the large-k asymptotics of equation (1.3).

We conclude this subsection with a discussion of the appropriate data needed to solve the RH problem.

PROPOSITION 2.3. *The RH problem for $\hat{Z}_p$ is uniquely specified by the following data, called scattering data:* (i) *$c(k)$ for $k \in \mathbb{R}^- \cup i\mathbb{R}^+$;* (ii) *$s_1^+(k)$ for $k \in \mathbb{R}$;* (iii) *the zeros of $s_2^+(k)$ for $k \in \mathbb{C}^+$;* (iv) *the zeros $\{k_j\}_1^N$ of $\rho(k)$ and the residues $\{c_j\}_1^N$ of $c(k)$ at these poles.*

Proof. Indeed, (ii) and (iii) together with equation (2.2.18) yield s_2^+ through the solution of a scalar RH problem. Then, $c(k)$, s_1^+, and s_2^+ specify all the jump conditions. The possible poles of $\hat{Z}_p$ can occur at the possible zeros of $\rho(k)$. The additional conditions on the residues of $\hat{Z}_p$ are given in the next section (equation 2.4.1)).

It is worth mentioning that $\rho(k) \neq 0$ for $k \in i\mathbb{R}^+$. Indeed, equation (2.2.5a) implies that for $k \in i\mathbb{R}^+$, $s_1^+(k) = \rho(k)\hat{\Psi}_1^+(0,k)$ and $s_2^+(k) = \rho(k)\hat{\Psi}_2^+(0,k)$. Thus $\rho(k) \neq 0$ since s_1^+ and s_2^+ cannot vanish simultaneously.

Remark 1. The above analysis is valid for the generic situation, i.e., we assume that all zeros of $\psi_2^+(0,t,k)$, of $s_2^+(k)$, and of $\rho(k)$ are simple, they do not coincide with each other, they do not lie on the cross $Imk^2 = 0$, and there is a finite number of them.

Remark 2. The function $c(k)$, $k \in \mathbb{R}^- \cup i\mathbb{R}^+$, satisfies certain restrictions. Indeed, since it is the boundary value of a function meromorphic in II and decreasing as $k \to \infty$, $c(k)$ satisfies the following infinite set of conditions:

$$\text{(2.3.9)} \qquad \frac{1}{2\pi i}\int_{\mathbb{R}^-\cup i\mathbb{R}^+} \frac{c(k')}{k'-k}dk' = \sum_{j=1}^{N} \frac{c_j}{k_j - k} + \sum_{j=1}^{M} \frac{c_j^+}{k_j^+ - k} \quad \forall k \notin \mathrm{II}$$

or

$$\text{(2.3.10)} \qquad \frac{1}{2\pi i}\int_{\mathbb{R}^-\cup i\mathbb{R}^+} c(k)(k-1)^{-n}dk = \sum_{j=1}^{N} c_j(k_j - 1)^{-n} + \sum_{j=1}^{M} c_j^+(k_j^+ - 1)^{-n},$$

where $k_j^+, j = 1, \ldots, M$ are the zeros of $s_2^+(k)$, and c_j^+ are the corresponding residues of $c(k)$. As a set of independent parameters for $c(k)$, one can take the *whole* set of its poles and the corresponding residues $\{k_j, k_j^+, c_j, c_j^+\}$ supplemented by

$$\alpha(k) \doteq Re\left\{ c(k) - \sum_{j=1}^{N} \frac{c_j}{k-k_j} - \sum_{j=1}^{M} \frac{c_j^+}{k-k_j^+} \right\}, \quad k \in \mathbb{R}^- \cup i\mathbb{R}^+.$$

Having $\{\alpha(k); k_j, k_j^+, c_j, c_j^+\}$, one can reconstruct $c(k)$ in *closed form.* Also, in order to reconstruct $s_1^+(k)$, one needs to know $\beta(k) \doteq Res_1^+(k)$. Thus the set of independent parameters for our RH problem are

$$\{\alpha(k); \beta(k); k_j, k_j^+, c_j, c_j^+\},$$

where $\alpha(k)$ and $\beta(k)$ are real-valued decreasing functions defined on the whole line, k_j, $k_j^+ \in \mathrm{II}$, c_j, $c_j^+ \in \mathbb{C}\backslash\{0\}$. In particular, this means that the functional dimension of the set can be represented by two complex-valued functions defined on the half-line.

2.4. Solitons. It turns out that the zeros of $\rho(k)$ give rise to solitons.

PROPOSITION 2.4. *Assume that $\rho(k)$ has a finite number of simple zeros for $k \in \mathrm{II}$. Let these zeros be denoted by $\{k_j\}_1^N$. Let $\{c_j\}_1^N$ be the corresponding residues of $c(k)$. Then the following hold:*

(i) *The first column $\hat{Z}_p^{(1)}(x,t,k)$ of $\hat{Z}_p(x,t,k)$ has a simple pole at k_j, $j = 1,\dots,N$, and the second column $\hat{Z}_p^{(2)}(x,t,k)$ of $\hat{Z}_p(x,t,k)$ has a simple pole at $\bar{k}_j$, $j = 1,\dots,N$; the corresponding residues satisfy the equations*

$$\begin{aligned} &\operatorname*{res}_{k_j} \hat{Z}_p^{(1)}(x,t,k) = c_j e^{\theta(k_j)} \hat{Z}_p^{(2)}(x,t,k_j), \qquad \theta(k) \doteq 2i(kx+2k^2t), \\ &\operatorname*{res}_{\bar{k}_j} \hat{Z}_p^{(2)}(x,t,k) = \lambda \bar{c}_j e^{-\theta(\bar{k}_j)} \hat{Z}_p^{(1)}(x,t,\bar{k}_j). \end{aligned} \tag{2.4.1}$$

(ii) *Equations* (2.4.1) *together with the jump condition* (1.19a) *and the asymptotic condition* (1.19b) *characterize the function $\hat{Z}_p(x,t,k)$ uniquely.*

(iii) *The RH problem* (1.19), (2.4.1) *can be solved as follows: Let $\hat{Z}(x,t,k)$ be the solution of an RH problem satisfying the same jump conditions as $\hat{Z}_p(x,t,k)$, but with $c(k)$ and $b(k)$ replaced by $c_0(k) \doteq c(k)\Pi_{j=1}^N(k-k_j)/(k-\bar{k}_j)$ and $b_0(k) = b(k)\Pi_{j=1}^N(k-\bar{k}_j)/(k-k_j)$. This RH problem is regular, i.e., $\hat{Z}(x,t,k)$ has no singularities in the complex k-plane. Then $\hat{Z}_p(x,t,k)$ can be found from $\hat{Z}$ through the equation*

(2.4.2)

$$\begin{aligned} &\hat{Z}_p(x,t,k) \\ &\quad = (kI+B_N)(kI+B_{N-1})\cdots(kI+B_1)\hat{Z}(x,t,k) \begin{pmatrix} \frac{1}{\Pi_{j=1}^N(k-k_j)} & 0 \\ 0 & \frac{1}{\Pi_{j=1}^N(k-\bar{k}_j)} \end{pmatrix}, \end{aligned}$$

where the 2×2 matrices $B_1,\dots,B_N$ are independent of k. These matrices can be determined recursively by solving the algebraic *equations*

$$\begin{aligned} &(k_j\mathrm{I} + B_j)\hat{Z}_{j-1}(x,t,k_j)\begin{pmatrix} 1 \\ -d_j(x,t) \end{pmatrix} = 0, \\ &(\bar{k}_jI + B_j)\hat{Z}_{j-1}(x,t,\bar{k}_j)\begin{pmatrix} -\lambda\bar{d}_j(x,t) \\ 1 \end{pmatrix} = 0, \end{aligned} \tag{2.4.3}$$

where

$$\hat{Z}_j(x,t,k) = (k\mathrm{I} + B_j)\hat{Z}_{j-1}(x,t,k), \quad j = 1,\dots,N-1, \quad \hat{Z}_0 = \hat{Z},$$

$$d_j(x,t) = c_j \frac{\Pi_{l=1,l\neq j}^{N}(k_j - k_l)}{\Pi_{l=1}^{N}(k_j - \bar{k}_l)} \exp(2ik_jx + 4ik_j^2 t). \tag{2.4.4}$$

The RH problem for $\hat{Z}$, and the algebraic equations for $B_1,\dots,B_N$ in the focusing case $\lambda = -1$, are always solvable.

Proof. $\hat{Z}_p(x,t,k)$ is given by equations (2.3.3) and (2.3.7) for k in the first and second quadrants, respectively. Since $s_2^+(k)$ has no zeros in I, $\hat{Z}_p$ is analytic for $k \in$ I, while the first column of $\hat{Z}_p$ has poles for $k \in$ II at the zeros of $\rho(k)$, i.e., at the points $k_1,\dots,k_N$, which are simultaneously the poles of $c(k)$. For $k \in i\mathbb{R}^+$,

$$\hat{Z}_p^-(x,t,k)\begin{pmatrix} 1 & 0 \\ -c(k)e^{\theta} & 1 \end{pmatrix} = \hat{Z}_p^+(x,t,k). \tag{2.4.5}$$

Because $\hat{\Phi}(t,k)$ is an entire function, both sides of equation (2.4.5) have an analytic continuation into the second quadrant of the complex k-plane; moreover, $\hat{Z}_p^+(x,t,k)$ has no singularities at $k_1,\dots,k_N$. This shows that the lhs of equation (2.4.5) has no singularities at $k_1,\dots,k_N$ as well. Similar considerations apply for $k \in$ III; thus

$$\hat{Z}_p(x,t,k)\begin{pmatrix} 1 \\ -c(k)e^{\theta} \end{pmatrix} \quad \text{and} \quad \hat{Z}_p(x,t,k)\begin{pmatrix} -\lambda\overline{c(\bar{k})}e^{-\theta(k)} \\ 1 \end{pmatrix}$$

have no singularities at k_j and $\bar{k}_j$, respectively. This implies equations (2.4.1).

In order to prove statement (ii), we suppose that $\hat{z}_p(x,t,k)$ is another solution of the RH problem (1.19), (2.4.1). The function $\hat{R}(k)$,

$$\hat{R} = \hat{z}_p\hat{Z}_p^{-1},$$

has no jumps at $k^2 \in \mathbb{R}$. Its only possible singularities ($\det \hat{Z}_p = 1$) are at the points k_j and $\bar{k}_j$. For $k \in$ II, one can rewrite $\hat{R}$ as

$$\hat{R}(k) = \hat{z}_p(k)\begin{pmatrix} 1 & 0 \\ -c(k)e^{\theta} & 1 \end{pmatrix}\left[\hat{Z}_p^-(k)\begin{pmatrix} 1 & 0 \\ -c(k)e^{\theta} & 1 \end{pmatrix}\right]^{-1},$$

which shows that $\hat{R}(k)$ has actually no singularities at k_j. Similar considerations for $k \in$ III show that $\hat{R}(k)$ has no singularities at $\bar{k}_j$ either. These facts together with (1.19b) imply that $\hat{R} = \mathrm{I}$.

The essence of deriving the statement (iii) is the usual idea of using Darboux transformations to solve an RH problem with singularities. We need to show that the function $\hat{Z}_p$ defined through (2.4.2)–(2.4.4) satisfies the RH problem (1.19), (2.4.1). In fact, by construction, the matrix $\hat{Z}_p$ satisfies the correct jump condition, and $\hat{Z}_p \to \mathrm{I}$ as $k \to \infty$. Also,

$$\underset{k_j}{res}\ \hat{Z}_p^{(1)}(x,t,k) = P_j(k_j)(k_j\mathrm{I} + B_j)\hat{Z}_{j-1}(x,t,k_j)\begin{pmatrix} 1 \\ 0 \end{pmatrix}\frac{1}{\Pi_{l=1,l\neq j}^{N}(k_j - k_l)},$$

$$\hat{Z}_p^{(2)}(x,t,k_j) = P_j(k_j)(k_j\mathrm{I} + B_j)\hat{Z}_{j-1}(x,t,k_j)\begin{pmatrix} 0 \\ 1 \end{pmatrix}\frac{1}{\Pi_{l=1}^{N}(k_j - \bar{k}_l)},$$

where

$$P_j(k) = (k\mathrm{I} + B_N)(k\mathrm{I} + B_{N-1}) \cdots (k\mathrm{I} + B_{j+1}).$$

This yields equation (2.4.1a) as a direct consequence of equation (2.4.3a); actually, since $\det P_j(k_j) \neq 0$, these two equations are equivalent. Similarly, equation (2.4.1b) follows from (2.4.3b).

Due to the underline symmetry of the jump conditions, the regular RH problem for $Z(x,t,k)$ is always solvable (i.e., there exists a vanishing lemma [7]). In order to discuss the solvability of the algebraic equations (2.4.3), we rewrite them in the form

$$B_j = -W_j \begin{pmatrix} k_j & 0 \\ 0 & \bar{k}_j \end{pmatrix} W_j^{-1},$$

$$W_j \doteqdot \left(\hat{Z}_{j-1}(x,t,k_j) \begin{pmatrix} 1 \\ -d_j(x,t) \end{pmatrix}, \ \hat{Z}_{j-1}(x,t,\bar{k}_j) \begin{pmatrix} -\lambda \overline{d_j}(x,t) \\ 1 \end{pmatrix} \right).$$

This implies that the system of the algebraic equations for the matrices $B_1, \ldots, B_N$ is solvable iff

$$\det W_j \neq 0, \quad j = 1, \ldots, N, \quad \forall x, t > 0. \tag{2.4.6}$$

For $\lambda = -1$ (focusing case), equations (2.4.6) are always valid. Indeed, the symmetry

$$\hat{Z}(x,t,k) = \sigma_2 \overline{\hat{Z}(x,t,\bar{k})} \sigma_2 \tag{2.4.7}$$

yields the formula

$$W_1 = \sigma_2 \bar{W}_1 \sigma_2, \tag{2.4.8}$$

or

$$\det W_1 = |(W_1)_{11}|^2 + |(W_1)_{21}|^2, \tag{2.4.9}$$

where $(W_1)_{11}$ and $(W_1)_{21}$ denote the 11 and 21 components of W_1. Since $\det \hat{Z} = 1$, $(W_1)_{11}$ and $(W_1)_{21}$ cannot be zero simultaneously, and equation (2.4.6) is valid for $j = 1$. The function $\hat{Z}_1$ satisfies the same symmetry condition (2.4.7) as $\hat{Z}$, and $\det \hat{Z}_1(x,t,k) = (k - k_1)(k - \bar{k}_1) \neq 0$ for $k = k_2, \bar{k}_2$; repeating the above arguments, it follows that (2.4.6) is valid for $j = 2$ together with the symmetry condition for $\hat{Z}_2$. Similar considerations apply to $j = 3, \ldots, N$. This completes the proof of Proposition 2.4.

In the defocusing case, $\lambda = 1$, the above arguments regarding the solvability of the algebraic system (2.4.3) are not valid. In this case one has to replace σ_2 in equations (2.4.7) and (2.4.8) by σ_1. This implies that the + sign in (2.4.9) is replaced by the − sign. Actually, the asymptotic analysis of §4 suggests that the solvability condition (2.4.6) does not hold in the defocusing case. In other words, solitons do not exist for $\lambda = 1$.

3. Rigorous considerations. We first discuss the x-problem. Let $\Phi(x,t,k)$ be defined by the integral equation (2.1.3). Let $\Psi(x,t,k)$ be defined by a similar integral equation with $\int_0^x$ replaced by $\int_\infty^x$. If $q(x,t) \in L_1(\mathbb{R}^+)$ in x, these integral equations have continuous and bounded solutions for fixed t and k. Furthermore,

Φ and Ψ have the analytic dependence in k indicated in equation (2.1.4). Also the Riemann–Lebesgue lemma implies that Φ, $\Psi \to \mathrm{I}$ as $k \to \infty$. If $q(x,t)$, $xq(x,t) \in L_1(\mathbb{R}^+) \cap L_2(\mathbb{R}^+)$ in x, then $\psi(0,t,k) \in H_1(\mathbb{R})$ in k.

In summary, if

$$(3.1) \qquad q(x,t), \;\; xq(x,t) \in L_1(\mathbb{R}^+) \cap L_2(\mathbb{R}^+) \;\text{ in }\; x,$$

then the formal results of §2.1 are justified. In particular the jump condition (1.13) is valid and the coefficients $\psi(0,t,k) - \mathrm{I}$ of the jump matrix (2.1.6) belong to $H_1(\mathbb{R})$.

We next consider the t-problem. Let $\hat{\Phi}(t,k)$ be defined by the integral equation (2.2.4). Let $\hat{\Psi}(t,k)$ be defined by a similar integral equation with $\int_0^t$ replaced by $\int_\infty^t$. If $q(0,t)$, $q_x(0,t) \in L_1(\mathbb{R}^+)$, these integral equations have continuous and bounded solutions for fixed k. Also, $\hat{\Phi}$ and $\hat{\Psi}$ have the analytic dependence in k indicated in equation (2.2.4) and $\hat{\Phi}, \hat{\Psi} \to I$ as $k \to \infty$. Furthermore, if in addition, $q(0,t) \in L_2(\mathbb{R}^+)$ and the first derivatives of $q(0,t)$ and $q_x(0,t) \in L_1(\mathbb{R}^+)$, then $\hat{\Psi}(0,k) - \mathrm{I} \in L_2$ for $k^2 \in \mathbb{R}$. Finally, if in addition, $tq(0,t)$, $tq_t(0,t)$, $tq_{tt}(0,t) \in L_1(\mathbb{R}^+)$, then $\hat{\Psi}(0,k) - \mathrm{I} \in H_1$ for $k^2 \in \mathbb{R}$.

In summary, let

$$(3.2) \qquad v(t) \div q(0,t), \quad w(t) \div q_x(0,t).$$

If

$$(3.3) \; v(t) \in L_1 \cap L_2(\mathbb{R}^+), \qquad v'(t), \;\; w(t), \;\; w'(t), \;\; tv(t), \;\; tv'(t), \;\; tv''(t) \in L_1(\mathbb{R}^+),$$

then $\hat{\Psi}(t,k)$ and $\hat{\Phi}(t,k)$ exist and are related through a matrix $\hat{\Psi}(0,k)$ such that $\hat{\Psi}(0,k) - \mathrm{I} \in H_1$ for $k^2 \in \mathbb{R}$.

In §2.2, we formulated an RH problem which involves $\psi(0,t,k)$ in addition to $\hat{\Phi}(t,k)$ and $\hat{\Psi}(t,k)$. The function $\psi(0,t,k)$ satisfies equation (2.2.1), which is uniquely determined in terms of $v(t)$ and $w(t)$. On the other hand, $\psi(x,0,k)$ satisfies equation (1.3a), which is uniquely defined in terms of $u(x) = q(x,0)$. The equality of $\psi(0,t,k)$ and $\psi(x,0,k)$ at $x = t = 0$ is guaranteed iff

$$(3.4) \qquad u(0) = v(0).$$

The relationship between $\psi(0,t,k)$, $\hat{\Phi}(t,k)$, and $\hat{\Psi}(t,k)$ involves the matrices $\psi(0,0,k)$ and $\hat{\Psi}(0,k)$, which under the assumptions (3.1)–(3.4) belong to H_1 and have the analytic dependence indicated by the appropriate superscripts.

THEOREM 3.1. *Assume the following:*

(i) *Equations* (3.1), (3.3), *and* (3.4) *are valid.*

(ii) *All the zeros of* $\psi_2^+(0,t,k)$, *of* $s_2^+(k)$ (*defined by equation* (1.15a)), *and of* $\rho(k)$ (*defined by equation* (1.15b)) *are simple, they do not coincide with each other, they do not lie on* $k^2 \in \mathbb{R}$, *and there is at most a finite number of them. The function* $\rho(k)$ *has no zeros if* $\lambda = 1$.

Then, the 2×2 *matrix* $\hat{Z}_p(x,t,k)$ *defined by equation* (1.18) *has unit determinant and satisfies the jump conditions* (1.19). *The jump matrices are defined in terms of* $\psi(0,0,k)$ *and* $\hat{\Psi}(0,k)$, *which satisfy* $\psi(0,0,k) - \mathrm{I}$, $\hat{\Psi}(0,k) - \mathrm{I} \in H_1$. *The matrix* $\hat{Z}_p(x,t,k)$ *is a meromorphic function in the complex* k*-plane cut along* $Im(k^2) = 0$. *Its possible poles can occur only at the zeros of* $\rho(k)$ *for* $k \in \mathrm{II}$ *and at the complex conjugate of these zeros. The matrix* $\hat{Z}_p(x,t,k)$ *can be obtained by solving the RH problem of Figure* 1.1. *This RH is always solvable. The solution* $q(x,t)$ *of the NLS can be obtained from* $q(x,t) = 2i \lim_{k \to \infty}(k\hat{Z}_p(x,t,k))_{12}$, $k \in \mathrm{I}$.

Proof. It follows from Propositions 2.1 and 2.2 and the following properties of the jump matrices: (a) they are H_1 functions; (b) their product equals unity, which guarantees continuity at $k = 0$; (c) they satisfy certain symmetry conditions involving complex conjugation and transposition. These symmetry conditions imply that there exists a vanishing lemma [7], i.e., the homogeneous RH problem has only the zero solution. □

The main technical difficulty associated with the scheme presented in this paper is to find conditions of $u(x)$ and $v(t)$ which guarantee the assumption (3.1) as well as the assumption that $w(t)$, $w'(t) \in L_1(\mathbb{R}^+)$.

It was shown in [13] that if

$$u(x) \in H_2(\mathbb{R}^+), \quad v \in C_2(\mathbb{R}^+), \quad \text{and} \quad u(0) = v(0), \tag{3.5}$$

then

$$t \to q(,t) \text{ is continuous from } \mathbb{R}^+ \text{ into } H_2(\mathbb{R}^+) \tag{3.6}$$

and

$$t \to q_t(,t) \text{ is continuous from } \mathbb{R}^+ \text{ into } L_2(\mathbb{R}^+). \tag{3.7}$$

Following Sung [14], we shall show that under some additional conditions on $u(x)$, assumption (3.1) is valid.

LEMMA 3.1. *Assume that in addition to assumptions* (3.5),

$$xu(x) \text{ and } x^2u(x) \in L_2(\mathbb{R}^+). \tag{3.8}$$

Then equation (3.1) *is valid.*

Proof. We shall first show that the assumption $xu(x) \in L_2(\mathbb{R}^+)$ implies that $q(x,t) \in L_1(\mathbb{R}^+)$ for fixed t.

The NLS equation and its complex conjugate imply

$$|q|_t^2 = i(q_x\bar{q} - \bar{q}_x q)_x.$$

Thus

$$\begin{aligned}\partial_t \int_0^\infty x^2|q|^2 e^{-\varepsilon x}dx &= i\textstyle\int_0^\infty x^2 e^{-\varepsilon x}(q_x\bar{q} - \bar{q}_x q)_x dx\\ &= -i\int_0^\infty (2x - \varepsilon x^2)e^{-\varepsilon x}(q_x\bar{q} - \bar{q}_x q)dx\\ &\le \int_0^\infty [4x|q||q_x|e^{-\varepsilon x} + 2\varepsilon x^2|q||q_x|e^{-\varepsilon x}]dx\\ &= 4\int_0^\infty \left|e^{-\frac{\varepsilon x}{2}}q_x\right| \left|xqe^{-\frac{\varepsilon x}{2}}\right| dx + 4\int_0^\infty \left|\frac{\varepsilon x}{2}e^{-\frac{\varepsilon x}{2}}q_x\right| \left|xqe^{-\frac{\varepsilon x}{2}}\right| dx.\end{aligned}$$

The maxima of $\exp(-\varepsilon x/2)$ and $\varepsilon x/2\exp(-\varepsilon x/2)$ are 1 and $1/e$, respectively; thus the lhs of the above equation is not greater than

$$4\left(1 + \frac{1}{e}\right)\int_0^\infty |q_x| \left|xqe^{-\frac{\varepsilon x}{2}}\right| dx.$$

Let

$$\varphi(t) \doteqdot \left\|xq(x,t)e^{-\frac{\varepsilon x}{2}}\right\|_{L_2^x(\mathbb{R}^+)},$$

where $L_2^x(\mathbb{R}^+)$ indicates L_2 in the variable x. Then

$$\partial_t(\varphi^2) \leq 4\left(1+\frac{1}{e}\right)\max_{t\in[0,\infty)}\|q_x\|_{L_2^x(\mathbb{R}^+)}\varphi. \tag{3.9}$$

Assuming that $xq(x,0) \in L_2(\mathbb{R}^+)$ and letting $\varepsilon \to 0$, equation (3.9) yields that $xq(x,t) \in L_2^x(\mathbb{R}^+)$.

Since $q(x,t)$ and $xq(x,t) \in L_2^x(\mathbb{R}^+)$, the identity

$$\int_0^\infty q(x,t)dx = \int_0^\infty [(1+x)q]\frac{1}{(1+x)}dx$$

implies that $q(x,t) \in L_1^x(\mathbb{R}^+)$.

We shall now show that the assumption $x^2u(x) \in L_2(\mathbb{R}^+)$ implies that $x^2q(x,t) \in L_2^x(\mathbb{R}^+)$. Let

$$\psi(t) \doteq \left\|x^2qe^{-\frac{\varepsilon x}{2}}\right\|_{L_2^x(\mathbb{R}^+)}.$$

Similar considerations as above yield

$$\partial_t\psi^2 \leq \left(8+\frac{4}{e}\right)\int_0^\infty |xq_x|\left|x^2qe^{-\frac{\varepsilon x}{2}}\right|dx \leq \left(8+\frac{4}{e}\right)\|xq_x\|_{L_2^x(\mathbb{R}^+)}\psi.$$

Thus if $x^2q(x,0) \in L_2(\mathbb{R}^+)$, ψ exists, and since $x^2q(x,t) \in L_2^x(\mathbb{R}^+)$, it follows that $xq(x,t) \in L_1^x(\mathbb{R}^+)$. □

4. The asymptotic analysis. In order to determine the large-t behavior of the solution $q(x,t)$, one needs to study the large-t asymptotic behavior of the oscillatory RH problem (1.19). The corresponding problem for integrable equations on the full line was first studied in [8] (see also [9], [12], and the review [10]). A rigorous and elegant approach to studying the asymptotic behavior of RH problems has recently been developed in [11]. In what follows we shall use this new approach.

We first study the solution $\hat{Z}(x,t,k)$ of the regular RH problem corresponding to the RH problem (1.19) (see Proposition 2.4 (iii)).

THEOREM 4.1. *Under the assumptions of Proposition* 2.4, *the solution* $\hat{Z}(x,t,k)$ *of the regular RH problem corresponding to* (1.19) *satisfies the asymptotic equation*

$$\hat{Z}(x,t,k) = \left(\mathrm{I}+O\left(t^{-\frac{1}{2}}\right)\right)(\delta(k))^{\sigma_3}, \quad t\to\infty, \quad 0<A\leq\frac{x}{t}\leq B<\infty, \tag{4.1}$$

uniformly for $|Imk| \geq \varepsilon > 0$. *The scalar function* $\delta(k)$ *is given by*

$$\delta(k) = \exp\left[\frac{1}{2\pi i}\int_{-\infty}^{k_0}\frac{\ln(1-\lambda|b(k')-\lambda\overline{c(k')}|^2)}{k'-k}dk'\right], \quad k_0 = -\frac{x}{4t}. \tag{4.2}$$

Proof. The method of [11] can be thought of as a nonlinear steepest-descent method. The stationary point and the directions of the steepest descent associated with $\exp 2i(kx+2k^2t)$ are given by $k_0 = -x/4t$ and $Im(i(k-k_0)^2) = 0$, respectively. This implies that we must deform the original RH problem to one defined on the above steepest-descent contours (see the solid lines of Figure 4.1). We now discuss how this deformation can be achieved.

The jump matrix along the positive real axis can be factorized into two triangular matrices. Using this factorization, we find

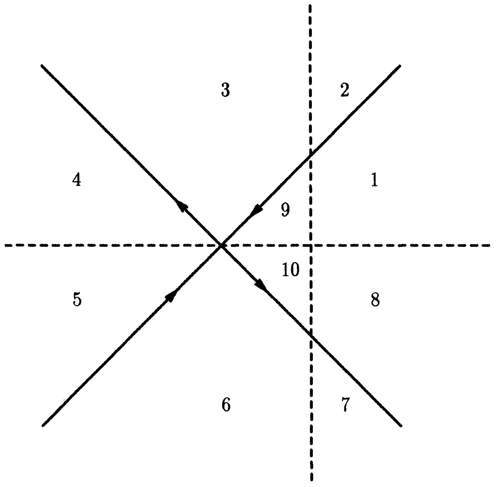


FIG. 4.1.

$$\hat{Z}^- \begin{pmatrix} 1 & b_0(k)e^{-\theta} \\ 0 & 1 \end{pmatrix} = \hat{Z}^+ \begin{pmatrix} 1 & 0 \\ \lambda\overline{b_0(k)}e^{\theta} & 1 \end{pmatrix}, \quad k \in \mathbb{R}^+. \tag{4.3}$$

Suppose that $b_0(k)$ is a rational function with appropriately chosen poles such that $b_0(k)$ is analytic in region 8. Since $e^{-\theta}$ and e^{θ} are analytic and decreasing in regions 8 and 1, respectively, it follows that the relevant jump matrices can be absorbed into $\hat{Z}^-$ and $\hat{Z}^+$. In this way, one can eliminate the jump along $\mathbb{R}^+$. Let c_0 and

$$r_0(k) \div b_0(k) - \lambda\overline{c_0(k)}, \tag{4.4}$$

be appropriate rational functions. Then regions 9 and 10 can be handled without difficulty since the factorization of the jump matrix for $k_0 < k < 0$ still has the "correct" triangularity. Also, the jumps along the broken lines separating regions 1 and 9 and regions 8 and 10 disappear. However, the factorization of the jump matrix for $k < k_0$ has the "wrong" triangularity,

$$\hat{Z}^- = \hat{Z}^+ \begin{pmatrix} 1 & r_0 e^{-\theta} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\lambda\bar{r}_0 e^{\theta} & 1 \end{pmatrix}, \tag{4.5}$$

i.e., the jump matrix involving $e^{-\theta}$ cannot be absorbed into $\hat{Z}^+$. To overcome this problem, one introduces the function $\delta(k)$. This function is analytic in the complex k-plane cut along $(-\infty, k_0]$; along this cut, it satisfies the jump condition

$$\delta_+(k) = \delta_-(k)(1 - \lambda|b(k) - \lambda\bar{c}(k)|^2), \quad k \in (-\infty, k_0]. \tag{4.6}$$

This jump is precisely chosen by the requirement of reversing the triangularity of the jump along $(-\infty, k_0)$. Indeed, if

$$\hat{W}(x,t,k) = \hat{Z}(x,t,k)(\delta(k))^{-\sigma_3} \tag{4.7}$$

and if G denotes the jump matrices associated with $\hat{Z}$, then the jump matrices associated with $\hat{W}$ are given by $G_0 = \delta^{\sigma_3} G \delta^{-\sigma_3}$. The jump matrices of the RH problem

$$\hat{W}^-(x,t,k) = \hat{W}^+(x,t,k)G_0(x,t,k), \quad k^2 \in \mathbb{R}, \quad \hat{W} \to \mathrm{I}, \quad k \to \infty, \tag{4.8}$$

possess the correct triangular factorizations:

$$G_0(x,t,k) = \begin{pmatrix} 1 & 0 \\ \lambda \bar{b}_0(k)\delta^{-2}(k)e^{\theta} & 1 \end{pmatrix} \begin{pmatrix} 1 & -b_0(k)\delta^2(k)e^{-\theta} \\ 0 & 1 \end{pmatrix}, \quad k \in \mathbb{R}^+,$$

$$G_0(x,t,k) = \begin{pmatrix} 1 & r_0(k)\delta^2(k)e^{-\theta} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\lambda \bar{r}_0(k)\delta^{-2}e^{\theta} & 1 \end{pmatrix}, \quad k \in [k_0, 0],$$

$$G_0(x,t,k) = \begin{pmatrix} 1 & 0 \\ \frac{-\lambda \overline{r_0(k)}\delta_-^{-2}(k)}{1-\lambda|r_0(k)|^2}e^{\theta} & 1 \end{pmatrix} \begin{pmatrix} 1 & \frac{r_0(k)}{1-\lambda|r_0(k)|^2}\delta_+^2(k)e^{-\theta} \\ 0 & 1 \end{pmatrix}, \quad k < k_0,$$

where $\delta_\pm(k) = \delta(k \pm i0)$, $k \in \mathbb{R}$. The first two equations above follow from equations (4.3) and (4.5). For the derivation of the third equation above, we use equation (4.6) and the fact that $|b(k) - \lambda\bar{c}(k)| = |r_0|$ (see the definitions of b_0 and c_0 in Proposition 2.4 (iii)).

The triangular factorizations of G_0 imply that the jumps along the real axis can be eliminated. The jump condition along the broken line separating regions 2 and 3 is given by

$$\hat{W}^- = \hat{W}^+ \begin{pmatrix} 1 & 0 \\ c_0\delta^{-2}e^{\theta} & 1 \end{pmatrix};$$

this shows that the jump matrix can be absorbed into $\hat{W}^+$. The proof continues similarly for the jump along the broken line separating regions 6 and 7. Hence, the function

$$X(x,t,k) = \hat{W}(x,t,k)K(x,t,k), \tag{4.9}$$

where $K(x,t,k)$ is given by

$$\begin{pmatrix} 1 & 0 \\ \lambda\overline{b_0(\bar{k})}\delta^{-2}(k)e^{\theta} & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ c_0(k)\delta^{-2}(k)e^{\theta} & 1 \end{pmatrix}, \ \mathrm{I}, \begin{pmatrix} 1 & -\frac{r_0(k)}{1-\lambda r_0(k)\overline{r_0(\bar{k})}}\delta^2(k)e^{-\theta} \\ 0 & 1 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 0 \\ -\lambda\frac{\overline{r_0(\bar{k})}}{1-\lambda r_0(k)\overline{r_0(\bar{k})}}\delta^{-2}(k)e^{\theta} & 1 \end{pmatrix}, \ \mathrm{I}, \ \begin{pmatrix} 1 & \lambda\overline{c_0(\bar{k})}\delta^2(k)e^{-\theta} \\ 0 & 1 \end{pmatrix},$$

$$\begin{pmatrix} 1 & b_0(k)\delta^2(k)e^{-\theta} \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ \lambda\overline{r_0(\bar{k})}\delta^{-2}(k)e^{\theta} & 1 \end{pmatrix}, \begin{pmatrix} 1 & r_0(k)\delta^2(k)e^{-\theta} \\ 0 & 1 \end{pmatrix}$$

for $k \in 1, 2, \ldots, 9, 10$ respectively is a sectionally holomorphic function satisfying the RH problem depicted in Figure 4.2 (also $X \to \mathrm{I}$ as $k \to \infty$). The RH problem for $X(x,t,k)$ has the crucial property that its jump matrices decay exponentially to the identity away from the stationary point k_0. This, just like the classical steepest-descent method, implies

$$X(x,t,k) = \mathrm{I} + O(t^{-\frac{1}{2}}) \tag{4.10}$$

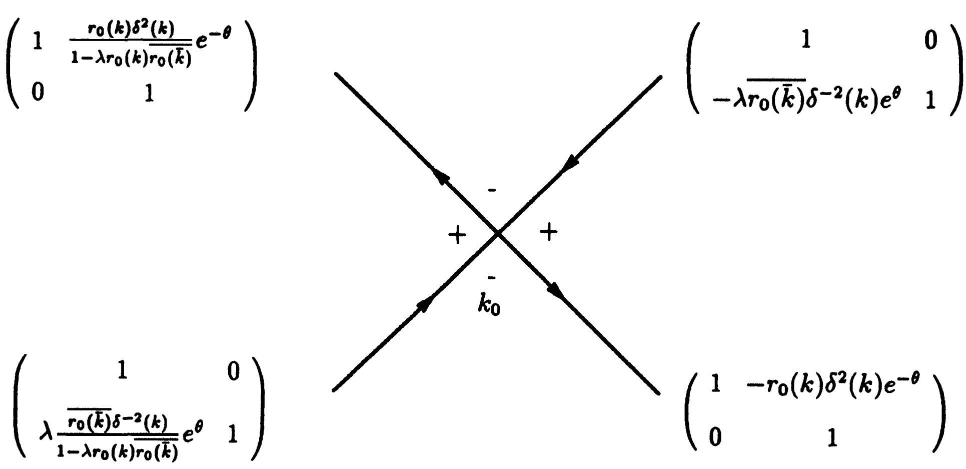

$\begin{pmatrix} 1 & \frac{r_0(k)\delta^2(k)}{1-\lambda r_0(k)\overline{r_0(\bar{k})}}e^{-\theta} \\ 0 & 1 \end{pmatrix}$
$\begin{pmatrix} 1 & 0 \\ -\lambda\overline{r_0(\bar{k})}\delta^{-2}(k)e^{\theta} & 1 \end{pmatrix}$
-
+
+
-
k_0
$\begin{pmatrix} 1 & 0 \\ \lambda\frac{\overline{r_0(\bar{k})}\delta^{-2}(k)}{1-\lambda r_0(k)\overline{r_0(\bar{k})}}e^{\theta} & 1 \end{pmatrix}$
$\begin{pmatrix} 1 & -r_0(k)\delta^2(k)e^{-\theta} \\ 0 & 1 \end{pmatrix}$

FIG. 4.2. *The deformed RH problem arising in the asymptotic analysis of the initial-boundary value problem of NLS. The x,t dependence enters only through $\theta(x,t) = 2i(kx + 2k^2t)$ and $\delta = \delta(k;k_0)$, $k_0 = -x/4t$.*

for k away from k_0. Indeed, if Γ and $\tilde{G}_0$ denote the contours and the jump matrices of Figure 4.2, then

$$X(x,t,k) = \mathrm{I} + \frac{1}{2\pi i}\int_\Gamma \frac{X^+(x,t,k')(\mathrm{I}-\tilde{G}_0(x,t,k'))}{k'-k}dk' \tag{4.11}$$

for k not on Γ. Taking into account the a priori estimate $|X^+(x,t,k)| \leq$ const and using the classical Laplace's method to the integral of the rhs of equation (4.11), one finds equation (4.10).

This concludes the proof of Theorem 4.1 in the case that b_0 and c_0 are rational. The general case can be reduced to this case following the construction of [11].

Remark 4.1. In the solitonless case, $\hat{Z}_p = \hat{Z}$, and Theorem 4.1 gives the asymptotic solution of the RH problem (1.19). Using equation (2.3.1b), which relates $q(x,t)$ to $\hat{Z}$, and the asymptotic relation (4.1), it follows that in the solitonless case,

$$q(x,t) = O(t^{-\frac{1}{2}}), \quad t \to \infty, \quad 0 \leq A \leq \frac{x}{t} \leq B. \tag{4.12}$$

We now consider the case that poles $\{k_j\}_1^N$ do exist. We assume $\lambda = -1$.

THEOREM 4.2. *Under the assumptions of Proposition* 2.4, *the solution $\hat{Z}_p(x,t,k)$ of the RH problem* (1.19) *satisfies the asymptotic equation*

$$\hat{Z}_p(x,t,k) = (\mathrm{I}+O(t^{-\frac{1}{2}}))\hat{Z}_S(x,t,k)(\delta(k))^{\sigma_3}, \quad t \to \infty, \quad 0 < A \leq \frac{x}{t} \leq B < \infty, \tag{4.13}$$

uniformly for $|Imk| \geq \varepsilon > 0$. The scalar function $\delta(k)$ is defined in equation (4.2). *The matrix function $\hat{Z}_S(x,t,k)$ is given by the algebraic equations* (2.4.2)–(2.4.4) *with $\lambda = -1$, $\hat{Z}$ replaced by* I, *and c_j replaced by $c_j(\delta(k_j))^{-2}$.*

Proof. Equation (4.13) is a direct consequence of Proposition 2.4, Theorem 4.1, and the equation $\overline{\delta(k_j)} = (\delta(\bar{k}_j))^{-1}$.

Remark 4.2. The function $\hat{Z}_S$ is nothing but the matrix eigenfunction corresponding to the pure N-soliton solution $q_s(x,t)$ of the NLS equation with parameters $\{k_j\}_1^N$ and $\{c_j(\delta(k_j))^{-2}\}_1^N$. This implies that

$$q(x,t) = q_s(x,t) + O(t^{-\frac{1}{2}}). \tag{4.14}$$

Using the well-known formula for the N-soliton solution of the NLS (or the algebraic system (2.4.3) directly), one can easily extract from (4.14) the asymptotics of $q(x,t)$ along the soliton rays, given in (1.4) and (1.5).

Remark 4.3. Just as in the classical steepest-descent method, one can improve estimate (4.10) by calculuating the contribution from the stationary point k_0 in closed form. It turns out [12] that the corresponding model RH problem can be solved explicitly in terms of parabolic cylindrical functions. The analysis is similar to that for the full-line problem presented in [10]. This leads to the following formula for the dispersive part of the asymptotics of $q(x,t)$:

(4.15)

$$q(x,t) = \alpha\left(-\frac{x}{4t}\right)t^{-1/2}\exp\left\{\frac{ix^2}{4t} + 2i\alpha^2\left(-\frac{x}{4t}\right)\log t + i\varphi_0\left(-\frac{x}{4t}\right)\right\} + 0(t^{-1/2}),$$

where the amplitude α and the phase φ_0 are given by

$$\alpha^2(k) = \frac{1}{4\pi}\log[1 + |b(k) + \overline{c(k)}|^2], \tag{4.16a}$$

$$\varphi_0(k) = 2\alpha^2(k)\log 2 + \frac{3\pi}{4} + \arg(b(k) + \overline{c(k)})$$

$$+\arg\Gamma(-2i\alpha^2(k)) + 4\int_{-\infty}^{k}\log|\mu - k|d\alpha^2(\mu) + 2\sum_{j=1}^{N}\arg(k_j - k)\mathrm{sign}(\xi_j - k). \tag{4.16b}$$

Remark 4.4. One can use equation (4.16a) to replace $\log[1 + |b(k) + \overline{c(k)}|^2]$ in the integral terms of equations (1.5) by $4\pi\alpha^2$. This shows that these terms indeed represent the interaction of solitons with the dispersive part.

Remark 4.5. In the case of the NLS on the full line, one can use equations (4.16) to solve b and c in terms of α and φ_0. However, in our case, this cannot be done since to define b and c, one needs two real functions defined on the whole linear (see Remark 2 in §2), while α and φ_0 are real functions defined in the half-line. This is a reflection of the fact that the information traveling towards the boundary is lost as $t \to \infty$. In the case of zero initial data, waves travel away from the boundary, no information is lost asymptotically, and the asymptotics of $q(x,t)$ can be used to recover c and hence $q(0,t)$.

Acknowledgments. The authors are grateful to L. Sung for many useful suggestions and to I. M. Gel'fand for his encouragement. The idea of using a combination of exact methods and of PDE techniques to solve initial-boundary value problems emerged after many discussions with J. Bona.

REFERENCES

[1] P. Lax, *Integrals of nonlinear equations of evolution and solitary waves*, Comm. Pure Appl. Math., 21 (1968), p. 467.

[2] J. Sheerin, J. C. Weatherall, D. R. Nicholson, G. L. Payne, M. V. Goldman, and P. J. Hansen, *Solitons and ionospheric modification*, J. Atmospheric Terr. Phys., 44 (1982), p. 1043.

[3] J. V. Moloney and A. C. Newell, *Theory of light-beam propagation at nonlinear interfaces*, Phys. Rev. A., 39 (1989), pp. 1809–1840.

[4] A. S. Fokas, *Initial-boundary value problems for soliton equations*, in Proc. III Potsdam–V Kiev International Workshop, Springer-Verlag, Berlin, 1992.

[5] A. S. Fokas and A. R. Its, *Soliton generation for initial-boundary value problems*, Phys. Rev. Lett., 68 (1992), p. 3117.

[6] V. Zakharov and A. Shabat, *Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media*, Soviet Phys. JETP, 34 (1972), p. 62.

[7] X. Zhou, *The Riemann–Hilbert problem and inverse scattering*, SIAM J. Math. Anal., 20 (1989), pp. 966–986.

[8] S. V. Manakov, *Nonlinear Fraunhofer diffraction*, Soviet Phys. JETP, 38 (1974), pp. 693–696.

[9] M. J. Ablowitz and H. Segur, *Asymptotic solution of the Korteweg–de Vries equation*, Stud. Appl. Math., 57 (1977), pp. 13–24.

[10] P. Deift, A. Its, and X. Zhou, *Long-time asymptotics for integrable nonlinear wave equations*, in Important Developments of Soliton Theory, A. S. Fokas and V. Zakharov, eds., Springer-Verlag, Berlin, 1993.

[11] P. Deift and X. Zhou, *A steepest descent method for oscillatory Riemann–Hilbert problems*, announcement in Bull. Amer. Math. Soc. (N.S.), 26 (1992), pp. 119–123; full paper in Ann. Math., 137 (1993), pp. 245–338.

[12] A. R. Its, *Asymptotics of solutions of the nonlinear Schrödinger equation and isomonodromic deformations of systems of linear differential equations*, Soviet Math. Dokl., 24 (1981), pp. 452–456.

[13] R. Caroll and C. Bu, *Solution of the forced NLS using PDE techniques*, Appl. Anal., 41 (1991), pp. 33–51.

[14] L. Sung, *Solution of the initial-boudnary value problem of NLS using PDE techniques*, preprint, Clarkson University, 1993.

[15] A. S. Fokas, *An initial-boundary value problem for the nonlinear Schrödinger equation*, Phys. D, 35 (1989), pp. 167–185.

[16] M. J. Ablowitz and H. Segur, *The inverse scattering transform: Semi-infinite interval*, J. Math. Phys., 16 (1975), p. 1054.

[17] E. K. Sklyanin, *Boundary conditions for integrable equations*, Funktsional Anal. i Prilozhen., 21 (1987), pp. 86–87 (in Russian); Functional Anal. Appl., 21 (1987), pp. 86-87 (in English).

[18] V. O. Tarasov, *An initial-boundary value problem for the nonlinear Schrödinger equation*, Zap. Nauchn. Sem. LOMI, 169 (1988), p. 151 (in Russian).

[19] R. F. Bikbaev and A. R. Its, *Algebro-geometric solutions of a boundary value problem for the nonlinear Schrödinger equation*, Mat. Zametki, 45 (1189), pp. 3–9 (in Russian); Math. Notes, 45 (1189), pp. 3–9 (in English).

[20] I. T. Habibullin, *Bäcklund transformations and integrable initial-boundary value problems*, in Nonlinear and Turbulent Processes, Vol. 1, World Scientific, Singapore, 1990 p. 259.

[21] J. Bona, W. G. Pritchard, and L. R. Scott, *An evaluation of a model equation for water waves*, Philos. Trans. Roy. Soc. London Ser. A, 302 (1981), pp. 457–510.

[22] J. Bona and R. Winther, *The Korteweg–de Vries equation, posed in a quarter-plane*, SIAM J. Math. Anal., 14 (1983), pp. 1056–1106.

[23] ———, *The Korteweg–de Vries equation in a quater-plane, continuous dependence result*, Differential Integral Equations, 2 (1989), pp. 228–250.

[24] R. L. Chou and C. K. Chu, *Solitons induced by boundary conditions from the Boussinesq equation*, Phys. Fluids A, 2 (1990), p. 1574.

[25] C. K. Chu and R. L. Chou, *Solitons induced by boundary conditions*, Adv. in Appl. Mech., 27 (1990), pp. 283–302.

[26] D. J. Kaup and P. Wycoff, *Time evolution of the scattering data for the forced Toda lattice*, Stud. Appl. Math., 81 (1989), pp. 7–20.

[27] E. S. Gutshabash and V. D. Lipovsky, *A boundary problem for the two dimensional elliptic sine-Gordon equation and its application to the theory of the stationary Josephson effect*, Zap. Nauchn. Sem. LOMI, 180 (1990), pp. 53–62.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 765–781, May 1996

009

STABILITY OF THE BUNSEN FLAME PROFILES IN THE KURAMOTO–SIVASHINSKY EQUATION*

DANIEL MICHELSON†

Abstract. The stability of the conical stationary solutions of the Kuramoto–Sivashinsky equation $u_t + \Delta^2 u + \Delta u + |\nabla u|^2 = c^2$ in one and two space dimensions is studied. It is shown that these solutions are unstable in the whole space. Next the problem is studied in the one-dimensional (1D) case in a bounded interval $|x| \leq l$ and in the 2D case in a disc $0 \leq r < l$ with natural boundary conditions. It is proved that for a large slope c the above stationary solutions are stable. In the 1D case part of the proof is computer assisted.

Key words. Bunsen flames, Kuramoto–Sivashinsky equation, computer-assisted proofs

AMS subject classifications. 34A34, 34A45, 35A50

1. Introduction.

It was shown in [2] and [3] that the Kuramoto–Sivashinsky equation

$$u_t + \Delta^2 u + \Delta u + |\nabla u|^2 = c^2, \quad u = u(x,t), \qquad x \in R^1 \quad \text{or} \quad x \in R^2 \tag{1.1}$$

possesses stationary conical solutions. In the context of combustion theory these solutions represent Bunsen flames on infinite linear or circular burners. From the mathematical point of view, in the one-dimensional case these are the solutions of the ordinary differential equation (ODE)

$$y''' + y' = c^2 - y^2, \quad y(x) = u'(x), \qquad -\infty < x < \infty, \tag{1.2}$$

which satisfy the boundary conditions

$$y(\pm\infty) = \mp c \tag{1.3}$$

and in the two-dimensional case these are the radial solutions $u(r)$ of the ODE

$$\left(\frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr}\right)\left(\frac{d}{dr} + \frac{1}{r}\right) y + \left(\frac{d}{dr} + \frac{1}{r}\right) y = c^2 - y^2, \qquad y(r) = u'(r), \tag{1.4}$$

with the boundary conditions

$$y(0) = y''(0) = 0, \qquad y(\infty) = -c. \tag{1.5}$$

In [2] it was proved analytically that for large c the problem (1.2), (1.3) has an odd solution. The result of [7] implies the uniqueness of a bounded solution of (1.2) for large c. Problem (1.4), (1.5) is apparently not amenable to analytical treatment. Nevertheless, the existence of solutions for (1.4), (1.5) was established in [3] for $0.27 \approx c_0 < c \leq \infty$ by a rigorous computer-assisted proof. The uniqueness for large c (for small c uniqueness is not expected) was not proved; however the computer program verified the following transversality condition.

Transversality condition. Let y_0 be a solution of (1.4), (1.5) such that

$$y_0(0) = y_0''(0) = 0 \tag{1.6}$$

*Received by the editors August 27, 1993; accepted for publication (in revised form) August 24, 1994.

†Department of Applied Mathematics and Computer Science, Weizmann Institute of Science, Rehovot 76100, Israel.

and let y be the solution of the linear problem

$$L[y_0]y = \left[\left(\frac{d^2}{dr} + \frac{1}{r}\frac{d}{dr}\right)\left(\frac{d}{dr} + \frac{1}{r}\right) + \left(\frac{d}{dr} + \frac{1}{r}\right) + 2y_0\right] y = 0, \tag{1.7}$$

$$y(0) = y''(0) = 0, \qquad y'(0) = 1. \tag{1.8}$$

Then, for large r the vector $\overline{y}(r) = (y(r), y'(r), y''(r))$ is transversal to the stable two-dimensional manifold $\mathcal{M}_2(r)$ of the flow defined by (1.7).

Note that the condition $y(0) = y''(0) = 0$ is imposed because we wish the corresponding u to be a solution of the partial differential equation in the plane. The above transversality condition was shown to hold for the computed radial solution $y_0(r)$ in the range $c_0 < c \leq \infty$. Our $y_0(r)$ is negative for $r > 0$. In [4] the computer also checked that for large c

$$y_0(r)/r + y_0'(r) < 0 \text{ for } r > 0. \tag{1.9}$$

We conjecture that the above condition defines y_0 uniquely. In the 1D case the transversality and the negativeness of y_0 for large c is proved analytically (e.g., see [6]).

Thus we take the existence of y_0 and the transversality condition for granted and pose the following question: *is the corresponding solution $u_0(x)$ stable in the sense of the evolutionary problem* (1.1) *with a fixed c*? Since in [4] it was shown that rotating solutions bifurcate from u_0, the answer should be negative. Actually, u_0 turns out to be linearly unstable also in the space of radial functions. The precise statement is as follows.

THEOREM 1. *For each s in the domain $\mathcal{D}$*

$$\mathcal{D} = \left\{ s \in \mathbb{C} | 0 < |\text{Ims}| < 2c, 0 < \text{Res} < \left(\frac{\text{Ims}}{2c}\right)^2 - \left(\frac{\text{Ims}}{2c}\right)^4 \right\} \tag{1.10}$$

the eigenvalue problem

$$\left(s + \Delta^2 + \Delta + 2u_0' \frac{\partial}{\partial r}\right) v = 0, \qquad v = v(x), \quad x \in R^2 \tag{1.11}$$

has radial exponentially decreasing solutions. In the 1D *case the corresponding function $v(x)$ depends on $x \in R^1$ and is even.*

As follows from the general theory of linear parabolic equations, the Cauchy problem for (1.1) when linearized at u_0 is well posed in a proper sense. Let the initial condition in (1.1) be $u = u_0 + \epsilon \text{Re } v$, where v is the eigenfunction in (1.11). Since the nonlinearity in (1.1) is weak, one can easily prove that the difference $u - u_0$ will grow exponentially in time for $t \leq \text{const} \cdot \log \epsilon^{-1}$.

Yet, numerical experiments show that the Bunsen flame profiles are stable for large c. The reason for it is the boundedness of the domain of x. Hence we consider (1.1) in a bounded domain

$$\Omega : 0 \leq r \leq l_0 \text{ in 2D} \quad \text{or} \quad |x| \leq l_0 \quad \text{in 1D}. \tag{1.12}$$

Let u_0 be the conical stationary solution of (1.1) in R^2 or R^1. We will impose the following boundary conditions:

$$u = u_0, \qquad \Delta u = \Delta u_0 \quad \text{at } \partial\Omega \tag{1.13}$$

or another pair

$$u = u_0, \qquad u_r = u_{0r} \quad \text{at } \partial\Omega. \tag{1.14}$$

In the 1D case Δu is replaced by u_{xx} and u_r by u_x. It is also possible to set similar boundary conditions that are independent of u_0

$$u = 0 \quad \text{and} \quad \Delta u = 0 \quad \text{at } \partial\Omega \tag{1.15}$$

or

$$u = 0 \quad \text{and} \quad u_r = -c \quad \text{at } \partial\Omega. \tag{1.16}$$

In the latter case one should first find a stationary radial solution $\tilde{u}_0$ that satisfies (1.15) or (1.16) and then study its stability. In the 1D case for large c we will consider more general nonlinear boundary conditions

$$S_\pm(\overline{u}_c(\pm l_0)) = 0, \tag{1.17}$$

where

$$\overline{u}_c(x) = c^{-2/3}(u(x), c^{-1/3}u'(x), c^{-2/3}u''(x), c^{-1}u'''(x)) \tag{1.18}$$

and $S_\pm : C^4 \to C^2$ are smooth mappings that satisfy

$$S_+(u, u', u'', u''') = S_-(u, -u', u'', -u''') \quad \text{and} \quad S_+(0, -1, 0, 0) = 0 \tag{1.19}$$

so that (1.17) is invariant under the transformation $x \to -x$. We will assume that the differential $dS_+[0, -1, 0, 0]$ of S_+ at $(0, -1, 0, 0)$ satisfies the Lopatinsky condition for the equation

$$\frac{d^4u}{dx^4} - \frac{2du}{dx} + su = 0, \qquad \text{Re}s \geq 0. \tag{1.20}$$

Namely, let λ_3, λ_4 be the two roots of the equation

$$\lambda^4 - 2\lambda + s = 0, \qquad \text{Re}s \geq 0 \tag{1.21}$$

in the half-plane Re $\lambda \geq 0$. Then we request that the vectors

$$dS_+[0, -1, 0, 0] \cdot (1, \lambda_i, \lambda_i^2, \lambda_i^3)^T, \qquad i = 3, 4 \tag{1.22}$$

are independent for all Re$s \geq 0$.

Concerning the existence of a stationary solution we will prove the following result in §3.

THEOREM 2. *Under the transversality condition,* (1.1) *with boundary conditions* (1.15) *or* (1.16) *and large* $l = l_0c^{1/3}$ *has a stationary solution* $\tilde{u}_0$ *close to* u_0. *The same result applies to the* 1D *problem with boundary condition* (1.17) *and large* c *and* l.

Concerning the stability of the stationary solution we will prove the following main result.

THEOREM 3. *For each* $l_0 > 0$ *there exists* $c^*(l_0)$ *such that for* $c > c^*(l_0)$ *the stationary radial solutions* $u_0(r)$ *and* $\tilde{u}_0(r)$ *are asymptotically stable in the sense of the corresponding evolutionary problem for* (1.1), *with boundary conditions* (1.13), (1.14) *and* (1.15), (1.16), *respectively. In the* 1D *case the same result holds for the* *boundary condition* (1.17) *and* $c > c^*, l_0 c^{1/3} > l^*$, *where* c^*, l^* *are some constants independent of* l_0.

The stability of the stationary solution for large c is linked to the eigenvalue problem

$$\left(s + \Delta^2 + 2y_0 \frac{\partial}{\partial r}\right) v = 0, \quad v = v(x), \quad x \in R^2 \text{ or } x \in R^1, \tag{1.23}$$

where y_0 is the solution of the equation

$$\left(\frac{d^2}{dr} + \frac{1}{r}\frac{d}{dr}\right)\left(\frac{d}{dr} + \frac{1}{r}\right) y = 1 - y^2, \qquad y(0) = y''(0) = 0, \quad y(\infty) = -1 \tag{1.24}$$

and in the 1D case, the solution of

$$y''' = 1 - y^2, \quad y(0) = y''(0) = 0, \quad y(\infty) = -1. \tag{1.25}$$

In the 2D case, (1.23) has no exponentially decreasing solutions as $|x| \to \infty$ with $\mathrm{Re}s \geq 0$. This follows from the property (1.9) of y_0. In the 1D case we would need $y_0'(x) < 0$ for $x > 0$ but this is not the case. Therefore we need the assistance of a computer. The proof is carried out in interval arithmetic and is completely rigorous module human or hardware errors. In §5 we show how to restrict problem (1.23) to a finite domain of $s : 0 \leq \mathrm{Re}s \leq .164, 0 \leq \mathrm{Im}s \leq 1.6$. Due to its length (about 800 lines, not counting the interval arithmetic library) the computer program is not included, but it can be obtained from the author upon request.

2. Instability in the unbounded domain.

Since we are mainly concerned with the case of large c, it is convenient to rescale problem (1.1) as follows:

$$u_{\text{new}} = c^{2/3} u_{\text{old}}, \quad x_{\text{new}} = c^{1/3} x_{\text{old}}, \quad t_{\text{new}} = c^{4/3} t_{\text{old}}. \tag{2.1}$$

Then (1.1) becomes

$$u_t + \Delta^2 u + \alpha \Delta u + |\nabla u|^2 = 1, \qquad 0 \leq \alpha = c^{-2/3} \leq \alpha_0. \tag{2.2}$$

The corresponding eigenvalue problem is

$$sv + \Delta^2 v + \alpha \Delta v + 2u_0' v_r = 0, \tag{2.3}$$

where $y_0 = u_0'$ is the solution of

$$\begin{gathered}\left(\frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr}\right)\left(\frac{d}{dr} + \frac{1}{r}\right) y + \alpha \left(\frac{d}{dr} + \frac{1}{r}\right) y = 1 - y^2, \\ y(0) = y''(0), \quad y(\infty) = -1.\end{gathered} \tag{2.4}$$

In the 1D case the equation for y_0 is

$$y''' + \alpha y' = 1 - y^2, \quad y(0) = y''(0) = 0, \quad y(\infty) = -1. \tag{2.5}$$

As $x \to \infty$ the limiting characteristic equation is

$$s + \lambda^4 + \alpha\lambda^2 - 2\lambda = 0. \tag{2.6}$$

For large $|s|$ with Re$s > 0$, two of the roots of (2.6), say λ_1, λ_2, have negative real parts and two, λ_3, λ_4, have Re $\lambda > 0$. If $\lambda = i\omega$ is imaginary, then s lies on the line

$$s = -\omega^4 + \alpha\omega^2 + 2i\omega \tag{2.7}$$

and for $0 < |\omega| < \alpha^{1/2}$ the corresponding s has Re$s > 0$. Denote by $\mathcal{D}$ the domain of s bounded by the above line and Re$s = 0$,

$$\mathcal{D} = \left\{ s \in \mathbb{C} | 0 < |\text{Im} s| < 2\alpha^{1/2}, 0 < \text{Re} s < \alpha \left(\frac{\text{Im} s}{2}\right)^2 - \left(\frac{\text{Im} s}{2}\right)^4 \right\}. \tag{2.8}$$

This is nothing but the domain $\mathcal{D}$ defined at (1.10) in the transformed variables. It is easy to see that for small $s \in \mathcal{D}$, (2.6) has three roots, $\lambda_1, \lambda_2, \lambda_3$ with Re $\lambda < 0$ and one λ_4 with Re $\lambda > 0$. By continuity this result holds for all $s \in \mathcal{D}$ while for s outside $\mathcal{D}$ and with Re$s > 0$, (2.6) has two roots λ_1, λ_2 with negative and two roots λ_3, λ_4 with positive real parts. Now, fix $s \in \mathcal{D}$. First consider the 2D case. Equation (2.3) for a radial function $v(r)$ becomes

$$sv + \left(\frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr}\right)^2 v + \alpha \left(\frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr}\right) v + 2y_0 \frac{dv}{dr} = 0 \tag{2.9}$$

with initial conditions

$$v'(0) = v'''(0) = 0. \tag{2.10}$$

The function y_0 has an asymptotic expansion

$$y_0 \sim -1 + \sum_{i=1}^{\infty} b_i r^{-i}, \qquad b_1 = -\alpha/2, b_2 = \alpha^2/8 \tag{2.11}$$

(see [3, (3.21)]). In the case when all roots λ_i of (2.7) are distinct and have different real parts, (2.10) has four independent solutions v_i with asymptotics

$$v_i(r) \sim e^{\lambda_i r} \sum_{n=0}^{\infty} a_{in} r^{-n}. \tag{2.12}$$

In our case λ_4 is always distinct and has a distinct real part, but $\lambda_1, \lambda_2, \lambda_3$ may have multiple real parts, e.g., Re λ_1 = Re λ_2 for Re$s = 0$. Still, one can choose four independent solutions such that v_1, v_2, v_3 are exponentially decreasing while v_4 is exponentially increasing (actually, v_4 is as in (2.12)). In the neighborhood of $r = 0$, $y_0(r)$ could be expanded into converging power series with odd powers of r (see [3]). This implies (see [4]) that the solution v of (2.9), (2.10) in a neighborhood of $r = 0$ could be expanded into a series

$$v(r) = \sum_{n=0}^{\infty} a_n r^{2n}, \tag{2.13}$$

where a_0 and a_1 are free parameters. We can choose two independent solutions φ_1 and φ_2 by selecting pairs $a_0 = 1, a_1 = 0$, or $a_0 = 0, a_1 = 1$. Since $\varphi_i = \sum_{j=1}^{4} c_{ij} v_j, i = 1, 2$ we can select a linear combination $\varphi = k_1\varphi_1 + k_2\varphi_2$ such that φ is a linear combination of v_1, v_2, v_3 only. Thus, for any $s \in \mathcal{D}$ we obtain a solution $v = \varphi$ of (2.9), (2.10) that decays exponentially at infinity.

In the 1D case y_0 approaches -1 along the stable 2D manifold defined by (2.5) (see [2]) and has the asymptotics

$$y_0(x) = -1 + O(e^{x\mathrm{Re}\lambda_{10}}), \tag{2.14}$$

where λ_{10} is the root of the equation

$$\lambda^3 + \alpha\lambda - 2 - 0 \tag{2.15}$$

with Re $\lambda < 0$. Again we have four independent solutions $v_i(x)$ of the equation

$$sv + \frac{d^4v}{dx^4} + \alpha\frac{d^2v}{dx^2} + 2y_0\frac{dv}{dx} = 0 \tag{2.16}$$

with v_4 exponentially increasing and v_1, v_2, v_3 exponentially decreasing. The solutions φ_1, φ_2 that satisfy (2.10) will be defined by initial conditions $\varphi_1(0) = 1, \varphi_1''(0) = 0$ and $\varphi_2(0) = 0, \varphi_2''(0) = 1$. Again we obtain a linear combination $\varphi = k_1\varphi_1 + k_2\varphi_2$ that decays exponentially as $x \to \infty$. This completes the proof of Theorem 1.

3. The existence of a stationary solution in a bounded domain. Let us consider a more general situation. Given a system of ODEs in R^n

$$\frac{dx}{dt} = f(x, t^{-1}), \qquad t_0 \le t < \infty, \tag{3.1}$$

where f depends smoothly on x and t^{-1}, and

$$f(0,0) = 0, \quad d_x f(0,0) = \begin{pmatrix} \lambda_1 & 0 \\ 0 & M_2 \end{pmatrix}, \qquad \lambda_1 > 0, \text{ Re } M_2 < 0. \tag{3.2}$$

Let $x_0(t)$ be a solution of (3.1) that tends to 0 as $t \to \infty$. We are looking for a solution of the problem

$$x' = f(x, t^{-1}), \quad t_0 \le t \le t_1, \quad x(t_0) \in \gamma, \quad S(x(t_1), t_1^{-1}) = 0, \tag{3.3}$$

where $\gamma = \gamma(p)$ is a smooth curve in R^n with $\gamma(0) = x_0(t_0)$ and $S : R^n \times R^1 \to R^1$ is a smooth map in a neighborhood of zero with $S(0) = 0$. It is assumed that $\gamma'(0)$ is transversal to the stable manifold $\mathcal{M}_{n-1}(t)$ of the system (3.1) at $t = t_0$ and that $d_x S(0,0)$ does not vanish on the eigenvector $e_1 = (1, 0, \dots, 0)^T$ of the matrix $d_x f(0,0)$.

LEMMA 3.1. *Under the above conditions, problem* (3.3) *for a sufficiently large* t_1 *has a solution* $x(t)$ *such that*

$$|x(t) - x_0(t)| < K\left(\frac{t}{t_1}\right)^m e^{\lambda_1(t-t_1)}|S(x_0(t_1), t_1^{-1})|, \tag{3.4}$$

where K *does not depend on* t_1 *and* m *is a constant (to be defined later).*

The proof is based on a standard normalization procedure for the solution map $\varphi[t_0, t_1] : x(t_0) \to x(t_1)$ in a neighborhood of $x_0(t_0)$. Let us partition vector x as $x = (x_1, x_2), x_1 \in R^1, x_2 \in R^{n-1}$. Without loss of generality we may assume that the stable manifold $\mathcal{M}_{n-1}(t)$ is given by the equation $x_1 = 0$. Then the differential $A(t^{-1}) = d_x f(x_0(t), t^{-1})$ has the block triangular form

$$A(t^{-1}) = \begin{pmatrix} a_1 & 0 \\ A_{21} & A_{22} \end{pmatrix}. \tag{3.5}$$

Since the function $x_0(t)$ has an asymptotic expansion

$$x_0(t) = c_1 t^{-1} + c_2 t^{-2} + \cdots = c_1 t^{-1} + O(t^{-2}), \tag{3.6}$$

the entries of $A(t^{-1})$ are

$$a_1 = \lambda_1 + m t^{-1} + O(t^{-2}), \quad A_{21} = O(t^{-1}), \quad A_{22} = M_2 + O(t^{-1}). \tag{3.7}$$

Now normalize the variables

$$x(t) = x_0(t) + \tilde{x}(t)\rho(t, t_1), \qquad \rho(t, t_1) = e^{\lambda_1(t-t_1)} \left(\frac{t}{t_1}\right)^m. \tag{3.8}$$

System (3.1) becomes

$$\tilde{x}'(t) = \tilde{A}(t^{-1})\tilde{x} + \rho(t, t_1)O(|\tilde{x}|^2), \tag{3.9}$$

where $\tilde{A}$ has the same form as in (3.5) but with

$$\tilde{a}_1 = O(t^{-2}) \quad \text{and} \quad \tilde{A}_{22} = \tilde{M}_2 + O(t^{-1}) = M_2 - \lambda_1 I + O(t^{-1}). \tag{3.10}$$

Solve (3.9) on an interval $t \in [t^*, t_1]$ with given $\tilde{x}(t^*)$. Let us represent the solution as

$$\tilde{x}(t) = e^{\tilde{A}(0)(t-t^*)}\tilde{x}(t^*) + \Delta\tilde{x}(t), \qquad \Delta\tilde{x}(t^*) = 0. \tag{3.11}$$

Then $\Delta\tilde{x}$ satisfies

$$\begin{aligned} \Delta\tilde{x}' = {} & \tilde{A}(t^{-1})\Delta\tilde{x} + (0, O(t^{-1}))^T \tilde{x}(t^*) + O(t^{-2})\tilde{x}(t^*) \\ & + \rho(t, t_1)O(|\tilde{x}|^2) = \tilde{A}(t^{-1})\Delta\tilde{x} + g. \end{aligned} \tag{3.12}$$

Note that

$$\operatorname{Re} \tilde{A}(t^{-1}) < \begin{pmatrix} O(t^{-2}) & 0 \\ 0 & -cI \end{pmatrix}. \tag{3.13}$$

Multiplication of (3.12) by $\Delta\tilde{x}$ and integration over the interval $I = [t^*, t]$ yields

$$\begin{aligned} \tfrac{1}{2}|\Delta\tilde{x}(t)|^2 \le {} & -c\|\Delta\tilde{x}_2\|_2^2 + K\|t^{-2}\|_1 \|\Delta\tilde{x}_1\|_\infty^2 + \|g_1\|_1 \|\Delta\tilde{x}_1\|_\infty \\ & + \|g_2\|_2 \|\Delta\tilde{x}_2\|_2, \end{aligned} \tag{3.14}$$

where $\|\cdot\|_p$ are the L_p norms on I and $g = (g_1, g_2)$ is partitioned according to x. Hence, for $t^* > (4K)^{-1}$,

$$\begin{aligned} \|\Delta\tilde{x}\|_\infty \le {} & K(\|g_1\|_1 + \|g_2\|_2) \le K((\|t^{-2}\|_1 + \|t^{-1}\|_2)|\tilde{x}(t^*)| \\ & + \|\rho(t, t_1)\|_1 (\|\Delta\tilde{x}\|_\infty + |\tilde{x}(t^*)|)^2). \end{aligned} \tag{3.15}$$

It is easy to see that the norm $\|\rho(t, t_1)\|_1$ is uniformly bounded for all $t_0 \le t^* < t_1 < \infty$ while $\|t^{-2}\|_1, \|t^{-1}\|_2 \to 0$ as $t^*, t_1 \to \infty$. Thus, for small $|\tilde{x}(t^*)| < \delta$, and large $t^* \ge t_0^*$

$$\|\Delta\tilde{x}\|_\infty \le K((t^*)^{-1/2}|\tilde{x}(t^*)| + |\tilde{x}(t^*)|^2) \tag{3.16}$$

uniformly in t_1. Now, if we differentiate (3.11), (3.12) with respect to the initial vector $\tilde{x}(t^*)$ and proceed as above, we obtain

$$\|\partial\Delta\tilde{x}/\partial\tilde{x}(t^*)\|_\infty \le K((t^*)^{-1/2} + |\tilde{x}(t^*)|). \tag{3.17}$$

Thus, for large t^* the map

$$\tilde{\varphi}[t^*,t] : \tilde{x}(t^*) \to \tilde{x}(t) \tag{3.18}$$

in the neighborhood of 0 approximates in the C^1 norm the linear map $\exp(\tilde{A}(0)(t-t^*))$ uniformly in the parameters $t^* \le t \le t_1 < \infty$.

Now return to the problem in (3.3). Rescale the curve $\gamma(p)$ as

$$\tilde{p} = \rho(t_0,t_1)^{-1}p, \qquad \tilde{\gamma}(\tilde{p}) = (\rho(t_0,t_1))^{-1}(\gamma(p)-\gamma(0)). \tag{3.19}$$

By the transversality assumption $\tilde{\gamma}'(0) = \gamma'(0)$ has a nonzero component $\tilde{\gamma}_1'(0)$. Without loss of generality we may assume that t_0 coincides with t_0^*. Denote by $\tilde{\gamma}(\tilde{p},t)$ the curve

$$\tilde{\gamma}(\tilde{p},t) = \tilde{\varphi}[t_0,t]\tilde{\gamma}(\tilde{p}), \qquad \tilde{\gamma} = (\tilde{\gamma}_1,\tilde{\gamma}_2). \tag{3.20}$$

It follows from (3.11) and (3.17) that

$$|\tilde{\gamma}_2'(\tilde{p},t)| \le c_1|\tilde{\gamma}_1'(\tilde{p},t)|, \qquad \tilde{\gamma}' = \partial\tilde{\gamma}/\partial\tilde{p} \tag{3.21}$$

for all $|\tilde{p}| < \delta_1$, where c_1 and δ_1 are independent of t and t_1. It follows then again from (3.11), (3.16), and (3.17) that for any $\epsilon > 0$ there exist δ_2 and t_2 such that for $t_2 \le t \le t_1$ and $|\tilde{p}| < \delta_2$

$$|\tilde{\gamma}_2'(\tilde{p},t)| < \epsilon|\tilde{\gamma}_1'(\tilde{p},t)|. \tag{3.22}$$

Since for large $t^*, d\tilde{\varphi}[t^*,t]$, when restricted to the first component, is close to identity, therefore

$$|\tilde{\gamma}_1'(\tilde{p},t)| > \delta_3 \quad \text{for all } |\tilde{p}| < \delta_2 \quad \text{and} \quad t_2 \le t \le t_1. \tag{3.23}$$

The problem in (3.3) could be rewritten as a single equation

$$S(x_0(t_1) + \tilde{\gamma}(\tilde{p},t_1), t_1^{-1}) = 0. \tag{3.24}$$

Recall that $x_0(t_1) = O(t_1^{-1})$. We can apply to (3.24) the implicit function theorem where $t_1^{-1} \approx 0$ serves as a parameter. Indeed,

$$\begin{aligned} &d_xS(x_0(t_1) + \tilde{\gamma}(\tilde{p},t_1), t_1^{-1})\tilde{\gamma}'(\tilde{p},t_1) \\ &\quad = (d_xS(0,0) + O(|\tilde{p}|,t_1^{-1}))\tilde{\gamma}_1'(\tilde{p},t_1)(e_1 + O(\epsilon)) \end{aligned} \tag{3.25}$$

is for small $|\tilde{p}|$ and t_1^{-1} of a constant sign and bounded away from zero. Hence, (3.24) has a locally unique solution

$$\tilde{p} = O^*(S(x_0(t_1), t_1^{-1})). \tag{3.26}$$

The corresponding trajectory $x(t) = x_0(t) + \rho(t,t_1)\tilde{\gamma}(\tilde{p},t)$ then satisfies estimate (3.4). □

The above lemma obviously applies to the equations in (2.4) and (2.5). The vector $x = (y,y',y'')^T, t = x$, or $t = r$. The curve $\gamma(p)$ is given by the vector $x(t_0)$ corresponding to the initial condition $y(0) = y''(0) = 0, y'(0) = p_0 + p$. In cases (1.15) and (1.16) the value of S is $(d/dr + r^{-1})y(r)$ and $y(r)+1$ correspondingly. The eigenvector e_1 corresponds to $x = (1,\lambda_3,\lambda_3^2)^T$ where $\lambda_3 = 2^{1/3}$ is the positive root of (1.21) for $s = 0$. Clearly the condition $d_x S(0,0)e_1 \neq 0$ is satisfied. In the case of the general boundary condition as in (1.17) one should eliminate the first component u, since $u = \int y\,dx$ is defined up to a constant. Since the Lopatinsky condition is assumed to hold also for $s = 0$, we obtain that the vectors $d_x S_+ \cdot (1,0,0,0)^T$ and $d_x S_+ \cdot (1, \lambda_3, \lambda_3^2, \lambda_3^3)^T$ are independent. Let ξ be a row vector orthogonal to $d_x S_+(1,0,0,0)^T$. Then the boundary operator $S = \xi \cdot S_+$ will satisfy the conditions of the lemma. Finally, note that all estimates of Lemma 3.1 apply uniformly to (2.4) and (2.5) for all $0 \le \alpha \le \alpha_0$. In particular the lower limit for $t_1 = l = l_0 c^{1/3}$ could be fixed for all α as above.

In the radial case the value of $S(x_0(t_1), t_1^{-1})$ corresponding to y_0 as in (2.11) is

$$y_0'(l) + y_0(l)/l = -l^{-1} + O(l^{-2}) \tag{3.27}$$

in the case of (1.15) and

$$y_0(l) + 1 = -\frac{\alpha}{2} l^{-1} + O(l^{-2}) \tag{3.28}$$

in the case of (1.16). In the one-dimensional case, as follows from (2.14),

$$\xi \cdot S(0, y_0'(l), y_0''(l), y_0'''(l)) = O(e^{l \mathrm{Re} \lambda_{10}}). \tag{3.29}$$

By (3.4) the derivative $\tilde{y}_0 = \tilde{u}_0'$ of the stationary solution $\tilde{u}_0$ will satisfy in the radial case

$$|\overline{\tilde{y}}_0(r) - \overline{y}_0(r)| < K l^{-1} e^{-(\lambda_{30} - \delta)(l-r)}, \qquad 0 \le r \le l, \tag{3.30}$$

where λ_{30} is the positive root of (2.15) and $\delta > 0$ is arbitrarily small while K depends on δ but not on l and α. Here and elsewhere we denote by $\overline{y}$ the vector

$$\overline{y} = (y, y', y''). \tag{3.31}$$

In the 1D case

$$|\overline{\tilde{y}}_0(x) - \overline{y}_0(x)| < K e^{l Re \lambda_{10}} \cdot e^{-(\lambda_{30} - \delta)(l-r)}, \qquad 0 \le x \le l. \tag{3.32}$$

If we make $\tilde{u}_0(r)$ and $u_0(r)$ coincide at $r = l$, then the difference $\tilde{u}_0(r) - u_0(r) = \int_l^r (\overline{y}_0(r) - y_0(r))dr$ will satisfy the same estimates as in (3.30) and (3.32). This completes the proof of Theorem 2.

Recall that the function y_0 satisfies (1.9). In the next section we will need the same inequality for $\tilde{y}_0$.

LEMMA 3.2. *In the radial case the function $\tilde{y}_0$ for small α and large l satisfies*

$$\tilde{y}_0(r)/r + \tilde{y}_0'(r) \le 0 \quad \text{for } 0 \le r \le l. \tag{3.33}$$

Proof. First consider the case of the boundary condition (1.16). In view of (3.28) we can replace the constant K in (3.30) by $K(\alpha + l^{-1})$. Since $y_0'(r) + y_0(r)r^{-1}$ has the asymptotics $-r^{-1} + O(r^{-2})$ it follows that for r sufficiently large, say $r \ge r_0$ and small $\alpha + l^{-1}$, inequality (3.33) holds. Now, for $0 \le r \le r_0$,

$$|(\tilde{y}_0(r) - y_0(r))r^{-1} + (\tilde{y}_0(r) - y_0(r))'| \le 2 \sup_{0 \le \xi \le r} |\tilde{y}_0'(\xi) - y_0'(\xi)| \tag{3.34}$$

and is estimated by the right-hand side of (3.30). Since at the same time $y_0(r)r^{-1} + y_0'(r) < -\delta < 0$, for large l we obtain (3.33).

In the case of boundary condition (1.15), the operator $S = d/dr + r^{-1}$ when applied to $\tilde{y}_0$ vanishes at $r = l$. In the notation of Lemma 3.1 the vector $\bar{\tilde{y}}_0$ is represented by $x(t)$

$$x(t) = x_0(t) + \rho(t, t_1)(e^{\tilde{A}(0)(t-t_0)}\tilde{x}(t_0) + \Delta\tilde{x}(t)). \tag{3.35}$$

By (3.16) and the block form of $\tilde{A}$ in (3.10), it follows that for large $t_0, \Delta\tilde{x}(t)$ and $\Delta\tilde{x}'(t)$ are $o(\tilde{x}(t_0))$. Recall that by (3.21), $\tilde{x}(t_0) = O(\tilde{x}_1(t_0))$. Now apply to (3.35) the operator $S(\cdot, t^{-1})$. For $t_1 \gg t_0$ we obtain

$$S(x(t_1), t_1^{-1}) = S(x_0(t_1), t_1^{-1}) + d_xS(0,0)e_1\tilde{x}_1(t_0) + o(\tilde{x}_1(t_0)) = 0. \tag{3.36}$$

Here $S(x_0(t_1), t_1^{-1})$ equals $y_0(t_1)t_1^{-1} + y_0'(t_1) = -t_1^{-1} + O(t_1^{-2}) < 0$ and hence $d_xS(0,0)e_1\tilde{x}_1(t_0) = t_1^{-1}(1 + o(1)) > 0$. The derivative with respect to t

$$\begin{aligned} S(x(t), t^{-1})' = S(x_0(t), t^{-1})' &+ \rho'(t, t_1)(d_xS(0,0)e_1\tilde{x}_1(t_0) \\ &+ o(\tilde{x}(t_0))) + \rho(t, t_1)o(\tilde{x}(t_0)). \end{aligned} \tag{3.37}$$

Now, $S(x_0(t), t^{-1})' = t^{-2} + O(t^{-3}) > 0$, and

$$\rho'(t, t_1) = \rho(t, t_1)(\lambda_1 + mt^{-1}) > 0 \quad \text{for large } t. \tag{3.38}$$

Since $\rho(t, t_1) \sim \rho'(t, t_1)$ and $|\tilde{x}(t_0)| \sim |\tilde{x}_1(t_0)| \sim |d_xS(0,0)e_1\tilde{x}_1(t_0)|$, the above derivative is positive and hence $S(x(t), t^{-1}) < 0$. Thus (3.33) holds for $r_0 \le r \le l$ where r_0 is large but independent of l. The range of $0 \le r \le r_0$ is treated as before.

4. The stability of the radial stationary solution.

Let $\tilde{u}_0(r)$ be the radial solution found in the previous section. Namely, $\tilde{y}_0 = \tilde{u}_0'(r)$ is a solution of the equation in (2.4) with initial condition $\tilde{y}_0(0) = \tilde{y}_0''(0) = 0$ and boundary conditions $\tilde{y}_0'(l) + r^{-1}\tilde{y}_0(l) = 0$ in the case of (1.15) and $\tilde{y}_0(l) = -1$ in the case of (1.16). We assume that $\tilde{y}_0(r)$ satisfies estimate (3.33). We will study the eigenvalue problem

$$(sI + L)v = \left(s + \Delta^2 + \alpha\Delta + 2\tilde{y}_0\frac{\partial}{\partial r}\right)v = 0, \qquad v = v(x), \quad x \in \Omega, \tag{4.1}$$

where Ω is a disk $r \le l$, with boundary conditions

$$\Delta v = v = 0 \text{ at } \partial\Omega \tag{4.2}$$

or

$$v_r = v = 0 \text{ at } \partial\Omega. \tag{4.3}$$

Denote by $\mathcal{L}$ the unbounded operator L that acts in the space of functions that satisfy the boundary conditions (4.2) or (4.3). Our main result is the following theorem.

THEOREM 4.1. *For $\alpha = c^{-2/3}$ and $l = l_0c^{1/3}$ with l_0 and c as in Theorem* 3 *the spectrum $\sigma(-\mathcal{L})$ lies in the half plane* Res $< -\delta_0$, *where $\delta_0 > 0$ depends on l_0 but not on $c > c^*$.*

Note that (4.1), (4.2) and (4.1), (4.3) are well-posed elliptic problems that satisfy the Lopatinsky condition. The adjoint problem is

$$(\bar{s}I + \bar{L})v = \left(\bar{s} + \Delta^2 + \alpha\Delta - r^{-1}\frac{\partial}{\partial r}(\tilde{y}_0r)\cdot\right)v = 0 \tag{4.4}$$

with the same boundary conditions as for the direct problem. Hence we should only check that both the direct and adjoint problems have no eigenvalues with $\mathrm{Re}s \geq 0$. Moreover, since for large positive s both operators $sI + \mathcal{L}$ and $(sI + \mathcal{L})^*$ have zero kernel, the index of $sI + \mathcal{L}$ is zero and, by continuity, it remains zero for all s. Hence it is enough to show that problems (4.1), (4.2) and (4.1), (4.3) have no nontrivial smooth solutions with $\mathrm{Re}s \geq 0$.

For that sake multiply (4.1) by v in $L_2(\Omega)$ and take real part. We obtain

$$\mathrm{Re}s\ \|v\|^2 + \|\Delta v\|^2 - \alpha\|\nabla v\|^2 + \int_\Omega (-\tilde{y}_0 r^{-1} - \tilde{y}_0')|v|^2 = 0, \tag{4.5}$$

where $\|\cdot\|$ is the usual $L_2(\Omega)$ norm. The imaginary part of the product yields

$$|\mathrm{Im}s|\|v\|^2 \leq 2\|\tilde{y}_0\|_\infty \|v\| \cdot \|\nabla v\| \leq \|\tilde{y}_0\|_\infty (b_1\|v\|^2 + b_1^{-1}\|\nabla v\|^2). \tag{4.6}$$

Multiply both sides of (4.6) by b_2 and add to (4.5). We obtain

$$\begin{aligned}(\mathrm{Re}s + b_2|\mathrm{Im}s| - b_2 b_1\|\tilde{y}_0\|_\infty)\|v\|^2 - (\alpha + b_2 b_1^{-1}\|\tilde{y}_0\|_\infty)\|\nabla v\|^2 \\ + \int_\Omega (-\tilde{y}_0 r^{-1} - \tilde{y}_0')|v|^2 + \|\Delta v\|^2 \leq 0.\end{aligned} \tag{4.7}$$

Recall that by Lemma 3.2 the last integral is positive. Since our problem does not depend on the polar angle φ we may assume that

$$v(x) = e^{in\varphi} v_n(r). \tag{4.8}$$

Recall that v vanishes on $\partial\Omega$. Hence

$$\|v\|^2 \leq (l/\lambda_{n,1})^2 \|\nabla v_n\|^2 \leq (l/\lambda_{n,1})^4 \|\Delta v_n\|^2, \tag{4.9}$$

where $\lambda_{n,1}$ is the first zero of the nth Bessel function $J_n(r)$. Therefore

$$\alpha\|\nabla v\|^2 \leq \alpha l^2 \lambda_{n,1}^{-2} \|\Delta v\|^2 = l_0^2 \lambda_{n,1}^{-2} \|\Delta v\|^2 < \tfrac{1}{2}\|\Delta v\|^2 \tag{4.10}$$

provided

$$n \approx \lambda_{n,1} > l_0\sqrt{2}. \tag{4.11}$$

Thus (4.5) implies that $v = 0$ for n as above and any s with $\mathrm{Re}s \geq 0$.

Now let b_1 and b_2 in (4.7) be $b_1 = \alpha^{1/2}, b_2 = \alpha^{3/2}/\|\tilde{y}_0\|_\infty$. Then we obtain

$$(\mathrm{Re}s + \alpha^{3/2}|\mathrm{Im}s|/\|\tilde{y}_0\|_\infty - \alpha^2)\|v\|^2 - 2\alpha\|\nabla v\|^2 + \|\Delta v\|^2 \leq 0. \tag{4.12}$$

But if

$$\mathrm{Re}s + \alpha^{3/2}|\mathrm{Im}s|/\|\tilde{y}_0\|_\infty > 2\alpha^2 \tag{4.13}$$

the left-hand side of (4.12) is positive. Thus it remains to study the eigenvalue problem (4.1) for v as in (4.8) with bounded

$$n \leq n_0 \approx l_0\sqrt{2} \tag{4.14}$$

and s in a neighborhood of zero. For this sake we will need the following elementary result.

LEMMA 4.1. *Given a system of linear ODEs*

$$x'(t) = A(t,p)x(t), \quad x(t) \in \mathbb{C}^n, \quad t_0 \leq t < \infty \tag{4.15}$$

$A(t,p)$ has a form

$$A(t,p) = A^{(0)}(p) + \Delta A(t,p), \tag{4.16}$$

where $A^{(0)}(p), \Delta A(t,p)$ depends continuously on a vector parameter p at $p = 0$ and $\|\Delta A(t,p)\| \to 0$ as $t \to \infty$ uniformly in p. Let $A^{(0)}(p)$ have the block form

$$A^{(0)}(p) = A_I^{(0)}(p) \oplus A_{II}^{(0)}(p), \tag{4.17}$$

where

$$\operatorname{Re} A_I^{(0)}(p) \leq \mu_1 < \mu_2 \leq \operatorname{Re} A_{II}^{(0)}(p) \tag{4.18}$$

and partition the vector $x = \binom{x_I}{x_{II}}$ according to the above blocks. Then, for any ϵ_1, ϵ_2 there exists t_1 such that the space X of solutions of (4.15) *for $t \geq t_1$ splits into a direct sum $X = X_I \oplus X_{II}$ with the following properties:*

(i) *if $x \in X_I$ then*

$$|x_{II}(t)| < \epsilon_1 |x_I(t)|, \qquad |x_I(t)| < e^{(\mu_1+\epsilon_2)(t-t_1)}|x_I(t_1)|; \tag{4.19}$$

(ii) *if $x \in X_{II}$ then*

$$|x_I(t)| < \epsilon_1 |x_{II}(t)|, \qquad |x_{II}(t)| > e^{(\mu_2-\epsilon_2)(t-t_1)}|x_{II}(t_1)|; \tag{4.20}$$

(iii) *the restrictions $X_I(t), X_{II}(t)$ of X_I, X_{II} to any $t \in [t_1, \infty)$ depend continuously on p at $p = 0$.*

Proof. By substitution $x(t) \to x(t)e^{\mu t}$ we may assume that $\mu_1 < 0 < \mu_2$. Let $x(t)$ be a solution of (4.15) that tends to 0 as $t \to \infty$. Multiply both sides of (4.15) by Rx in $L_2[t_1, \infty)$ where $R = cI \oplus (-I)$ and take real part. We obtain

$$\frac{1}{2}|x_{II}(t_1)|^2 + \min(-c\mu_1, \mu_2)\|x\|_2^2 \leq \frac{c}{2}|x_I(t_1)|^2 + \|\Delta A\|_\infty \|x\|_2^2, \tag{4.21}$$

where $\|\cdot\|_p$ is the $L_p[t_1, \infty)$ norm. Take $c < \epsilon_1^2$ and t_1 large enough so that $\|\Delta A\|_\infty < \frac{1}{2}\min(-c\mu_1, \mu_2)$. Then $|x_{II}(t)| < \epsilon_1 |x_I(t)|$ for all $t \geq t_1$ and the solution $x(t)$ is uniquely determined by $x_I(t_1)$. The existence of a solution in $L_2[t_1, \infty)$ for any $x_I(t_1)$ could be proved by iteration scheme $(d/dt - A^{(0)}(p))x^{(n)} = \Delta A x^{(n-1)}$. The second inequality in (4.19) follows from the estimate

$$\begin{aligned} \frac{1}{2}\frac{d|x_I(t)|^2}{dt} &\leq \mu_1 |x_I(t)|^2 + |\Delta A||x_I(t)| \cdot |x(t)| \\ &\leq (\mu_1 + \|\Delta A\|_\infty (1 + \epsilon_1))|x_I(t)|^2 \end{aligned} \tag{4.22}$$

provided $\|\Delta A\|_\infty (1 + \epsilon_1) < \epsilon_2$. Thus X_I is identified with the space of the above solutions $x(t)$. Denote by $x(\cdot, p) \in X_I$ the solution of (4.15) with given $x_I(t_1) = a_I$ and consider the difference $\Delta x = x(\cdot, p) - x(\cdot, 0)$. Then Δx satisfies

$$(\Delta x)' - A(t,0)\Delta x = f = (A(t,p) - A(t,0))x(t,p). \tag{4.23}$$

By applying to (4.23) the same procedure as in (4.21) we prove that $|\Delta x_{II}(t_1)| \to 0$ as $\|A(\cdot, p) - A(\cdot, 0)\|_\infty \to 0$. Hence $X_I(t)$ depends continuously on $p \to 0$. The space X_{II}

could be taken as the space of solutions x of (4.15) with initial condition $x_I(t_1) = 0$ and arbitrary $x_{II}(t_1)$. Then

$$\frac{1}{2}(\epsilon_1^2|x_{II}(t)|^2 - |x_I(t)|^2)' \geq (-\mu_1)|x_I(t)|^2 + \epsilon_1^2\mu_2|x_{II}(t)|^2 - |\Delta A| \cdot |x(t)|^2 \geq 0 \tag{4.24}$$

provided $\|\Delta A\|_\infty < \min(-\mu_1, \epsilon_1^2\mu_2)$. Hence the estimate $|x_I(t)| < \epsilon_1|x_{II}(t)|$ is valid for all $t \geq t_1$. The second estimate in (4.20) follows as in (4.22). The continuous dependence of $X_{II}(t)$ on p follows from the definition of $X_{II}(t_1)$.

Now let us apply the lemma to our situation. Extend the function $\tilde{y}_0(r)$ to be equal $y_0(r)$ for $r > l$ and consider (4.1) for v as in (4.8) with n bounded as in (4.14). We obtain the equation

$$\left(s + \Delta_n^2 + \alpha\Delta_n + 2\tilde{y}_0\frac{d}{dr}\right)v_n = 0, \quad \Delta_n = \left(\frac{d}{dr}\right)^2 + r^{-1}\frac{d}{dr} + n^2r^{-2}. \tag{4.25}$$

Then the vector $\bar{v}_n = (v_n, v_n', v_n'', v_n''')^T$ corresponds to x and r to t. The small parameter $p = (s, \alpha, l^{-1})$. At $p = 0$ we obtain the equation

$$\left(\Delta_n^2 + 2y_0\frac{d}{dr}\right)v_n = 0. \tag{4.26}$$

The leading part $A^{(0)}(p)$ corresponds to the operator

$$s + \left(\frac{d}{dr}\right)^4 + \alpha\left(\frac{d}{dr}\right)^2 - 2\frac{d}{dr}. \tag{4.27}$$

At $p = 0$ the eigenvalues of $A^{(0)}$ are

$$\lambda_{1,2} = 2^{1/3}(-\tfrac{1}{2} \pm i\tfrac{\sqrt{3}}{2}), \quad \lambda_3 = 2^{1/3}, \quad \lambda_4 = 0.$$

We define

$$\mu_1 = -\tfrac{1}{2} \cdot 2^{1/3} + \delta, \qquad \mu_2 = -\delta. \tag{4.28}$$

Then, for small p, the condition in (4.18) is satisfied. By (3.30), ΔA also satisfies the conditions of Lemma 4.1. In the sequel we will follow the notation of Lemma 4.1.

Let X_I and X_{II} be defined as in Lemma 4.1 and $P_I(t,p), P_{II}(t,p)$ the projectors corresponding to the direct sum $\mathbb{C}^4 = X_I(t,p) \oplus X_{II}(t,p)$. Denote by $X_0(t,p)$ the 2D space of vectors $x(r)$ that correspond to the solutions of (4.25) with initial conditions (2.10). Clearly $X_0(t,p)$ depends continuously on $p \to 0$. Note that $X_0(t,0) \cap X_I(t,0) = 0$. Indeed, otherwise (4.26) would have a nontrivial solution that decreases exponentially as $t \to \infty$. This would violate estimate (4.5) with $s = \alpha = 0$ and $\tilde{y}_0 = y_0$. By continuity, for a fixed t_1 and small $|p|$,

$$|P_{II}(t_1,p)x(t_1)| > \delta_1|x(0)|, \qquad |P_I(t_1,p)x(t_1)| < K_1|x(0)| \tag{4.29}$$

for all the above solutions $x \in X_0$. The vector functions P_Ix and $P_{II}x$ are also solutions of (4.15) and hence their components x_I and x_{II} satisfy estimates (4.19), (4.20), respectively. Now assume that $x(t)$ satisfies the boundary condition

$$Sx(l) = 0, \tag{4.30}$$

where S corresponds to (4.2) or (4.3). It is easy to see that S satisfies the Lopatinsky condition at infinity. Namely, S does not vanish on the vectors x with the zero component $x_I = 0$. (Recall that the decomposition of $x = (x_I, x_{II})$ corresponds to the eigenvalue decomposition λ_1, λ_2 and λ_3, λ_4.) Hence, for small $|p|$ and ϵ_1

$$|Sx| > \delta_2|x| \quad \text{for all } x \in X_{II}(l). \tag{4.31}$$

Now,

$$0 = Sx(l) = SP_I(l,p)x(l) + SP_{II}(l,p)x(l) \tag{4.32}$$

where by (4.19), (4.20), and (4.29)

$$\begin{aligned} |P_I(l,p)x(l)| &< |P_I(t_1,p)x(t_1)|e^{(\mu_1+\epsilon_2)(l-t_1)} \\ &\leq K|x(0)|e^{(\mu_1+\epsilon_2)(l-t_1)} \end{aligned} \tag{4.33}$$

and

$$\begin{aligned} |P_{II}(l,p)x(l)| &> |P_{II}(t_1,p)x(t_1)|e^{(\mu_2-\epsilon_2)(l-t_1)} \\ &> \delta_1|x(0)|e^{(\mu_2-\epsilon_2)(l-t_1)} \end{aligned} \tag{4.34}$$

and therefore

$$|SP_{II}(l,p)x(l)| > \delta_1\delta_2|x(0)|e^{(\mu_2-\epsilon_2)(l-t_1)}. \tag{4.35}$$

Since $\mu_2 - \epsilon_2 > \mu_1 + \epsilon_2$, for large l independent of x and p the second term in the right hand side of (4.32) will overweigh the first one. This completes the proof of Theorem 4.1. Clearly, our proof also applies to the case of the boundary conditions (1.13) and (1.14). One should merely replace $\tilde{y}_0$ everywhere by y_0.

Finally, let us remark that it was essential in our proof that the boundary condition is such that the energy estimate (4.5) holds. Thus we could restrict the problem to a finite number n_0 of ODEs. This is essential, since the above value r_1 where the connection between the inner and outer solutions is made grows very fast with n. The alternative is to restrict l and use (4.9) for $n > Kl$. But then the asymptotic expansion of Lemma 4.1 would not be applicable to $v_n(l)$.

5. The stability of the one-dimensional stationary solution. In the 1D case the integral in (4.5) is replaced by $\int(-y_0')|v|^2\,dx$. The graph of the function $-y_0'$ is displayed below in Fig. 1. Thus the integral is not positive. Actually our computation shows that it is negative for $v = x$. Since $y_0' \to 0$ exponentially as $|x| \to \infty$, the function $v = x$ could be modified so that $v \to 0$ exponentially as $x \to \infty$ and still

$$\|v''\|_2^2 + \int_{-\infty}^{\infty}(-y_0')|v|^2\,dx < 0. \tag{5.1}$$

Moreover, one can also construct an even function v that decreases exponentially as $|x| > \infty$ and for which the above quadratic form is negative. Our (noninterval) computations also show that if the norm $\|v''\|_2^2$ in (5.1) is increased by a factor $K \geq 1.3$ then the resulting quadratic form is positive in the space of even function in $H^2(R)$. We mention these results here without proof only to explain why the energy method does not work in the 1D case.

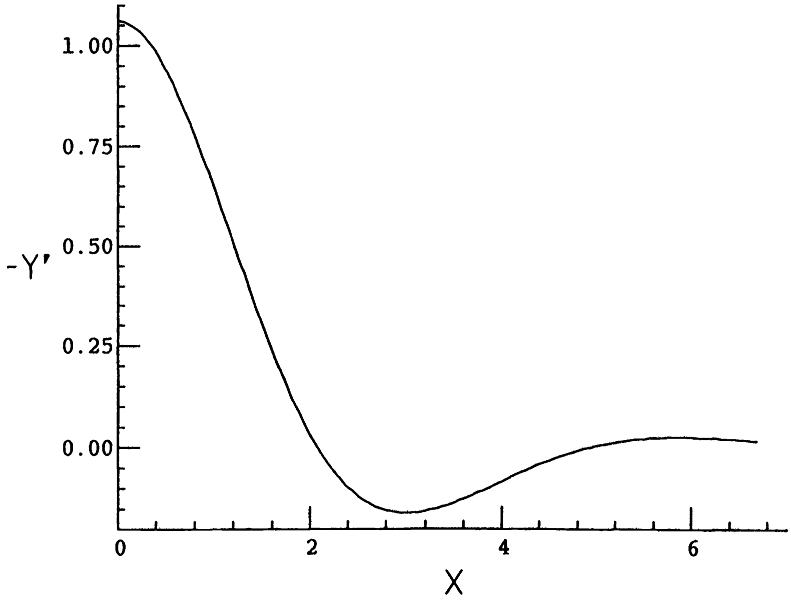


FIG. 1.

Recall that our rescaled eigenvalue problem is

$$L(s,\alpha)v = sv + v^{(4)} + \alpha v'' + 2\tilde{y}_0(x)v' = 0, \qquad |x| \le l \tag{5.2}$$

with boundary condition

$$dS_\pm[\bar{\tilde{u}}_0(l)]\bar{v}(l) = 0, \tag{5.3}$$

where $dS_\pm[\bar{\tilde{u}}_0(l)]$ are the differentials of $S_\pm$ at the points $\bar{\tilde{u}}_0(\pm l)$. Recall that $\tilde{y}_0(x)$ is an odd function of x. Since $S_\pm$ are related as in (1.19) it follows that $v(-x)$ is also a solution of (5.2), (5.3). Hence $v_1 = v(x) + v(-x)$ and $v_2 = v(x) - v(-x)$ are correspondingly even and odd solutions of (5.2), (5.3). Thus, without loss of generality we can consider the equivalent problems

$$L(s,\alpha)v = 0, \qquad 0 \le x \le l \tag{5.4}$$

with boundary condition at $x = l$

$$dS_+[\bar{\tilde{u}}_0(l)]\bar{v}(l) = 0 \tag{5.5}$$

and the boundary condition at $x = 0$

(5.6) (a) $v'(0) = v'''(0) = 0$

or

(b) $v(0) = v''(0) = 0.$

Since the problems (5.4)–(5.6) are uniformly Lopatinsky well posed for all large l and small α, it follows that for $|s| > K_0$ with K_0 independent of α and l the above problems have only trivial solutions. Now, for $|s| \leq K_0$ and small $\alpha + l^{-1}$ we wish to reduce these problems to the equation

$$sv + v^{(4)} + 2y_0 v' = 0, \qquad 0 \leq x < \infty, \tag{5.7}$$

with boundary conditions (5.6). Regarding the latter problem we make the following claim.

THEOREM 5.1. *Problem* (5.7), (5.6) *has no exponentially decreasing solutions for all* $s \in \mathbb{C}$ *with* $\operatorname{Re} s \geq 0$.

The reduction is carried out in the same way as in §4. The parameter $p = (s, \alpha, l^{-1})$, but the central point $p = 0$ is replaced by $p_0 = (s_0, 0, 0)$. The roots $\lambda_i(s)$ of (1.21) for $|s| < K_0$ and $\operatorname{Re} s \geq 0$ are separated: $\operatorname{Re}\lambda_{1,2}(s) < 0$ and $\operatorname{Re}\lambda_{3,4}(s) \geq 0$. Hence in a neighborhood of p_0 we may choose μ_1, μ_2 so that (4.18) holds. The constant matrix S in (4.30) is replaced by $dS_+[\overline{u}_0(l)]$. Still, since $dS_+[\overline{u}_0(\infty)]$ by assumption satisfies the Lopatinsky condition (see (1.22)), estimate (4.31) holds for large l. The reduction to problem (5.6), (5.7) then follows from (4.32)–(4.35).

Thus it remains to prove Theorem 5.1. Since $y_0(0) = 0$ and v satisfies the conditions in (5.6), we obtain the one-dimensional analog of (4.7). Namely

$$\begin{aligned}&(\operatorname{Re} s + b_2|\operatorname{Im} s| - b_2 b_1 \|y_0\|_\infty)\|v\|^2 - b_2 b_1^{-1}\|y_0\|_\infty \|v'\|^2 \\ &\quad + \int_0^\infty (-y_0')|v|^2\,dx + \|v''\|^2 \leq 0.\end{aligned} \tag{5.8}$$

Our computer program verified that

$$-y_0' > -\delta_0 = -0.164 \quad \text{and} \quad \|y_0\|_\infty < 1.23. \tag{5.9}$$

For the left-hand side of (5.8) to be positive definite it is sufficient that

$$\frac{1}{4}(b_2 b_1^{-1}\|y_0\|_\infty)^2 < \operatorname{Re} s - \delta_0 + b_2|\operatorname{Im} s| - b_2 b_1 \|y_0\|_\infty. \tag{5.10}$$

It is easy to check that the lowest bound on $\operatorname{Re} s$ is

$$\operatorname{Re} s > \delta_0 - \left(\frac{|\operatorname{Im} s|}{2\|y_0\|_\infty}\right)^4 \tag{5.11}$$

and is achieved when

$$b_1 = \frac{|\operatorname{Im} s|}{2\|y_0\|_\infty}, \qquad b_2 = \frac{|\operatorname{Im} s|^3}{4\|y_0\|_\infty^4}. \tag{5.12}$$

Since problem (5.7)–(5.6) is invariant under complex conjugation we may assume that $\operatorname{Im} s \geq 0$.

Thus, it remains to prove Theorem 5.1 in the domain

$$s \in \mathcal{D} : 0 \leq \operatorname{Re} s \leq \delta_0 = 0.164, \qquad 0 \leq \operatorname{Im} s \leq 2\|y_0\|_\infty \delta_0^{1/4} < 1.6. \tag{5.13}$$

We believe that this could be done by tedious hand computations in a way similar to [1]. Namely, one can expand y_0 and v into a power series near $x = 0$ that is valid up to $x \sim 4$ and to match it with exponential asymptotics of y_0 and v for large x. Instead we preferred a computer-based proof that involves millions of operations but requires that the reader only check a conceptually simple program.

The procedure of the proof is similar to the one in [3] and [4]. Unlike the real parameter α in [3] we have now a complex s. This forced us to modify all elementary interval arithmetic subroutines to include also the complex intervals. The domain $\mathcal{D}$ was subdivided into 40 small rectangles and for each of them the program ran separately. To manage rectangles in $\mathcal{D}$ of size $\delta \sim 0.08$ the dependence on s was expressed by Taylor formula of order 2. Also the estimates of the asymptotics of v for large x were carried out uniformly in s. Interested readers can obtain the program and the relevant analytical formulas and estimates from the author upon request.

REFERENCES

[1] N. Kopell and L. N. Howard, *Bifurcations and trajectories joining critical points*, Adv. Math., 18 (1975), pp. 306–358.

[2] D. Michelson, *Steady solutions of the Kuramoto–Sivashinsky equation*, Phys. D, 19 (1986), pp. 89–111.

[3] ———, *Bunsen flames as steady solutions of the Kuramoto–Sivashinsky equation*, SIAM J. Math. Anal., 23 (1991), pp. 364–386.

[4] ———, *Rotating Bunsen flames as solutions of the Kuramoto–Sivashinsky equation*, J. Dynamics Differential Equations, 5 (1993), pp. 375–416.

[5] ———, *Elementary particles as solutions of the Sivashinsky equation*, Phys. D, 44 (1990), pp. 502–556.

[6] ———, *Discrete shocks for difference approximations to systems of conservation laws*, Adv. Appl. Math., 5 (1984), pp. 433–469.

[7] C. K. McCord, *Uniqueness of connecting orbits in the equation $y^{(3)} = y^2 - 1$*, J. Math. Anal. Appl., 114 (1986), pp. 584–592.

[8] W. Troy, *The existence of steady solutions of the Kuramoto–Sivashinsky equation*, J. Differential Equations, 82 (1989), pp. 269–313.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 782–804, May 1996

ASYMPTOTIC EXPANSIONS WITH ERROR BOUNDS FOR THE COEFFICIENTS OF CAPACITY AND INDUCTION OF TWO SPHERES*

ANDREW H. VAN TUYL†

Abstract. Asymptotic expansions are obtained for the coefficients of capacity and induction of two spheres which hold as the distance ε between the spheres tends to zero, starting from expressions in terms of definite integrals obtained earlier [A. H. Van Tuyl, Electrostatic problems for two conducting spheres, *SIAM J. Math. Anal.*, 20 (1989), pp. 1293–1320]. Bounds for the remainders are given which hold uniformly for all ratios of the radii of the spheres. These asymptotic expansions are used to obtain an asymptotic expansion for the capacity of the spheres with respect to the infinite sphere with a uniform bound for the remainder and to find the asymptotic behavior of the capacity of two spheres. The asymptotic expansion for the capacity of two spheres with respect to the infinite sphere is also found directly from a definite-integral representation involving elliptic functions, leading to a smaller uniform bound for the remainder. Finally, the asymptotic behavior of the coefficients of potential is obtained, and the behavior of the potential difference and charge density as two spheres approach contact with given total charges is found.

Key words. coefficients of capacity, coefficients of capacity and induction, capacity, coefficients of potential, two spheres, charge density, asymptotic expansions

AMS subject classification. 31B20

1. Introduction.

The potential outside two charged conducting spheres and the charge density and total charge on each sphere were first given by Poisson [18] for both separated and tangent spheres. Further work was carried out by Plana [17] using Poisson's methods. Kirchhoff [11] corrected errors in [17] and [18] concerning the charge densities at the inner axial points as two spheres approach contact with equal radii and potentials and transformed some of Poisson's series to more rapidly convergent forms. Maxwell [14] discussed both separated and tangent spheres, obtaining new forms for the series expansions of the coefficients of capacity and induction in terms of dipolar coordinates. Barnes [2] expressed the coefficients of capacity and induction of two spheres in terms of the logarithmic derivative of his double gamma function [3]. Russell [20]–[25] carried out extensive investigations in order to facilitate the calculation of the coefficients of capacity and induction and related quantities. The latter quantities include the capacity of two spheres with respect to the infinite sphere and the capacity of two spheres. A summary of some of Russell's results is included in [12].

The capacity of two spheres has been of interest in connection with the conductivity of granular materials (Keller [10] and Batchelor and O'Brien [4]). The leading term of the asymptotic expansion of the capacity as the distance ε between the spheres tends to zero was obtained in [10], and an approximate constant term was added in [4]. Additional terms were obtained by Jeffrey in [9] by the use of the method of matched asymptotic expansions. Love [13] and Rawlins [19] have obtained the complete asymptotic expansion by different methods when the radii of the spheres are equal. However, both results are in error by the same factor due to misprints in [28].

Russell [20] obtained several terms of the asymptotic expansions of the coefficients of capacity and induction of two spheres by use of a generalization of a result due to Schlömilch [26]. However, results of the present paper show that some of his

*Received by the editors October 20, 1993; accepted for publication (in revised form) August 19, 1994.

†1000 West Nolcrest Drive, Silver Spring, MD 20903-1040 (avt@clark.net).

coefficients are in error. Buchholz [5] obtained the complete asymptotic expansions when the radii of the spheres are equal by use of the Mellin transform, but without an estimate for the remainder. The method of [5] was generalized in [8] to the case of unequal spheres, and numerical calculations were carried out in [8] to determine the optimum number of terms of the asymptotic expansion for a given separation of the spheres.

In §4 of the present paper, asymptotic expansions are obtained for the coefficients of capacity and induction starting from integral representations involving elliptic functions [27]. Bounds for the remainders are found which hold uniformly for $0 \leq r_1/r_2 \leq \infty$ in specified ranges of ε of the form $0 < \varepsilon \leq \varepsilon_0$, where r_1 and r_2 are the radii of the spheres. In the first part of §5, the results of §4 are used to obtain an asymptotic expansion of the capacity of the spheres with respect to the infinite sphere with a uniform bound for the remainder. A separate derivation of this asymptotic expansion is then carried out starting from an integral representation involving elliptic functions, and an improved bound for the remainder is obtained. The results of §5 also give the known expression for the capacity of two tangent spheres with respect to the infinite sphere and an integral representation involving hyperbolic functions.

In §6, the asymptotic behavior of the capacity of two spheres as the spheres approach tangency is found. When the radii are equal, the results of §4 are used to obtain the asymptotic expansion considered in [13] and [19] with a uniform bound for the remainder. In §7, the asymptotic behavior of the coefficients of potential as the spheres approach tangency is obtained. Also, the asymptotic behavior of the potential difference as the spheres approach contact with given total charges is found. Finally, in §8, the results of §7 are used to find the behavior of the charge density at the inner axial points as the spheres approach tangency with given total charges. A special case of interest is that in which the total charges are in the same ratio as the corresponding charges on two tangent spheres. These results are given in equivalent forms in [17] and [18], and agreement with the present results is found after correction of some misprints in each reference. Several misprints in [18] have been pointed out in [17].

2. Dipolar coordinates. As in [27], dipolar coordinates η, θ, and ϕ are defined by

$$x + i\rho = ia \cot \frac{1}{2}(\theta + i\eta), \tag{2.1}$$

$$y = \rho \cos \phi, \quad z = \rho \sin \phi, \tag{2.2}$$

with $a > 0$, $\rho > 0$. The coordinate surface η = constant is a sphere with $(a, 0, 0)$ and $(-a, 0, 0)$ as inverse points. The sphere $\eta = \eta_1$ has radius $a \operatorname{csch}|\eta_1|$ and center at $x = a \coth \eta_1$, $y = z = 0$. Hence, when $\eta_1 > 0 > \eta_2$, the sphere $\eta = \eta_1$ contains the point $(a, 0, 0)$ in its interior, and $\eta = \eta_2$ contains $(-a, 0, 0)$.

Let the spheres $\eta = \eta_1$ and $\eta = \eta_2$ have radii r_1 and r_2, respectively, and let the distance between the spheres be ε. Then as shown in [27], we have

$$a = \frac{\sqrt{\varepsilon(\varepsilon + 2r_1)(\varepsilon + 2r_2)(\varepsilon + 2r_1 + 2r_2)}}{2(\varepsilon + r_1 + r_2)} \tag{2.3}$$

and

$$\eta_1 = \sinh^{-1} \frac{a}{r_1}, \quad \eta_2 = -\sinh^{-1} \frac{a}{r_2}. \tag{2.4}$$

When $r_1 = r_2 = r$, (2.3) simplifies to

$$a = \frac{1}{2}\sqrt{\varepsilon(\varepsilon + 4r)}. \tag{2.5}$$

As $\varepsilon \to 0$, we have

$$a = \sqrt{\sigma\varepsilon}[1 + \mathrm{O}(\varepsilon)], \tag{2.6}$$

$$\eta_1 = \frac{\sqrt{\sigma\varepsilon}}{r_1}[1 + \mathrm{O}(\varepsilon)], \tag{2.7}$$

$$\eta_2 = -\frac{\sqrt{\sigma\varepsilon}}{r_2}[1 + \mathrm{O}(\varepsilon)], \tag{2.8}$$

and

$$\eta_1 - \eta_2 = 2\sqrt{\frac{\varepsilon}{\sigma}}[1 + \mathrm{O}(\varepsilon)], \tag{2.9}$$

where

$$\sigma = \frac{2r_1 r_2}{r_1 + r_2}. \tag{2.10}$$

3. Integral representations for the total charge. The total charges on the spheres in the first problem are given by

$$\begin{aligned} Q_1 &= C_{11}V_1 + C_{12}V_2, \\ Q_2 &= C_{12}V_1 + C_{22}V_2, \end{aligned} \tag{3.1}$$

where C_{11} and C_{22} are the coefficients of capacity and C_{12} is the coefficient of induction. From [15, p. 89], we obtain

$$C_{11} = a\sum_{n=0}^{\infty} \frac{e^{-N(\eta_1+\eta_2)}}{\sinh N\delta}, \tag{3.2}$$

$$C_{12} = -a\sum_{n=0}^{\infty} \frac{e^{-N\delta}}{\sinh N\delta}, \tag{3.3}$$

and

$$C_{22} = a\sum_{n=0}^{\infty} \frac{e^{N(\eta_1+\eta_2)}}{\sinh N\delta}, \tag{3.4}$$

where $\delta = \eta_1 - \eta_2$ and $N = n + 1/2$.

In [27], these series have been expresed in terms of definite integrals involving elliptic functions. Let $q = \exp(-\delta)$, and let the modulus k be defined implicitly by

$$-\log q = \eta_1 - \eta_2 = \pi K'/K, \tag{3.5}$$

where K is the complete elliptic integral of the first kind and $K' = K(k')$, $k'^2 = 1-k^2$. Also, let

$$q' = e^{-\pi K/K'} = e^{\pi^2/\log q}. \tag{3.6}$$

It follows from (2.9) and (3.5) that $q \sim \exp(-2\sqrt{\varepsilon/\sigma})$ and $q' \sim \exp(-2^{-1}\pi^2\sqrt{\sigma/\varepsilon})$ as $\varepsilon \to 0$. As in [27, Eq. (6.5)–(6.7)], we have

$$C_{11} = \frac{aKk}{\pi^2}\int_0^\pi \operatorname{Re}\,\operatorname{sn}\frac{K}{\pi}[t+i(\eta_1+\eta_2)]\csc\frac{t}{2}\,dt + \frac{iaKk}{\pi}\operatorname{sn}\frac{iK}{\pi}(\eta_1+\eta_2), \tag{3.7}$$

$$C_{12} = -\frac{aK}{\pi^2}\int_0^\pi\left(\operatorname{ns}\frac{Kt}{\pi} - \frac{\pi}{2K}\csc\frac{t}{2}\right)\csc\frac{t}{2}\,dt, \tag{3.8}$$

and

$$C_{22} = C_{11} - \frac{2iaKk}{\pi}\operatorname{sn}\frac{iK}{\pi}(\eta_1+\eta_2). \tag{3.9}$$

4. Asymptotic expansions for the coefficients of capacity and induction. Writing (3.5) in the form $\eta_2 = \eta_1 - \pi K'/K$ and substituting in (3.7), using the addition theorem for sn$u$ and Jacobi's imaginary transformation, we obtain

$$C_{11} = -\frac{aK}{\pi^2}\int_0^\pi \operatorname{Im}\,\operatorname{cs}\left[\frac{K}{\pi}(2\eta_1+it),\ k'\right]\csc\frac{t}{2}\,dt + \frac{aK}{\pi}\operatorname{cs}\left(\frac{2K\eta_1}{\pi},\ k'\right). \tag{4.1}$$

Referring to [29, p. 512], with q' defined by (3.6), we have

$$\begin{aligned} C_{11} = \frac{aK}{2\pi K'}\Bigg\{&\int_0^\pi \frac{\sinh\frac{Kt}{K'}\csc\frac{t}{2}\,dt}{\cosh\frac{Kt}{K'} - \cos\frac{2K\eta_1}{K'}} \\ &+4\sum_{n=1}^\infty \frac{q'^{2n}}{1+q'^{2n}}\cos\frac{2nK\eta_1}{K'}\int_0^\pi \sinh\frac{nKt}{K'}\csc\frac{t}{2}\,dt\Bigg\} \\ &+\frac{aK}{2K'}\left\{\cot\frac{K\eta_1}{K'} - 4\sum_{n=1}^\infty \frac{q'^{2n}}{1+q'^{2n}}\sin\frac{2nK\eta_1}{K'}\right\}, \end{aligned} \tag{4.2}$$

which converges for $\varepsilon > 0$. We find that

$$\begin{aligned} \left|\int_0^\pi \sinh\frac{nKt}{K'}\csc\frac{t}{2}\,dt\right| &< \pi\int_0^\pi \frac{\sinh\frac{nKt}{K'}\,dt}{t} \\ &< \pi\int_0^{nK\pi/K'}\cosh t\,dt = \pi\sinh\frac{nK\pi}{K'} \\ &< \frac{\pi}{2}q'^{-n}. \end{aligned} \tag{4.3}$$

Hence, denoting the first sum on the right of (4.2) by S_1, we have

$$|S_1| < \frac{\pi}{2}\sum_{n=1}^\infty q'^n = \frac{\pi}{2}\frac{q'}{1-q'}. \tag{4.4}$$

Similarly, denoting the second sum by S_2, we see that

$$|S_2| < \sum_{n=1}^{\infty} q'^{2n} = \frac{q'^2}{1-q'^2}. \tag{4.5}$$

We therefore have

$$C_{11} = \frac{aK}{2\pi K'}\left\{ \int_0^\pi \frac{\sinh\frac{Kt}{K'}\csc\frac{t}{2}\,dt}{\cosh\frac{Kt}{K'} - \cos 2\mu} + \pi\cot\mu + R^{(1)} \right\}, \tag{4.6}$$

where

$$\mu = K\eta_1/K' = \frac{\pi\eta_1}{\eta_1 - \eta_2} \tag{4.7}$$

by (3.5) and

$$|R^{(1)}| < 4|S_1| + 4\pi|S_2|. \tag{4.8}$$

It follows from (2.4) that $0 \le \mu \le \pi$, with $\mu = 0$ when $r_2/r_1 = 0$, and $\mu = \pi$ when $r_1/r_2 = 0$. We see that this range of μ is sufficient, since (4.6) is periodic in μ with period π. We see that $|R^{(1)}| = O(q')$ as $\varepsilon \to 0$.

We can verify that

$$\begin{aligned} &\int_0^\pi \frac{\sinh\frac{Kt}{K'}\csc\frac{t}{2}\,dt}{\cosh\frac{Kt}{K'} - \cos 2\mu} \\ &= 2\int_0^{\pi K/K'} \left(\frac{\sinh u}{\cosh u - \cos 2\mu} - 1\right)\left(\frac{K'u}{2K}\csc\frac{K'u}{2K} - 1\right)\frac{du}{u} \\ &+2\int_0^{\pi K/K'} \left[\frac{\sinh u}{(\cosh u - \cos 2\mu)u} - \frac{1-e^{-2u}}{u}\right] du \\ &+2\int_0^{\pi K/K'} \frac{1-e^{-2u}}{u}\,du \\ &+2\int_0^{\pi K/K'} \left(\frac{K'u}{2K}\csc\frac{K'u}{2K} - 1\right)\frac{du}{u}. \end{aligned} \tag{4.9}$$

We obtain

$$\int_0^{\pi K/K'} \left(\frac{K'u}{2K}\csc\frac{K'u}{2K} - 1\right)\frac{du}{u} = \log\frac{4}{\pi} \tag{4.10}$$

by direct integration, and we find from [1, Eq. 5.1.39] that

$$\begin{aligned} \int_0^{\pi K/K'} \frac{1-e^{-2u}}{u}\,du &= \log\frac{2\pi K}{K'} + \gamma + E_1\left(\frac{2\pi K}{K'}\right) \\ &= \log\left(-\frac{2}{\log q}\right) + 2\log\pi + \gamma + R^{(2)}/2. \end{aligned} \tag{4.11}$$

By use of the inequality

$$\int_a^\infty \frac{e^{-u}}{u}\,du < \frac{e^{-a}}{a}, \quad a > 0 \tag{4.12}$$

[1, Eq. 5.1.19], we obtain

$$\begin{aligned} |R^{(2)}| &= 2\int_{2\pi K/K'}^{\infty} \frac{e^{-u}}{u}\,du \\ &< \frac{K'}{\pi K} e^{-2\pi K/K'} = -\frac{q'^2}{\pi^2}\log q. \end{aligned} \tag{4.13}$$

We have

$$\begin{aligned} \left|\frac{\sinh u}{(\cosh u - \cos 2\mu)\,u} - \frac{1-e^{-2u}}{u}\right| &= \left(\frac{1-e^{-2u}}{u}\right)\left|\frac{2\cos 2\mu - e^{-u}}{1-2e^{-u}\cos 2\mu + e^{-2u}}\right| e^{-u} \\ &< \frac{(1+e^{-u})(2+e^{-u})}{1-e^{-u}}\cdot\frac{e^{-u}}{u} \\ &< 2.000031\frac{e^{-u}}{u} \end{aligned} \tag{4.14}$$

when $0 \le \mu \le \pi$ and $u \ge 12$. Hence,

$$\int_0^{\pi K/K'} \left[\frac{\sinh u}{(\cosh u - \cos 2\mu)\,u} - \frac{1-e^{-2u}}{u}\right] du = I + R^{(3)}, \tag{4.15}$$

where

$$I = \int_0^{\infty} \left[\frac{\sinh u}{(\cosh u - \cos 2\mu)\,u} - \frac{1-e^{-2u}}{u}\right] du \tag{4.16}$$

and by (4.12) and (4.14),

$$\begin{aligned} |R^{(3)}| &< 2.000031\int_{\pi K/K'}^{\infty} \frac{e^{-u}}{u}\,du \\ &< -4.00007q'\left(\frac{\log q}{2\pi^2}\right) \end{aligned} \tag{4.17}$$

for $0 \le \mu \le \pi$ and $-\log q \le \pi^2/12$. As in [27, Eq. (13.2)], we have

$$\begin{aligned} \frac{\sinh u}{(\cosh u - \cos 2\mu)^2} = \frac{4u}{\sin 2\mu}\Bigg\{ &\sum_{n=0}^{\infty} \frac{2\pi n + 2\mu}{[u^2 + (2\pi n + 2\mu)^2]^2} \\ &- \sum_{n=1}^{\infty} \frac{2\pi n - 2\mu}{[u^2 + (2\pi n - 2\mu)^2]^2}\Bigg\}. \end{aligned} \tag{4.18}$$

Substituting (4.18) into

$$\frac{dI}{d\mu} = -2\sin 2\mu \int_0^{\infty} \frac{\sinh u\,du}{(\cosh u - \cos 2\mu)^2\,u} \tag{4.19}$$

and interchanging summation and integration, we obtain

$$\begin{aligned} \frac{dI}{d\mu} &= -\frac{\pi}{2}\left[\sum_{n=0}^{\infty} \frac{1}{(\pi n + \mu)^2} - \sum_{n=1}^{\infty} \frac{1}{(\pi n - \mu)^2}\right] \\ &= -\frac{1}{2}\frac{d}{d\mu}\left[\psi\left(\frac{\mu}{\pi}\right) + \psi\left(1 - \frac{\mu}{\pi}\right)\right], \end{aligned} \tag{4.20}$$

where $\psi(z) = d \log \Gamma(z)/dz$. Hence,

$$I = -\psi\left(\frac{\mu}{\pi}\right) - \frac{\pi}{2}\cot\mu + c, \tag{4.21}$$

where c is a constant to be determined. The substitution $\mu = \pi/4$ yields

$$c = -\int_0^\infty \frac{\tanh u}{u} e^{-2u}\, du + \psi\left(\frac{1}{4}\right) + \frac{\pi}{2}. \tag{4.22}$$

We obtain

$$\int_0^\infty \frac{\tanh u}{u} e^{-2u}\, du = \log \frac{\pi}{2} \tag{4.23}$$

from [7, Eq. 3.551.9], and as in [7, Eq. 8.366.4], we have

$$\psi\left(\frac{1}{4}\right) = -\gamma - 3\log 2 - \frac{\pi}{2}. \tag{4.24}$$

Substitution of the preceding in (4.21) gives

$$I = -\psi\left(\frac{\mu}{\pi}\right) - \frac{\pi}{2}\cot\mu - \gamma - \log(4\pi). \tag{4.25}$$

Finally, we can obtain an asymptotic expansion for the first integral on the right-hand side of (4.9) by use of the identity

$$\begin{aligned} x\csc x = {} & 2\sum_{m=0}^{n} (-1)^m \frac{x^{2m}}{(2m)!}(1-2^{2m-1})B_{2m} \\ & + \frac{(-1)^{n+1}(2x)^{2n+2}}{(2n+1)!\sin x}\int_0^{1/2} B_{2n+1}(t)\sin 2xt\, dt \end{aligned} \tag{4.26}$$

[16, p. 32], which holds for real $x \neq \pm n\pi$, $n \geq 1$. Noting that $(-1)^{n+1}B_{2n+1}(t) > 0$ for $0 < t < 1/2$ and referring to [16, pp. 19 and 22], we find the inequalities

$$\begin{aligned} 0 < (-1)^{n+1}\int_0^{1/2} B_{2n+1}(t)\sin 2xt\, dt & < (-1)^{n+1}x\int_0^{1/2} B_{2n+1}(t)\, dt \\ & = (-1)^{n+1}\frac{x}{2n+2}B_{2n+2}\left(\frac{1}{2}\right) \\ & = (-1)^{n+1}x(2^{-2n-1}-1)\frac{B_{2n+2}}{2n+2}. \end{aligned} \tag{4.27}$$

Hence,

$$\begin{aligned} & \int_0^{\pi K/K'} \left(\frac{\sinh u}{\cosh u - \cos 2\mu} - 1\right)\left(\frac{K'u}{2K}\csc\frac{K'u}{2K} - 1\right)\frac{du}{u} \\ & = 2\sum_{m=1}^{n} \frac{(-1)^{m+1}(2^{2m-1}-1)B_{2m}}{(2m)!}\left(\frac{K'}{2K}\right)^{2m} \\ & \cdot \int_0^{\pi K/K'} \left(\frac{\cos 2\mu - e^{-u}}{\cosh u - \cos 2\mu}\right) u^{2m-1}\, du + R_n^{(4)}, \end{aligned} \tag{4.28}$$

where

$$
\begin{aligned}
|R_n^{(4)}| &< \frac{2(2^{2n+1}-1)|B_{2n+2}|}{(2n+2)!}\left(\frac{K'}{2K}\right)^{2n+3} \\
&\quad\cdot \int_0^{\pi K/K'} \left|\frac{\cos 2\mu - e^{-u}}{\cosh u - \cos 2\mu}\right| \frac{u^{2n+2}\,du}{\sin\frac{K'u}{2K}} \\
&< \frac{\pi(2^{2n+1}-1)|B_{2n+2}|}{(2n+2)!}\left(\frac{K'}{2K}\right)^{2n+2} \\
&\quad\cdot \int_0^{\pi K/K'} \left|\frac{\cos 2\mu - e^{-u}}{\cosh u - \cos 2\mu}\right| u^{2n+1}\,du.
\end{aligned}
\tag{4.29}
$$

Noting that

$$(\cos 2\mu - e^{-u})^2 \le 1 - 2e^{-u}\cos 2\mu + e^{-2u},$$

we have

$$
\begin{aligned}
\left|\frac{\cos 2\mu - e^{-u}}{\cosh u - \cos 2\mu}\right| &= \frac{2|\cos 2\mu - e^{-u}|e^{-u}}{1-2e^{-u}\cos 2\mu + e^{-2u}} \\
&\le \frac{2e^{-u}}{\sqrt{1-2e^{-u}\cos 2\mu + e^{-2u}}} \\
&< \frac{2e^{-u}}{1-e^{-u}}
\end{aligned}
\tag{4.30}
$$

for $0 \le \mu \le \pi$, $u > 0$. Hence, referring to [1, Eq. 23.2.7], we obtain

$$
\begin{aligned}
&\int_0^{\pi K/K'} \left|\frac{\cos 2\mu - e^{-u}}{\cosh u - \cos 2\mu}\right| u^{2n+1}\,du \\
&\quad < 2\int_0^{\infty} \frac{u^{2n+1}e^{-u}}{1-e^{-u}}\,du = 2\,\zeta(2n+2)\,(2n+1)!
\end{aligned}
\tag{4.31}
$$

for $0 \le \mu \le \pi$, $n \ge 1$, where $\zeta(s)$ is the Riemann zeta function. It follows from (4.29) and (4.31) that

$$|R_n^{(4)}| < 2\pi\zeta(2n+2)\frac{(2^{2n+1}-1)|B_{2n+2}|}{2n+2}\left(\frac{K'}{2K}\right)^{2n+2} \tag{4.32}$$

for $0 \le \mu \le \pi$, $n \ge 1$. Finally, by use of the inequality

$$\frac{(-1)^{n+1}(2^{2n-1}-1)B_{2n}}{(2n)!} < \pi^{-2n} \tag{4.33}$$

[1, Eq. 23.1.15], we obtain the simpler bound

$$
\begin{aligned}
|R_n^{(4)}| &< \frac{2\pi\zeta(2n+2)(-1)^n(2^{2n+1}-1)B_{2n+2}}{(2n+2)!}(2n+1)!\left(\frac{K'}{2K}\right)^{2n+2} \\
&< 2\zeta(2n+2)\pi^{-2n-1}(2n+1)!\left(\frac{K'}{2K}\right)^{2n+2} \\
&< 2.165\pi^{-2n-1}(2n+1)!\left(\frac{K'}{2K}\right)^{2n+2},
\end{aligned}
\tag{4.34}
$$

$0 \le \mu \le \pi$, $n \ge 1$, since $\zeta(2n+2)$ decreases monotonically toward 1 as $n \ge 1$ increases. In (4.34), the inequality $\zeta(4) < 1.0825$ is used. In terms of the variable $\log q = -\pi K'/K$, we have

$$|R_n^{(4)}| < 2.165\pi(2n+1)!\left(\frac{\log q}{2\pi^2}\right)^{2n+2} \tag{4.35}$$

for $0 \le \mu \le \pi$, $n \ge 1$. It follows that (4.28) is an asymptotic expansion which holds uniformly with respect to μ in the interval $0 \le \mu \le \pi$. We see that it is also a convergent expansion, since integration and summation can be interchanged in (4.28).

As in [27], we obtain a more convenient, but divergent, asymptotic expansion by writing

$$\begin{aligned}
&\sum_{m=1}^{n} \frac{(-1)^{m+1}(2^{2m-1}-1)B_{2m}}{(2m)!}\left(\frac{K'}{2K}\right)^{2m} \\
&\quad\cdot \int_0^{\pi K/K'} \left(\frac{\cos 2\mu - e^{-u}}{\cosh u - \cos 2\mu}\right) u^{2m-1}\,du \\
&= \sum_{m=1}^{n} \frac{(-1)^{m+1}(2^{2m-1}-1)B_{2m}}{(2m)!}\left(\frac{K'}{2K}\right)^{2m} \\
&\quad\cdot \int_0^{\infty} \left(\frac{\cos 2\mu - e^{-u}}{\cosh u - \cos 2\mu}\right) u^{2m-1}\,du + R_n^{(5)},
\end{aligned} \tag{4.36}$$

where

$$\begin{aligned}
R_n^{(5)} = &-\sum_{m=1}^{n} \frac{(-1)^{m+1}(2^{2m-1}-1)B_{2m}}{(2m)!}\left(\frac{K'}{2K}\right)^{2m} \\
&\cdot \int_{\pi K/K'}^{\infty} \left(\frac{\cos 2\mu - e^{-u}}{\cosh u - \cos 2\mu}\right) u^{2m-1}\,du.
\end{aligned} \tag{4.37}$$

The inequality

$$\int_x^{\infty} t^n e^{-t}\,dt < 2x^n e^{-x}, \tag{4.38}$$

$x \ge 2n$, $n \ge 0$, follows either from the asymptotic expansion with remainder of the incomplete gamma function, or by integration by parts and induction. From (4.30) and (4.38), we obtain

$$\begin{aligned}
\left|\int_{\pi K/K'}^{\infty} \left(\frac{\cos 2\mu - e^{-u}}{\cosh u - \cos 2\mu}\right) u^{2m-1}\,du\right| &< 2.00002 \int_{\pi K/K'}^{\infty} u^{2m-1} e^{-u} du \\
&< 4.00004\left(\frac{\pi K}{K'}\right)^{2m-1} e^{-\pi K/K'}
\end{aligned} \tag{4.39}$$

when $\pi K/K' \ge \max(12, 4m-2)$, $m \ge 1$. Hence, replacing $\pi K'/K$ by $-\log q$ and using (4.39), we have

$$|R_n^{(5)}| < -4.00004 \sum_{m=1}^{n} \frac{(-1)^{m+1}(2^{2m-1}-1)B_{2m}}{(2m)!}\left(\frac{\pi}{2}\right)^{2m-1} q' \frac{\log q}{\pi}$$

$$\begin{aligned} &< -4.00004 \sum_{m=1}^{n} 2^{1-2m} q' \frac{\log q}{\pi^2} \\ &< -4.00004 \left(\frac{2}{3}\right) q' \frac{\log q}{\pi^2} \\ &< -5.34q' \left(\frac{\log q}{2\pi^2}\right) \end{aligned} \tag{4.40}$$

when $\pi K/K' \geq \max(12, 4n-2)$, $n \geq 1$. From [6, p. 38], we have

$$\int_0^{\infty} \frac{\cos 2\mu - e^{-u}}{\cosh u - \cos 2\mu} u^{2m-1}\, du = \frac{(-1)^{m+1}(2\pi)^{2m}}{2m} B_{2m}\left(\frac{\mu}{\pi}\right). \tag{4.41}$$

Finally, from (4.6) through (4.41), we obtain

$$\begin{aligned} C_{11} = -\frac{a}{\log q} \Bigg\{ &\log\left(\frac{-2}{\log q}\right) - \psi\left(\frac{\mu}{\pi}\right) \\ &+ \sum_{m=1}^{n} \frac{(2^{2m-1}-1) B_{2m} B_{2m}(\frac{\mu}{\pi})}{(2m)!\, m} (\log q)^{2m} + R_n \Bigg\}, \end{aligned} \tag{4.42}$$

where

$$R_n = R^{(1)} + R^{(2)} + R^{(3)} + R_n^{(4)} + R_n^{(5)}. \tag{4.43}$$

Similarly, starting from (3.9) and using the same transformations as in (4.1), we obtain

$$\begin{aligned} C_{22} &= C_{11} - \frac{aK}{K'} \cot \mu \\ &= -\frac{a}{\log q} \Bigg\{ \log\left(\frac{-2}{\log q}\right) - \psi\left(1 - \frac{\mu}{\pi}\right) \\ &\qquad + \sum_{m=1}^{n} \frac{(2^{2m-1}-1) B_{2m} B_{2m}(\frac{\mu}{\pi})}{(2m)!\, m} (\log q)^{2m} + R_n \Bigg\} \end{aligned} \tag{4.44}$$

We see that when $\eta_2 = 0$ is substituted into (3.2), we obtain the expression for $-C_{12}$ with $q = e^{-\eta_1}$. We then have $\pi K/K' = \eta_1$, and hence $\mu = K\eta_1/K' = \pi$. It follows that we can obtain the asymptotic expansion for C_{12} from that for $-C_{11}$ by substituting $\mu = \pi$. Proceeding in this way, we find that

$$\begin{aligned} C_{12} = \frac{a}{\log q} \Bigg\{ &\log\left(\frac{-2}{\log q}\right) + \gamma \\ &+ \sum_{m=1}^{n} \frac{(2^{2m-1}-1)(B_{2m})^2}{(2m)!\, m} (\log q)^{2m} + R_n\big|_{\mu=\pi} \Bigg\}. \end{aligned} \tag{4.45}$$

When the radii of the spheres are equal, we have

$$\begin{aligned} C_{11} = C_{22} = -\frac{a}{\log q} \Bigg\{ &\log\left(\frac{-2}{\log q}\right) + \gamma + 2\log 2 \\ &- 2\sum_{m=1}^{n} \frac{(2^{2m-1}-1)^2 (B_{2m})^2}{(2m)! m} \left(\frac{\log q}{2}\right)^{2m} + R_n\big|_{\mu=\pi/2} \Bigg\}, \end{aligned} \tag{4.46}$$

since $\psi(1/2) = -\gamma - 2\log 2$ [1, Eq. 6.3.3] and $B_{2m}(1/2) = -(1-2^{1-2m})B_{2m}$ [1, Eq. 23.1.21].

When ε is sufficiently small, we can obtain more convenient bounds for the exponentially small remainder terms $R^{(1)}$, $R^{(2)}$, $R^{(3)}$, and $R_n^{(5)}$ as in [27, §9]. From the inequality

$$n! > n^n e^{-n}\sqrt{2\pi n} \tag{4.47}$$

for $n \geq 1$, we obtain

$$\begin{aligned} \frac{x^{2n+1}e^{-x/2}}{(2n+1)!} &< \left(\frac{x}{2n+1}\right)^{2n+1} e^{2n+1-x/2}\frac{1}{\sqrt{2\pi(2n+1)}} \\ &< \left(\frac{8}{e^3}\right)^3 \frac{1}{\sqrt{6\pi}} < 0.0145536, \qquad x > 8(2n+1), \\ &< \left(\frac{16}{e^7}\right)^3 \frac{1}{\sqrt{6\pi}} < 7.7.2\times 10^{-7}, \qquad x > 16(2n+1), \end{aligned} \tag{4.48}$$

$n \geq 1$, since $x^{2n+1}e^{-x/2}$ is decreasing in each case. Similarly, we find that

$$\begin{aligned} \frac{x^{2n+1}e^{-x}}{(2n+1)!} &< \left(\frac{x}{2n+1}\right)^{2n+1} e^{2n+1-x}\frac{1}{\sqrt{2\pi(2n+1)}} \\ &< \left(\frac{8}{e^7}\right)^3 \frac{1}{\sqrt{6\pi}} < 9\times 10^{-8} \end{aligned} \tag{4.49}$$

when $x \geq 8(2n+1)$ and $n \geq 1$. Multiplying both sides of the first line of (4.48) by x, we obtain

$$\frac{x^{2n+2}e^{-x/2}}{(2n+1)!} < \left(\frac{x}{2n+1}\right)^{2n+2} e^{2n+1-x/2}\sqrt{\frac{2n+1}{2\pi}} \tag{4.50}$$

under the same conditions as (4.48). For $u > 7$ and $n \geq 1$, the inequality

$$u^{2n+2}e^{-(u/2-1)(2n+1)} < u^4 e^{-3(u/2-1)}\sqrt{\frac{3}{2\pi}} \tag{4.51}$$

follows from the fact that the left-hand side is then a decreasing function of n. From (4.50) and (4.51), we find that

$$\begin{aligned} \frac{x^{2n+2}e^{-x/2}}{(2n+1)!} &< 8^4 e^{-9}\sqrt{\frac{3}{2\pi}} < 0.34929, \qquad x > 8(2n+1), \\ &< 16^4 e^{-21}\sqrt{\frac{3}{2\pi}} < 3.43374\times 10^{-5}, \qquad x > 16(2n+1), \end{aligned} \tag{4.52}$$

$n \geq 1$.

With $x = -2\pi^2/\log q$, we have $q' = e^{-x/2}$. When $x \geq 8(2n+1)$ and $n \geq 1$, we see that $q' \leq e^{-12}$. From (4.4), (4.5), (4.8), and (4.52), it follows that

$$\begin{aligned} |R^{(1)}| &< \frac{2\pi q'}{1-q'}(1+2q') \\ &< 6.283302(x^{2n+2}e^{-x/2})x^{-2n-2} \\ &< 2.1947(2n+1)!x^{-2n-2} \end{aligned} \tag{4.53}$$

when $x \geq 8(2n+1)$ and $n \geq 1$. Similarly, from (4.13), (4.17), (4.40), (4.48), and (4.49), we obtain

$$|R^{(2)}| < 0.0000002(2n+1)!x^{-2n-2}, \tag{4.54}$$

$$|R^{(3)}| < 0.05822(2n+1)!x^{-2n-2}, \tag{4.55}$$

and

$$|R_n^{(5)}| < 0.0778(2n+1)!x^{-2n-2} \tag{4.56}$$

when $x \geq 8(2n+1)$ and $n \geq 1$. Finally, from (4.34), (4.43), and (4.53)–(4.56), we have

$$|R_n| < 9.14\,(2n+1)!\left(\frac{\log q}{2\pi^2}\right)^{2n+2} \tag{4.57}$$

for $0 \leq \mu \leq \pi$ when $-\log q \leq \pi^2/4(2n+1)$ and $n \geq 1$. As we further restrict the interval in which ε lies, the exponentially small remainder terms eventually become negligible with respect to $R_n^{(4)}$. In particular, we obtain

$$|R_n| < 6.80159\,(2n+1)!\left(\frac{\log q}{2\pi^2}\right)^{2n+2}. \tag{4.58}$$

when $0 \leq \mu \leq \pi$ and $-\log q \leq \pi^2/8(2n+1)$, $n \geq 1$. We see that the coefficient in this bound differs from that in the bound for $R_n^{(4)}$ by 4 in the fifth decimal place. We find that $|R_1| < 54.9(1/24)^4 = 0.00017$ when $-\log q < \pi^2/12$. When $\log q = -\pi^2/12$ and $r_1 = r_2 = 1$, we have $\varepsilon = 0.172$.

To obtain the explicit dependence of C_{11}, C_{22}, and C_{12} on ε, we find from (3.5), (4.7), (2.7), and (2.9) that

$$-\log q = 2\sqrt{\varepsilon/\sigma}[1 + O(\varepsilon)] \tag{4.59}$$

and

$$\mu = \beta[1 + O(\varepsilon)], \tag{4.60}$$

where

$$\beta = \frac{\pi r_2}{r_1 + r_2}, \tag{4.61}$$

and σ is given by (2.10). We therefore have

$$C_{11} = \frac{\sigma}{4}\left\{\log\left(\frac{\sigma}{\varepsilon}\right) - 2\psi\left(\frac{\beta}{\pi}\right) + O(\varepsilon\log\varepsilon)\right\}, \tag{4.62}$$

$$C_{22} = \frac{\sigma}{4}\left\{\log\left(\frac{\sigma}{\varepsilon}\right) - 2\psi\left(1 - \frac{\beta}{\pi}\right) + O(\varepsilon\log\varepsilon)\right\}, \tag{4.63}$$

and

$$C_{12} = -\frac{\sigma}{4}\left\{\log\left(\frac{\sigma}{\varepsilon}\right) - 2\gamma + O(\varepsilon\log\varepsilon)\right\} \tag{4.64}$$

as $\varepsilon \to 0$.

Terms of the asymptotic expansions for C_{11}, C_{12}, and C_{22} corresponding to the preceding have been obtained by Russell [20], but without an investigation of the remainders. It can be shown that the terms of the asymptotic expansions for C_{11} and C_{22} in [20] agree with (4.42) and (4.44), respectively. However, the asymptotic expansion for C_{12} in [20] is given only through the first five terms, with the coefficients given as rational fractions. Comparison with (4.45) shows that the second, third, and fourth coefficients in [20] are too small by a factor of 1/2.

5. The capacity of two spheres with respect to the infinite sphere. Let $V_1 = V_2 = V$, and let Q_1 and Q_2 be the charges on spheres 1 and 2, respectively. Then the capacity of the spheres with respect to the infinite sphere is given by

$$C = \frac{Q_1 + Q_2}{V}. \tag{5.1}$$

It follows from (3.1) that

$$C = C_{11} + 2C_{12} + C_{22}. \tag{5.2}$$

We find from (5.2) and (3.9) that

$$Q_1/V = C_{11} + C_{12} = \frac{1}{2}C + \frac{iaKk}{\pi}\mathrm{sn}\frac{i(\eta_1 + \eta_2)}{\pi}, \tag{5.3}$$

$$Q_2/V = C_{22} + C_{12} = \frac{1}{2}C - \frac{iaKk}{\pi}\mathrm{sn}\frac{i(\eta_1 + \eta_2)}{\pi}, \tag{5.4}$$

and

$$(Q_1 - Q_2)/V = \frac{2iaKk}{\pi}\mathrm{sn}\frac{iK(\eta_1 + \eta_2)}{\pi} = \frac{2aK}{\pi}\mathrm{cs}\left(\frac{2K\eta_1}{\pi}, k'\right). \tag{5.5}$$

5.1. An asymptotic expansion for *C*. It follows immediately from (4.42)–(4.45) and (4.57) that

$$\begin{aligned} C = -\frac{a}{\log q}\bigg\{ & -\psi\left(\frac{\mu}{\pi}\right) - \psi\left(1 - \frac{\mu}{\pi}\right) - 2\gamma \\ & + 2\sum_{1}^{n} \frac{(2^{2m-1} - 1)}{(2m)!m}\left[B_{2m}B_{2m}\left(\frac{\mu}{\pi}\right) - B_{2m}^2\right](\log q)^{2m} + R_n\bigg\}, \end{aligned} \tag{5.6}$$

where

$$|R_n| < 36.6(2n+1)!\left(\frac{\log q}{2\pi^2}\right)^{2n+2} \tag{5.7}$$

for $0 \leq \mu \leq \pi$ when $-\log q \leq \pi^2/4(2n+1)$ and $n \geq 1$.

5.2. A smaller bound for $|R_n|$. We can obtain a smaller bound for $|R_n|$ by starting from the integral representation

$$C = \frac{2aK}{\pi^2}\int_0^{\pi}\left[\mathrm{Re}\,\mathrm{ns}\frac{K}{\pi}(t + 2i\eta_2) - \mathrm{ns}\frac{Kt}{\pi} + \frac{\pi}{2K}\csc\frac{t}{2}\right]\csc\frac{t}{2}\,dt, \tag{5.8}$$

which follows from (3.7) through (3.9). Proceeding as in §4, we find that

$$C = \frac{2aK}{\pi^2}\int_0^\pi \left\{ -\operatorname{Im}\operatorname{cs}\left[\frac{K}{\pi}(2\eta_1 + it), k'\right] + \operatorname{Im}\operatorname{cs}\left(\frac{iKt}{\pi}, k'\right) + \frac{\pi}{2K}\csc\frac{t}{2}\right\} dt$$

$$(5.9)\quad = \frac{aK}{\pi K'}\left\{ \int_0^\pi \left(\frac{\sinh\frac{Kt}{K'}}{\cosh\frac{Kt}{K'} - \cos 2\mu} - \coth\frac{Kt}{K'} + \frac{K'}{K}\csc\frac{t}{2}\right)\csc\frac{t}{2}\,dt \right.$$
$$\left. + 4\sum_{n=1}^{\infty}\frac{q'^{2n}}{1+q'^{2n}}\left(\cos\frac{2nK\eta_1}{K'} - 1\right)\int_0^\pi \sinh\frac{nKt}{K'}\cos\frac{t}{2}\,dt\right\},$$

where the series converges for all $\varepsilon > 0$.

Denoting the sum in (5.9) by S and using (4.3), noting that the integrand is ≥ 0, we have

$$(5.10)\qquad |S| < 2\sum_{n=1}^{\infty} q'^{2n}\int_0^\pi \sinh\frac{nKt}{K'}\cos\frac{t}{2}\,dt$$
$$< \pi\sum_{n=1}^{\infty} q'^{n} = \frac{\pi q'}{1-q'}.$$

Hence,

$$(5.11)\quad C = \frac{aK}{\pi K'}\left\{\int_0^\pi \left(\frac{\sinh\frac{Kt}{K'}}{\cosh\frac{Kt}{K'} - \cos 2\mu} - \coth\frac{Kt}{2K'} + \frac{K'}{K}\csc\frac{t}{2}\right)\csc\frac{t}{2}\,dt + R^{(1)}\right\},$$

where

$$(5.12)\qquad |R^{(1)}| < \frac{4\pi q'}{1-q'}.$$

We find that

$$\int_0^\pi \left(\frac{\sinh\frac{Kt}{K'}}{\cosh\frac{Kt}{K'} - \cos 2\mu} - \coth\frac{Kt}{2K'} + \frac{K'}{K}\csc\frac{t}{2}\right)\csc\frac{t}{2}\,dt$$
$$= 2\int_0^{\pi K/K'}\left(\frac{\sinh u}{\cosh u - \cos 2\mu} - \coth\frac{u}{2}\right)\left(\frac{K'u}{2K}\csc\frac{K'u}{2K} - 1\right)\frac{du}{u}$$
$$(5.13)\qquad +2\int_0^\infty \left(\frac{\sinh u}{\cosh u - \cos 2\mu} - \coth\frac{u}{2} + \frac{2}{u}\right)\frac{du}{u}$$
$$-2\int_{\pi K/K'}^\infty \left(\frac{\sinh u}{\cosh u - \cos 2\mu} - \coth\frac{u}{2}\right)\frac{du}{u}$$
$$-4\int_{\pi K/K'}^\infty \frac{du}{u^2} + \int_0^{\pi K/K'}\left[\left(\frac{K'}{K}\right)^2\csc^2\frac{K'u}{2K} - \frac{4}{u^2}\right]du.$$

We first have

$$(5.14)\quad -4\int_{\pi K/K'}^\infty \frac{du}{u^2} + \int_0^{\pi K/K'}\left[\left(\frac{K'}{K}\right)^2\csc^2\frac{K'u}{2K} - \frac{4}{u^2}\right]du = -\frac{4K}{\pi K'} + \frac{4K}{\pi K'} = 0.$$

Next, writing

$$(5.15)\qquad R^{(2)} = 2\int_{\pi K/K'}^\infty \left(\coth\frac{u}{2} - \frac{\sinh u}{\cosh u - \cos 2\mu}\right)\frac{du}{u},$$

we find that

$$\begin{aligned}
0 \le \coth\frac{u}{2} - \frac{\sinh u}{\cosh u - \cos 2\mu} &= \frac{(1-\cos 2\mu)(1+e^{-u})e^{-u}}{(1-e^{-u})(1-2e^{-u}\cos 2\mu + e^{-2u})} \\
&\le \frac{2(1+e^{-u})e^{-u}}{(1-e^{-u})^3}, \quad u \ge 0 \\
&\le 2.0001\, e^{-u}, \qquad u \ge 12.
\end{aligned} \tag{5.16}$$

Hence, by using (4.12), we obtain

$$\begin{aligned}
|R^{(2)}| &< 4.0002 \int_{\pi K/K'}^{\infty} \frac{e^{-u}\,du}{u} \\
&< -8.0004 q'\left(\frac{\log q}{2\pi^2}\right),
\end{aligned} \tag{5.17}$$

when $\pi K/K' \ge 12$.

While the integral

$$I = \int_0^\infty \left(\coth\frac{u}{2} - \frac{\sinh u}{\cosh u - \cos 2\mu} + \frac{2}{u}\right)\frac{du}{u} \tag{5.18}$$

can be evaluated by comparison with (5.6), it can also be evaluated directly by use of (4.21). We have

$$I = -\frac{1}{2}\left[\psi\left(\frac{\mu}{\pi}\right) + \psi\left(1 - \frac{\mu}{\pi}\right)\right] + c, \tag{5.19}$$

where c is a constant. Substituting $\mu = \pi/2$, we obtain

$$c = 2\int_0^\infty \left(\frac{1}{u} - \operatorname{csch} u\right)\frac{du}{u} + \psi\left(\frac{1}{2}\right). \tag{5.20}$$

From [7, Eq. 3.529.1], we have

$$\int_0^\infty \left(\frac{1}{u} - \operatorname{csch} u\right)\frac{du}{u} = \log 2. \tag{5.21}$$

Finally, substituting $\psi(1/2) = -\gamma - 2\log 2$, we obtain

$$I = -\frac{1}{2}\left[\psi\left(\frac{\mu}{\pi}\right) + \psi\left(1 - \frac{\mu}{\pi}\right) + 2\gamma\right]. \tag{5.22}$$

Proceeding as in §4, using (4.26) and (4.27), we have

$$\begin{aligned}
&\int_0^{\pi K/K'} \left(\frac{\sinh u}{\cosh u - \cos 2\mu} - \coth\frac{u}{2}\right)\left(\frac{K'u}{2K}\csc\frac{K'u}{2K} - 1\right)\frac{du}{u} \\
&= 2\sum_{m=1}^{n} \frac{(-1)^{m+1}(2^{2m-1}-1)B_{2m}}{(2m)!}\left(\frac{K'}{2K}\right)^{2m} \\
&\cdot \int_0^{\pi K/K'} \left(\frac{\sinh u}{\cosh u - \cos 2\mu} - \coth\frac{u}{2}\right)u^{2m-1}\,du + R_n^{(3)},
\end{aligned} \tag{5.23}$$

where

$$(5.24)\qquad |R_n^{(3)}| < \frac{\pi(2^{2n+1}-1)|B_{2n+2}|}{(2n+2)!}\left(\frac{K'}{2K}\right)^{2n+2} \cdot \int_0^{\pi K/K'} \left(\coth\frac{u}{2} - \frac{\sinh u}{\cosh u - \cos 2\mu}\right) u^{2n+1}\,du.$$

We note that the integrand in the last integral is positive when $u > 0$.

Denoting the integral in (5.24) by I and using (4.41), we obtain

$$(5.25)\qquad \begin{aligned} |I| &< \int_0^{\infty} \left(\coth\frac{u}{2} - \frac{\sinh u}{\cosh u - \cos 2\mu}\right) u^{2n+1}\,du \\ &= \int_0^{\infty} \left(\frac{1-e^{-u}}{\cosh u - 1} - \frac{\cos 2\mu - e^{-u}}{\cosh u - \cos 2\mu}\right) u^{2n+1}\,du \\ &= \frac{(-1)^n(2\pi)^{2n+2}}{2n+2}\left[B_{2n+2} - B_{2n+2}\left(\frac{\mu}{\pi}\right)\right]. \end{aligned}$$

From [16, p. 22] and (4.33), we have

$$(5.26)\qquad \begin{aligned} \left|B_{2n+2} - B_{2n+2}\left(\frac{\mu}{\pi}\right)\right| &\le \left|B_{2n+2} - B_{2n+2}\left(\frac{1}{2}\right)\right| = (2-2^{-2n-1})|B_{2n+2}| \\ &< \frac{2(2\pi)^{-2n-2}(2-2^{-2n-1})(2n+2)!}{1-2^{-2n-1}}. \end{aligned}$$

Finally, from (5.24)–(5.26) and (3.5), we obtain

$$(5.27)\qquad \begin{aligned} |R_n^{(3)}| &< 4\pi\frac{1-2^{-2n-2}}{1-2^{-2n-1}}(2n+1)!\left(\frac{\log q}{2\pi^2}\right)^{2n+2} \\ &< \frac{30}{7}\pi(2n+1)!\left(\frac{\log q}{2\pi^2}\right)^{2n+2}, \end{aligned}$$

$0 \le \mu \le \pi, n \ge 1$.

As in the case of (4.28), (5.23) is both an asymptotic and a convergent expansion. Proceeding as in (4.36) and (4.37), we write

$$(5.28)\qquad \begin{aligned} &\sum_{m=1}^{n} \frac{(-1)^{m+1}(2^{2m-1}-1)B_{2m}}{(2m)!}\left(\frac{K'}{2K}\right)^{2m} \cdot \int_0^{\pi K/K'} \left(\frac{\sinh u}{\cosh u - \cos 2\mu} - \coth\frac{u}{2}\right) u^{2m-1}\,du \\ &= \sum_{m=1}^{n} \frac{(-1)^{m+1}(2^{2m-1}-1)B_{2m}}{(2m)!}\left(\frac{K'}{2K}\right)^{2m} \cdot \int_0^{\infty} \left(\frac{\sinh u}{\cosh u - \cos 2\mu} - \coth\frac{u}{2}\right) u^{2m-1}\,du + R_n^{(4)}, \end{aligned}$$

where

$$(5.29)\qquad R_n^{(4)} = -\sum_{m=1}^{n} \frac{(-1)^{m+1}(2^{2m-1}-1)B_{2m}}{(2m)!}\left(\frac{K'}{2K}\right)^{2m} \cdot \int_{\pi K/K'}^{\infty} \left(\frac{\sinh u}{\cosh u - \cos 2\mu} - \coth\frac{u}{2}\right) u^{2m-1}\,du.$$

By use of (5.16) and (4.38), we find

$$
\begin{aligned}
0 &< \int_{\pi K/K'}^{\infty} \left(\coth\frac{u}{2} - \frac{\sinh u}{\cosh u - \cos 2\mu}\right) u^{2m-1}\,du \\
&< 2.0001 \int_{\pi K/K'}^{\infty} u^{2m-1} e^{-u}\,du \\
&< 4.0002 \left(\frac{\pi K}{K'}\right)^{2m-1} e^{-\pi K/K'}
\end{aligned}
\tag{5.30}
$$

when $\pi K/K' \geq 12$. Proceeding as in (4.40), we then obtain

$$
\begin{aligned}
|R_n^{(4)}| &< \frac{-2.0001 q' \log q}{\pi} \sum_{m=1}^{n} \frac{(-1)^{m+1}(2^{2m-1}-1)B_{2m}}{(2m)!} \left(\frac{\pi}{2}\right)^{2m-1} \\
&< \frac{-2.0001}{\pi^2} q' \log q \sum_{m=1}^{\infty} 2^{1-2m} \\
&< -1.3334 q' \left(\frac{\log q}{2\pi^2}\right)
\end{aligned}
\tag{5.31}
$$

when $-\log q \leq \pi^2/4(2n+1)$, $n \geq 1$. Next, Proceeding as in (4.53)–(4.56), we find that

$$
\begin{aligned}
|R^{(1)}| &< 4.38934\,(2n+1)!\,x^{-2n-2}, \\
|R^{(2)}| &< 0.11644\,(2n+1)!\,x^{-2n-2}, \\
|R_n^{(3)}| &< 13.4640\,(2n+1)!\,x^{-2n-2}, \\
|R^{(4)}| &< 0.01941\,(2n+1)!\,x^{-2n-2}
\end{aligned}
\tag{5.32}
$$

when $x \geq 8(2n+1)$, $n \geq 1$, where $x = 2\pi K/K'$. Finally, with $R_n = R^{(1)} + R^{(2)} + R_n^{(3)} + R^{(4)}$, we have

$$
|R_n| < 18.0\,(2n+1)! \left(\frac{\log q}{2\pi^2}\right)^{2n+2}
\tag{5.33}
$$

when $-\log q \leq \pi^2/4(2n+1)$, $n \geq 1$. The exponentially small remainder terms $R^{(1)}$, $R^{(2)}$, and $R^{(4)}$ become neglibible with respect to $R_n^{(3)}$ as we further restrict the interval in which ε lies. When $0 \leq \mu \leq \pi$ and $-\log q < \pi^2/8(2n+1)$, $n \geq 1$, we obtain

$$
|R_n| < 13.4644\,(2n+1)! \left(\frac{\pi^2}{2\pi}\right)^{(2n+2)}.
\tag{5.34}
$$

We see that the coefficient in this bound differs from that in the bound for $R_n^{(3)}$ by 4 in the fourth decimal place.

5.3. The capacity of two tangent spheres with respect to the infinite sphere. From (2.6), (2.9), (2.10), and (3.5), we have

$$
\lim_{\varepsilon \to 0} \frac{aK}{\pi K'} = \frac{r_1 r_2}{r_1 + r_2} = \sigma/2,
\tag{5.35}
$$

and from (4.60) and (4.61),

$$\lim_{\varepsilon\to 0}\mu = \beta. \tag{5.36}$$

Noting that μ and $\log q$ are continuous functions of ε for $\varepsilon \geq 0$, we find from (5.6) and (5.7) that C is a continuous function of ε at $\varepsilon = 0$. Hence, the capacity of two tangent spheres with respect to the infinite sphere is given by $C^{(0)} = \lim_{\varepsilon\to 0} C$, and we find from (5.11), (5.18), (5.22), (5.35), and (5.36) that

$$C^{(0)} = \sigma\int_0^\infty \left(\frac{\sinh u}{\cosh u - \cos 2\beta} - \coth\frac{u}{2} + \frac{2}{u}\right)\frac{du}{u} \tag{5.37}$$

$$= \frac{\sigma}{2}\left[-\psi\left(\frac{\beta}{\pi}\right) - \psi\left(1-\frac{\beta}{\pi}\right) - 2\gamma\right] \tag{5.38}$$

when $0 < \beta < \pi$. When $r_1 = r_2 = r$, (5.38) simplifies to

$$C^{(0)} = 2r\log 2. \tag{5.39}$$

Finally, the well-known results

$$Q_1^{(0)}/V = \lim_{\varepsilon\to 0} Q_1/V = \frac{\sigma}{2}\left[-\gamma - \psi\left(\frac{\beta}{\pi}\right)\right], \tag{5.40}$$

$$Q_2^{(0)}/V = \lim_{\varepsilon\to 0} Q_2/V = \frac{\sigma}{2}\left[-\gamma - \psi\left(1-\frac{\beta}{\pi}\right)\right], \tag{5.41}$$

and

$$(Q_1^{(0)} - Q_2^{(0)})/V = \lim_{\varepsilon\to 0}(Q_1 - Q_2)/V = \frac{\pi\sigma}{2}\cot\beta \tag{5.42}$$

follow from (5.3)–(5.5).

Poisson [18, pp. 56–59] has given expressions for $Q_1^{(0)}/V$, $Q_2^{(0)}/V$, and $C^{(0)}$ involving integrals between 0 and 1. When these integrals are evaluated, his expressions reduce to (5.40), (5.41), and (5.38), respectively. Also, the result (5.42) is given on page 59 of [18]. Expressions for $Q_1^{(0)}$ and $Q_2^{(0)}$ in terms of $\psi(z)$ are given in the second and third editions of Maxwell's treatise [14] but not in the first edition (1873).

6. The capacity of two spheres. With $Q_1 = -Q_2 = Q$, the capacity of two spheres is defined by

$$C_0 = \frac{Q}{V_1 - V_2}. \tag{6.1}$$

From (3.1) and (6.1), we have

$$C_0 = D/C, \tag{6.2}$$

where

$$D = C_{11}C_{22} - C_{12}^2 \tag{6.3}$$

and C is the capacity of the spheres with respect to the infinite sphere. When the radii of the spheres are equal, and hence $C_{11} = C_{22}$, we have

$$C_0 = \frac{1}{2}(C_{11} - C_{12}). \tag{6.4}$$

Let

$$C_{11} = -\frac{a}{\log q}\left[\log\left(\frac{-2}{\log q}\right) + \bar{C}_{11}\right], \tag{6.5}$$

$$C_{22} = -\frac{a}{\log q}\left[\log\left(\frac{-2}{\log q}\right) + \bar{C}_{22}\right], \tag{6.6}$$

and

$$C_{12} = \frac{a}{\log q}\left[\log\left(\frac{-2}{\log q}\right) + \bar{C}_{12}\right], \tag{6.7}$$

where $\bar{C}_{11}, \bar{C}_{22}$, and $\bar{C}_{12}$ are defined by (4.42), (4.44), and (4.45), respectively. It follows from (5.2) and the preceding that

$$C = -\frac{a}{\log q}(\bar{C}_{11} - 2\bar{C}_{12} + \bar{C}_{22}). \tag{6.8}$$

From (6.3) and (6.5)–(6.7), we obtain

$$D = \left(\frac{a}{\log q}\right)^2\left[(\bar{C}_{11} - 2\bar{C}_{12} + \bar{C}_{22})\log\left(\frac{-2}{\log q}\right) + \bar{C}_{11}\bar{C}_{22} - \bar{C}_{12}^2\right]. \tag{6.9}$$

Finally, from (6.2), (6.8), and (6.9), we have

$$C_0 = -\frac{a}{\log q}\left[\log\left(\frac{-2}{\log q}\right) + \frac{\bar{C}_{11}\bar{C}_{22} - \bar{C}_{12}^2}{\bar{C}_{11} - 2\bar{C}_{12} + \bar{C}_{22}}\right]. \tag{6.10}$$

For sufficiently small ε, we can obtain close upper and lower bounds for $\bar{C}_{11}, \bar{C}_{22}$, and $\bar{C}_{12}$ from (4.42) through (4.45) and (4.57). Similarly, bounds for C can be obtained from (5.6) and (5.33).

When the radii of the spheres are not equal, it follows from (4.42)–(4.45) and (6.10) that

$$C_0 = -\frac{a}{\log q}\left\{\log\left(\frac{-2}{\log q}\right) - \frac{\psi(\mu/\pi)\psi(1-\mu/\pi) - \gamma^2}{\psi(\mu/\pi) + \psi(1-\mu/\pi) + 2\gamma} + \mathrm{O}[(\log q)^2]\right\} \tag{6.11}$$

as $\varepsilon \to 0$. When the radii of the spheres are equal, we find from (4.42), (4.45), and (6.4) that

$$\begin{aligned} C_0 = \frac{-a}{\log q}\Bigg\{ &\log\left(\frac{-2}{\log q}\right) + \gamma \\ &+ \sum_{m=1}^{n} \frac{(2^{2m-1}-1)(B_{2m})^2}{(2m)!m}\left(\frac{\log q}{2}\right)^{2m} + R_n\Bigg\}, \end{aligned} \tag{6.12}$$

where, by (4.57),

$$|R_n| < 18.3(2n+1)!\left(\frac{\log q}{2\pi^2}\right)^{2n+2} \tag{6.13}$$

when $-\log q \le \pi^2/4(2n+1)$ and $n \ge 1$.

The asymptotic expansion (6.12) has been derived by different methods in [13] and [19], but the result obtained in each case differs from (6.12) by the factor 4. Agreement is obtained after correcting some misprints in [28, p. 232][1] and expressing the capacity in the same units as in (6.12).

[1] The factor 2 in (31) should be replaced by 1/2, and both right-hand sides of (32) should be multiplied by 4π.

7. Asymptotic behavior of the coefficients of potential as $\varepsilon \to 0$. The potentials V_1 and V_2 are given in terms of the total charges on the spheres by

$$\begin{aligned} V_1 &= p_{11}Q_1 + p_{12}Q_2, \\ V_2 &= p_{12}Q_1 + p_{22}Q_2, \end{aligned} \tag{7.1}$$

where p_{11}, p_{12}, and p_{22} are the coefficients of potential of the spheres. It follows from (3.1) that

$$p_{11} = \frac{C_{22}}{D}, \qquad p_{12} = -\frac{C_{12}}{D}, \qquad p_{22} = \frac{C_{11}}{D}, \tag{7.2}$$

where D is given by (6.3). As in the case of C_0, we can obtain upper and lower bounds for p_{11}, p_{12}, and p_{22} for sufficiently small ε by use of the results in §4.

From (4.62)–(4.64) and (5.38), we find that

$$p_{11} = \frac{1}{C^{(0)}}\left\{1 + \frac{\sigma}{C^{(0)}}\left[\psi\left(1 - \frac{\beta}{\pi}\right) + \gamma\right]^2 \left[\log\left(\frac{\sigma}{\varepsilon}\right)\right]^{-1} + O[(\log\varepsilon)^{-2}]\right\}, \tag{7.3}$$

$$p_{12} = \frac{1}{C^{(0)}}\left\{1 + \frac{\sigma}{C^{(0)}}\left[\psi\left(\frac{\beta}{\pi}\right) + \gamma\right]\left[\psi\left(1 - \frac{\beta}{\pi}\right) + \gamma\right]\left[\log\left(\frac{\sigma}{\varepsilon}\right)\right]^{-1} + O[(\log\varepsilon)^{-2}]\right\}, \tag{7.4}$$

and

$$p_{22} = \frac{1}{C^{(0)}}\left\{1 + \frac{\sigma}{C^{(0)}}\left[\psi\left(\frac{\beta}{\pi}\right) + \gamma\right]^2 \left[\log\left(\frac{\sigma}{\varepsilon}\right)\right]^{-1} + O[(\log\varepsilon)^{-2}]\right\}. \tag{7.5}$$

It follows from (7.1) and the preceding that both V_1 and V_2 are asymptotically equal to $(Q_1 + Q_2)/C^{(0)}$ as $\varepsilon \to 0$ and, hence, that $V_1 - V_2 \to 0$ as $\varepsilon \to 0$. When $r_1 = r_2$ and $Q_1 = Q_2$, we see that $V_1 = V_2$ for all ε.

From (7.2) and (7.1) together with (4.42), (4.44), and (4.45), we obtain

$$\begin{aligned} (V_1 - V_2)D = -\frac{a}{\log q}\Big\{ &-\left[\psi\left(1 - \frac{\mu}{\pi}\right) + \gamma\right] Q_1 + \left[\psi\left(\frac{\mu}{\pi}\right) + \gamma\right] Q_2 \\ &+ \frac{B_2}{2}\left[B_2\left(\frac{\mu}{\pi}\right) - B_2\right](Q_1 - Q_2)(\log q)^2 + \mathrm{O}[(\log q)^4]\Big\}. \end{aligned} \tag{7.6}$$

Using (4.59)–(4.61), we find that

$$V_1 - V_2 = \left\{-\left[\psi\left(1 - \frac{\beta}{\pi}\right) + \gamma\right] Q_1 + \left[\psi\left(\frac{\beta}{\pi}\right) + \gamma\right] Q_2 + \mathrm{O}(\varepsilon)\right\} \Big/ [D^{(0)} + \mathrm{O}(\varepsilon)], \tag{7.7}$$

where

$$D^{(0)} = C^{(0)}\left\{\frac{1}{2}\log(\sigma/\varepsilon) - \frac{\psi(\frac{\beta}{\pi})\psi(1 - \frac{\beta}{\pi}) - \gamma^2}{\psi(\frac{\beta}{\pi}) + \psi(1 - \frac{\beta}{\pi}) + 2\gamma}\right\}. \tag{7.8}$$

When $r_1 = r_2 = r$ while $Q_1 \neq Q_2$, noting that all higher-order terms in (7.6) are multiplied by $Q_1 - Q_2$, we find that (7.6) simplifies to

$$V_1 - V_2 = \frac{(Q_1 - Q_2)[1 + \mathrm{O}(\varepsilon)]}{r[\frac{1}{2}\log(r/\varepsilon) - \gamma - \log 2]}. \tag{7.9}$$

We see that the sum of the first two terms of the numerator of (7.7) vanishes when

$$Q_2/Q_1 = Q_2{}^{(0)}/Q_1{}^{(0)} = \frac{\psi(1-\beta/\pi)+\gamma}{\psi(\beta/\pi)+\gamma}, \tag{7.10}$$

where $Q_1{}^{(0)}$ and $Q_2{}^{(0)}$ are defined in (5.40) and (5.41), respectively. Noting that

$$\frac{\mu-\beta}{\pi} = \frac{r_1-r_2}{3(r_1+r_2)^2}\varepsilon + \mathrm{O}(\varepsilon^2), \tag{7.11}$$

$$(\log q)^2 = \frac{2(r_1+r_2)}{r_1r_2}\varepsilon + \mathrm{O}(\varepsilon^2), \tag{7.12}$$

and

$$\begin{aligned} \frac{1}{2}B_2[B_2(\mu/\pi)-B_2] &= \frac{1}{12}\frac{\beta}{\pi}\left(\frac{\beta}{\pi}-1\right)+\mathrm{O}(\varepsilon) \\ &= -\frac{1}{12}\frac{r_1r_2}{(r_1+r_2)^2}+\mathrm{O}(\varepsilon), \end{aligned} \tag{7.13}$$

we then find that

$$\begin{aligned} V_1 - V_2 = \frac{Q_1}{3(r_1+r_2)}\left\{\left[\left(\frac{r_1-r_2}{r_1+r_2}\right)\left\{\psi'\left(1-\frac{\beta}{\pi}\right)+\frac{Q_2{}^{(0)}}{Q_1{}^{(0)}}\psi'\left(\frac{\beta}{\pi}\right)\right\}\right.\right. \\ \left.\left. -\frac{1}{2}\left\{1-\frac{Q_2{}^{(0)}}{Q_1{}^{(0)}}\right\}\right]\varepsilon+\mathrm{O}(\varepsilon^2)\right\}\Big/[D^{(0)}+\mathrm{O}(\varepsilon)], \end{aligned} \tag{7.14}$$

where $D^{(0)}$ is given by (7.8). We see that this is the potential difference which results when two charged tangent spheres are separated slightly. When $r_1 = r_2$, the potentials remain equal for all values of ε.

8. Asymptotic behavior of the charge density at the inner axial points with Q_1 and Q_2 held constant. As in [27], the charge density at the inner axial point of sphere 1 is given by

$$D_1|_{\theta_1=\pi} = \left(\frac{V_1+V_2}{2}\right)D_{11}|_{\theta_1=\pi} + \left(\frac{V_1-V_2}{2}\right)D_{12}|_{\theta=\pi}, \tag{8.1}$$

where

$$D_{11}|_{\theta_1=\pi} = (\pi/\sigma)\sin\beta\,(\sigma/\varepsilon)^{3/2}\exp\left[-2^{-1}\pi^2(\sigma/\varepsilon)^{1/2}\right][1+\mathrm{O}(\varepsilon)] \tag{8.2}$$

and

$$D_{12}|_{\theta_1=\pi} = (2\pi\varepsilon)^{-1}[1+\mathrm{O}(\varepsilon)] \tag{8.3}$$

as $\varepsilon \to 0$. When (7.10) does not hold, it follows from (7.7)–(7.9) and (8.1)–(8.3) that

$$D_1|_{\theta=\pi} = \frac{1}{2\pi\varepsilon}\left\{-\left[\psi\left(1-\frac{\beta}{\pi}\right)+\gamma\right]Q_1+\left[\psi\left(\frac{\beta}{\pi}\right)+\gamma\right]Q_2+\mathrm{O}(\varepsilon)\right\}\Big/[D^{(0)}+\mathrm{O}(\varepsilon)] \tag{8.4}$$

when $r_1 \neq r_2$ and

$$D_1|_{\theta=\pi} = \frac{(Q_1 - Q_2)[1 + \mathrm{O}(\varepsilon)]}{2\pi\varepsilon r[\frac{1}{2}\log(r/\varepsilon) - \gamma - \log 2]} \tag{8.5}$$

when $r_1 = r_2 = r$ and $Q_1 \neq Q_2$.

When (7.10) is satisfied and $r_1 \neq r_2$, we find from (7.14) and (8.1)–(8.3) that

$$\begin{aligned} D_1|_{\theta=\pi} = \frac{Q_1}{6\pi(r_1+r_2)}\Bigg\{ \left(\frac{r_1 - r_2}{r_1 + r_2}\right) \left[\psi'\left(1 - \frac{\beta}{\pi}\right) + \frac{Q_2{}^{(0)}}{Q_1{}^{(0)}}\psi'\left(\frac{\beta}{\pi}\right)\right] \\ - \frac{1}{2}\left[1 - \frac{Q_2{}^{(0)}}{Q_1{}^{(0)}}\right] + \mathrm{O}(\varepsilon)\Bigg\} \Big/ [D^{(0)} + \mathrm{O}(\varepsilon)]. \end{aligned} \tag{8.6}$$

We see that

$$D_1|_{\theta=\pi} = \mathrm{O}[(\varepsilon\log\varepsilon)^{-1}] \tag{8.7}$$

when (7.10) does not hold and that

$$D_1|_{\theta=\pi} = \mathrm{O}[(\log\varepsilon)^{-1}] \tag{8.8}$$

when (7.10) holds and $r_1 \neq r_2$. When (7.10) holds and $r_1 = r_2 = r$, we have

$$D_1|_{\theta=\pi} = \mathrm{O}\left\{(r/\varepsilon)^{3/2}\exp\left[-2^{-1}\pi^2(r/\varepsilon)^{1/2}\right]\right\} \tag{8.9}$$

To the order of approximation shown in (8.4)–(8.6), we have

$$D_1|_{\theta=\pi} = -D_2|_{\theta=\pi} \tag{8.10}$$

However, when (7.10) is satisfied and $r_1 = r_2$, we see that

$$D_1|_{\theta=\pi} = D_2|_{\theta=\pi} \tag{8.11}$$

for all values of ε.

Results corresponding to (8.4) and (8.6) are given in [17] and [18]. However, both [17] and [18] contain misprints. After correction of these misprints and evaluation of some definite integrals, it can be shown that the results in [17] and [18] are identical to (8.4) and (8.6). While the first few terms of the asymptotic expansions of the coefficients of capacity and induction enter into these results, it was not known in [17] and [18] that these expansions are divergent. Also, the existence of the exponentially small contribution to $D_1|_{\theta=\pi}$ was not known. As noted in the introduction, the asymptotic behavior of $D_{11}|_{\theta=\pi}$ was first found by Kirchhoff in [11].

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, NBS Applied Mathematics Series 55, National Bureau of Standards, Washington, DC, 1964.

[2] E. W. Barnes, *On the coefficients of capacity of two spheres*, Quart. J. Pure Appl. Math., 35 (1904), pp. 155–175.

[3] ———, *The theory of the double gamma-functon*, Phil. Trans. Roy. Soc. London Ser. A, 196 (1901), pp. 265–397.

[4] G. K. Batchelor and R. W. O'Brien, *Thermal or electrical conduction through a granular material*, Proc. Roy. Soc. London Ser. A, 355 (1977), pp. 312–333.

[5] H. BUCHHOLZ, *Electrische und Magnetische Potentialfelder*, Springer-Verlag, Berlin, 1957.

[6] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F. G. TRICOMI, *Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[7] I. S. GRADSHTEYN AND I. M. RYSHIK, *Table of Integrals, Series, and Products*, Academic Press, New York, London, 1980.

[8] I. GUIASU AND H. RASZILLIER, *Optimal approximations of the electrostatic capacity matrix of two conducting spheres by a short-distance asymptotic expression*, IMA J. Appl. Math., 43 (1989), pp. 185–193.

[9] D. J. JEFFREY, *The temperature field or electric potential around two almost touching spheres*, J. Inst. Math. Appl., 22 (1978), pp. 337–351.

[10] J. B. KELLER, *Conductivity of a medium containing a dense array of perfectly conducting spheres or cylinders or nonconducting cylinders*, J. Appl. Phys., 34 (1963), pp. 991–993.

[11] G. KIRCHHOFF, *Die Vertheilung der Elektricität auf zwei leitenden Kugeln*, J. Reine Angew. Math., 59 (1861), pp. 89–110.

[12] F. KOTTLER, *Elektrostatik der Leiter*, in Handbuch der Physik, Vol. 12, Springer-Verlag, Berlin, 1927, Chapter 4.

[13] J. D. LOVE, *A note on the capacitance of two closely separated spheres*, J. Inst. Math. App., 24 (1979), pp. 255–257.

[14] J. C. MAXWELL, *A Treatise on Electricity and Magnetism*, Vol. 1, 3rd edition, Oxford University Press, London, 1892.

[15] E. NEUMANN, *Zur Poissonsche Theorie der Electrostatik, insbesondere über die elektrische Vertheilung auf einem von drei Kugelflächen begrenzten Conductor*, J. Reine Angew. Math., 120 (1899), pp. 60–98, 277–304.

[16] N. E. NÖRLUND, *Vorlesungen über Differenzrechnung*, Chelsea, New York, 1954.

[17] G. A. A. PLANA, *Mémoire sur la distribution de l'électricité à la surface de deux sphéres conductrices complèments isolées*, Mem. R. Accad. Sci. Torino Ser. 2, 7 (1845), pp. 71–401.

[18] S. D. POISSON, *Mémoire sur la distribution de l'électricité à la surface des corps conducteurs*, Mém. Classe de Sci. Math. Phys., l'Institut Impériale de France, 1811, part one, pp. 1–162, part two, pp. 16–274.

[19] A. D. RAWLINS, *Note on the capacitance of two closely separated spheres*, IMA J. Appl. Math., 34 (1985), pp. 119–120.

[20] A. RUSSELL, *The coefficients of capacity and the mutual attractions or repulsions of two electrified spherical conductors when close together*, Proc. Roy. Soc. London Ser. A, 82 (1909), pp. 524–531.

[21] ———, *The capacity coefficients of spherical electrodes*, Proc. Phys. Soc. London, 23 (1911), pp. 352–360.

[22] ———, *The mutual attractions or repulsions of two electrified spherical conductors*, J. Inst. Elec. Engrg., 48 (1911), pp. 257–268.

[23] ———, *The capacity coefficients of spherical conductors*, Proc. Roy. Soc. London Ser. A 97 (1920), pp. 160–172.

[24] ———, *The problem of two electrified spheres*, Proc. Phys. Soc. London, 35 (1922), pp. 10–29.

[25] ———, *The electrostatic capacity of two spheres when touching one another*, Proc. Phys. Soc. London, 37 (1925), pp. 282–286.

[26] O. SCHLÖMILCH, *Ueber die Lambertsche Reihe*, Z. Math. Phys., 6 (1860), pp. 407–415.

[27] A. H. VAN TUYL, *Electrostatic problems for two conducting spheres*, SIAM J. Math. Anal., 20 (1989), pp. 1293–1329.

[28] E. WEBER, *Electromagnetic Theory: Static Fields and Their Mapping*, Dover Publications, New York, 1965.

[29] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, 4th edition, Cambridge University Press, Cambridge, 1927, reprinted 1963.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 805–822, May 1996

011

INFINITE TOEPLITZ AND HANKEL MATRICES WITH OPERATOR-VALUED ENTRIES*

ALBRECHT BÖTTCHER† AND BERND SILBERMANN‡

Abstract. Infinite Toeplitz matrices with operator-valued entries arise, for example, when interpreting Wiener–Hopf integral operators on $L^2(0,\infty)$ as matrices acting on the direct sum of countably many copies of $L^2(0,1)$. This paper concerns the question of asymptotically inverting such infinite Toeplitz matrices by having recourse to their finite principal sections. As expected from the corresponding theories for the scalar and matrix-valued cases, this problem leads to the investigation of compactness properties of infinite Hankel matrices. By introducing the concept of Q_n-compact operators on spaces of square-summable sequences with values in a separable Hilbert space, criteria for the applicability of the finite section method to Toeplitz operators with symbols in $C + H^\infty$, in PC, or with locally sectorial symbols are established.

Key words. infinite matrices, Toeplitz operators, Hankel operators, projection methods

AMS subject classifications. 47B35, 15A06, 47A56, 65J10, 65R20

1. Introduction.

Let $\mathcal{H}$ be a separable Hilbert space and let $l^2(\mathcal{H})$ stand for the Hilbert space of all sequences $f = (f_n)_{n=0}^{\infty}$ with values $f_n \in \mathcal{H}$ for which

$$\|f\|^2 := \sum_{n=0}^{\infty} \|f_n\|_{\mathcal{H}}^2 < \infty.$$

A function a defined on the complex unit circle $\mathbf{T}$ and taking on values in $\mathcal{L}(\mathcal{H})$, the C^*-algebra of all bounded linear operators on $\mathcal{H}$, is said to belong to $L^\infty(\mathcal{L}(\mathcal{H}))$ if it is weakly measurable and

$$\|a\|_\infty := \operatorname*{ess\,sup}_{t\in\mathbf{T}} \|a(t)\|_{\mathcal{L}(\mathcal{H})} < \infty.$$

Each function $a \in L^\infty(\mathcal{L}(\mathcal{H}))$ induces both a Toeplitz operator $T(a)$ and a Hankel operator $H(a)$ on $l^2(\mathcal{H})$. If we denote the Fourier coefficients of a by $\{a_n\}_{n=-\infty}^{\infty}$,

$$a_n := \frac{1}{2\pi} \int_{-\pi}^{\pi} a(e^{i\theta}) e^{-in\theta}\, d\theta,$$

then $T(a)$ and $H(a)$ are the bounded operators on $l^2(\mathcal{H})$ given by the infinite matrices $(a_{j-k})_{j,k=0}^{\infty}$ and $(a_{j+k+1})_{j,k=0}^{\infty}$, respectively. The function a is in this context usually referred to as the *symbol* of the operators $T(a)$ and $H(a)$.

Toeplitz and Hankel operators have been studied for a long time in the scalar ($\dim \mathcal{H} = 1$) and matrix ($\dim \mathcal{H} < \infty$) cases; see [12], [5], [6] for "centennial" interim reports on the development. Much less is known in the operator case ($\dim \mathcal{H} = \infty$). The pioneering works in this direction are certainly the papers by Rabindranathan

*Received by the editors July 22, 1994; accepted for publication (in revised form) September 20, 1994.

†Fachbereich Mathematik, Technische Universität Chemnitz–Zwickau, PSF 964, D-09107 Chemnitz, Germany (aboettch@mathematik.tu-chemnitz.de). The research of this author was supported by the Alfried Krupp Förderpreis für junge Hochschullehrer of the Krupp Foundation.

‡Fachbereich Mathematik, Technische Universität Chemnitz–Zwickau, PSF 964, D-09107 Chemnitz, Germany.

[17] and Page [16]. Recent interest in this topic has come up with works by Treil [22], Gohberg and Kaashoek [13], [14] and the authors [7].

An important question in Toeplitz theory concerns the replacement of the equation $T(a)f = g$, i.e., the discrete Wiener–Hopf equation

$$\sum_{k=0}^{\infty} a_{j-k} f_k = g_j \qquad (j = 0, 1, 2, \ldots), \tag{1}$$

by its truncations (= finite sections) $T_n(a)f^{(n)} = P_n g$,

$$\sum_{k=0}^{n} a_{j-k} f_k^{(n)} = g_j \qquad (j = 0, 1, \ldots, n). \tag{2}$$

Here, by $P_n : l^2(\mathcal{H}) \to l^2(\mathcal{H})$ we denote the projections defined by

$$P_n : (g_0, g_1, g_2, \ldots) \mapsto (g_0, g_1, \ldots, g_n, 0, 0, \ldots), \tag{3}$$

and we let $T_n(a)$ stand for the compression $P_n T(a) P_n | \mathrm{Im} P_n$. If there is an $n_0 \geq 0$ such that equations (2) have a unique solution $f^{(n)} \in \mathrm{Im} P_n$ for every $g \in l^2(\mathcal{H})$ and every $n \geq n_0$ and if $f^{(n)}$ converges in $l^2(\mathcal{H})$ to a solution $f \in l^2(\mathcal{H})$ of (1), then the *finite-section method* is said to be applicable to $T(a)$. We write $T(a) \in \Pi\{P_n\}$ in this case. Some authors speak of the "projection method" instead of the "finite-section method"; we prefer the latter name, since there are many other projections methods (including various Galerkin–Petrov methods) one might apply to solve (1) approximately.

Here now is also the place to remark that Toeplitz matrices with operator-valued entries are not considered for academic purposes only. First, quarter-plane Toeplitz operators (with scalar-valued symbols) are nothing but Toeplitz matrices on $l^2(l^2)$ whose entries are themselves Toeplitz matrices. Secondly, in [13] it was pointed out that Wiener–Hopf integral operators may be "discretized" to become Toeplitz operators with operator-valued entries. This observation enabled Gohberg and Kaashoek to establish a first Szegö limit theorem for the Fredholm determinants of truncated Wiener–Hopf integral operators. Moreover, by having recourse to Toeplitz operators with operator-valued entries, the authors and Harold Widom [8] were able to prove a certain continuous analogue of the Fisher–Hartwig formula for Toeplitz determinants—up to now no other way of obtaining this analogue is known.

In the matrix case ($\dim \mathcal{H} < \infty$), there is a well-known strong interplay between Fredholm criteria and the finite-section method for Toeplitz operators on the one hand and compactness properties of Hankel operators on the other (see, e.g., [12] and [6]). Page [16] showed that studying compact Hankel operators on $l^2(\mathcal{H})$ with $\dim \mathcal{H} = \infty$ is also of interest; he proved the following operator-valued version of the celebrated Hartman theorem:

$$H(a) \in \mathcal{K}(l^2(\mathcal{H})) \qquad \Longleftrightarrow \qquad a \in C(\mathcal{K}(\mathcal{H})) + \overline{H^\infty(\mathcal{L}(\mathcal{H}))}. \tag{4}$$

Here $\mathcal{K}$ stands for the ideal of all compact operators, $C(\mathcal{K}(\mathcal{H}))$ denotes the continuous functions of $\mathbf{T}$ into $\mathcal{K}(\mathcal{H})$, and $\overline{H^\infty(\mathcal{L}(\mathcal{H}))}$ is the algebra of all $a \in L^\infty(\mathcal{L}(\mathcal{H}))$ for which $a_n = 0$ for all $n \geq 1$. In what follows, we abbreviate the set on the right of (4) to $C_\mathcal{K} + \overline{H^\infty}$. Using (4) it is easy to realize that $C_\mathcal{K} + \overline{H^\infty}$ is, in fact, a Banach subalgebra of $L^\infty(\mathcal{L}(\mathcal{H}))$ (remember R. Douglas and D. Sarason for the scalar case!), and this almost immediately yields a Fredholm theory for Toeplitz operators $T(a)$ with $a \in C_{\mathcal{K}} + \overline{H^\infty}$ on $l^2(\mathcal{H})$ (see, e.g., [5, p. 90]). The finite-section method for Toeplitz operators generated by $C_{\mathcal{K}} + \overline{H^\infty}$ functions was disposed of in [7]; note that if a is of the form identity ($\in \overline{H^\infty(\mathcal{L}(\mathcal{H}))}$) plus trace-class operator ($\in C(\mathcal{K}(\mathcal{H}))$), which is the situation we are usually confronted with when studying determinants, then $C_{\mathcal{K}} + \overline{H^\infty}$ is just the algebra that satisfies us.

Nevertheless, when dealing with operators on $l^2(\mathcal{H})$ in the $\dim \mathcal{H} = \infty$ case, the concept of compactness should be replaced by what we will call Q_n-compactness. With P_n given by (3), put $Q_n = I - P_n$, that is,

$$Q_n : (g_0, g_1, g_2, \ldots) \mapsto (0, \ldots, 0, g_{n+1}, g_{n+2}, \ldots).$$

We say that an operator $K \in \mathcal{L}(l^2(\mathcal{H}))$ is *Q_n-compact* (and write $K \in \mathcal{Q}$) if $Q_n K \rightrightarrows 0$ and $KQ_n \rightrightarrows 0$ as $n \to \infty$, where here and in what follows, $\rightrightarrows$ denotes uniform convergence (= convergence in the norm of $\mathcal{L}(l^2(\mathcal{H}))$). Since $Q_n = Q_n^*$ and $Q_n \to 0$ strongly as $n \to \infty$, every compact operator is necessarily Q_n-compact. The archetypal example of a Q_n-compact but not compact operator on $l^2(\mathcal{H})$ is the Hankel operator $H(\varphi)$ generated by the function $\varphi(e^{i\theta}) = e^{i\theta} I$. We have

$$H(\varphi) = \begin{pmatrix} I & 0 & 0 & \ldots \\ 0 & 0 & 0 & \ldots \\ 0 & 0 & 0 & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix},$$

and if $\dim \mathcal{H} = \infty$, then $H(\varphi)$ is clearly not compact, while $Q_n H(\varphi) = H(\varphi) Q_n = 0$ for all n, implying that $H(\varphi)$ is Q_n-compact.

It is easy to see that even $H(a) \in \mathcal{Q}$ for every $a \in C(\mathcal{L}(\mathcal{H}))$. Indeed, if we denote by $\sigma_n a$ the nth Fejer–Cesaro mean of a, then

$$\|(a - \sigma_n a)(e^{ix})\| \le \frac{1}{2\pi} \int_{-\pi}^{\pi} \|a(e^{i(x+\theta)}) - a(e^{ix})\| \left(\frac{\sin \frac{n+1}{2}\theta}{\sin \frac{\theta}{2}} \right)^2 \frac{d\theta}{n+1},$$

from which we infer that $\|a - \sigma_n a\|_\infty \to 0$ as $n \to \infty$, implying that

$$\|Q_n H(a)\| = \|Q_n H(a - \sigma_n a)\| \le \|a - \sigma_n a\|_\infty = o(1) \ (n \to \infty).$$

(A different argument of verifying that $H(a)$ is Q_n-compact for every $a \in C(\mathcal{L}(\mathcal{H}))$ was given in [14].) We will show that, in fact,

$$(5) \qquad H(a) \in \mathcal{Q} \quad \Longleftrightarrow \quad a \in C(\mathcal{L}(\mathcal{H})) + \overline{H^\infty(\mathcal{L}(\mathcal{H}))}.$$

A consequence of (5) is that $C + \overline{H^\infty}$ (= abbreviation for the set on the right-hand side of (5)) is a closed subalgebra of $L^\infty(\mathcal{L}(\mathcal{H}))$. We will use (5) in order to establish the following result on the applicability of the finite-section method.

If $a \in C + \overline{H^\infty}$, then $T(a) \in \Pi\{P_n\}$ if and only if both $T(a)$ and $T(\tilde{a})$ are invertible.

Here and throughout what follows, $\tilde{a}$ denotes the function obtained from a by $\tilde{a}(e^{i\theta}) := a(e^{-i\theta})$. It is well known that in the case $\dim \mathcal{H} \ge 2$, the invertibility of $T(\tilde{a})$ is in no way related to the invertibility of $T(a)$.

Functions in $C + \overline{H^\infty}$ cannot have jumps. To provide results for Toeplitz operators induced by functions with jumps and other discontinuities, we use the localization technique introduced in [19] and developed further in [5] and [6] to establish the following criterion.

If $a \in L^\infty(\mathcal{L}(\mathcal{H}))$ is locally sectorial (in a sense that will be specified below), then $T(a) \in \Pi\{P_n\}$ if and only if both $T(a)$ and $T(\tilde{a})$ are invertible.

The preceding theorem is in particular applicable to certain piecewise continuous functions, i.e., to functions $a \in L^\infty(\mathcal{L}(\mathcal{H}))$ with the property that the one-sided limits $a(\tau \pm 0) \in \mathcal{L}(\mathcal{H})$ exist for every $\tau \in \mathbf{T}$; this class of functions will henceforth be denoted by $PC(\mathcal{L}(\mathcal{H}))$. In the $\dim \mathcal{H} < \infty$ case, the local sectoriality of a function $a \in PC(\mathcal{L}(\mathcal{H}))$ can be deduced from the Fredholmness (and thus all the more from the invertibility) of $T(a)$. We have not been able to prove such a result for $a \in PC(\mathcal{L}(\mathcal{H}))$. However, using results on the finite section method for quarter-plane Toeplitz operators with piecewise continuous symbols established in [3] and [6] or employing the "numerical symbol" constructed in [20], we can prove the following theorem.

If $a \in PC(\mathbf{C} + \mathcal{K}(\mathcal{H}))$, i.e., if $a \in PC(\mathcal{L}(\mathcal{H}))$ is a function with values in

$$\mathbf{C} + \mathcal{K}(\mathcal{H}) := \{\alpha I + K : \alpha \in \mathbf{C},\, K \in \mathcal{K}(\mathcal{H})\},$$

then $T(a) \in \Pi\{P_n\}$ if and only if both $T(a)$ and $T(\tilde{a})$ are invertible.

2. Continuous symbols. The purpose of this section is to prove that if $a \in C(\mathcal{L}(\mathcal{H}))$, then $T(a) \in \Pi\{P_n\}$ if and only if $T(a)$ and $T(\tilde{a})$ are invertible. The proofs we will give are straightforward and are not based on any sort of heavy machinery. A few general remarks are nevertheless in order.

It is well known (see, e.g., [12]) that if $a \in L^\infty(\mathcal{L}(\mathcal{H}))$ and $T(a) \in \Pi\{P_n\}$, then $T(a)$ and $T(\tilde{a})$ are necessarily invertible. If, on the other hand, $T(a)$ is known to be invertible, then the applicability of the finite-section method is equivalent to the existence of an $n_0 \geq 0$ such that $T_n(a) : \mathrm{Im} P_n \to \mathrm{Im} P_n$ is invertible for all $n \geq n_0$ and such that

$$\sup_{n \geq n_0} \|T_n^{-1}(a) P_n\| < \infty.$$

We may express this also in the following form: if $T(a)$ is invertible, then $T(a) \in \Pi\{P_n\}$ if and only if there are sequences $\{R_n\}$, $\{R_n'\}$ and $\{C_n\}$, $\{C_n'\}$ of operators in $\mathcal{L}(\mathrm{Im} P_n)$ such that

$$R_n T_n(a) = P_n + C_n, \quad T_n(a) R_n' = P_n + C_n',$$

$$\sup_n \|R_n\| < \infty, \quad \sup_n \|R_n'\| < \infty, \quad \|C_n\| \to 0 \text{ and } \|C_n'\| \to 0 \text{ as } n \to \infty.$$

A key role in the investigation of the finite-section method is played by the operators W_n $(n = 0, 1, 2, \ldots)$ which are defined $l^2(\mathcal{H})$ by

$$W_n : (g_0, g_1, g_2, \ldots) \mapsto (g_n, g_{n-1}, \ldots, g_0, 0, 0, \ldots).$$

For example, with these operators at hand, one may easily understand why $T(\tilde{a})$ must be invertible if $T(a) \in \Pi\{P_n\}$. We have $W_n^2 = P_n$ and $W_n T_n(a) W_n = T_n(\tilde{a})$, whence

$$\sup_{n \geq n_0} \|T_n^{-1}(\tilde{a})\| = \sup_{n \geq n_0} \|W_n T_n^{-1}(a) W_n\| \leq \sup_{n \geq n_0} \|T_n^{-1}(a)\|.$$

We finally remark that the Hartman–Wintner theorem also holds in the operator-valued case (see, e.g., [7]): if $T(a)$ is invertible in $\mathcal{L}(l^2(\mathcal{H}))$, then a is invertible in $L^\infty(\mathcal{L}(\mathcal{H}))$.

Let us now turn to continuous symbols. Recall that the set $\mathcal{Q}$ of Q_n-*compact operators* is defined by

$$\mathcal{Q} = \{K \in \mathcal{L}(l^2(\mathcal{H})) : \quad Q_nK \rightrightarrows 0 \quad \text{and} \quad KQ_n \rightrightarrows 0 \quad \text{as} \quad n \to \infty\}.$$

As pointed out in the introduction, we have $H(a) \in \mathcal{Q}$ whenever $a \in C(\mathcal{L}(\mathcal{H}))$.

The following theorem is a special version of Corollary 8 of Devinatz and Shinbrot's paper [10]. It was independently established in [7] for symbols in $C(\mathcal{K}(\mathcal{H})) + H^\infty(\mathcal{L}(\mathcal{H}))$ and in [14] in the form presented here. The two proofs given below are proofs in the spirit of [5].

THEOREM 2.1. *Let $a \in C(\mathcal{L}(\mathcal{H}))$. Then $T(a) \in \Pi\{P_n\}$ if and only if $T(a)$ and $T(\tilde{a})$ are both invertible operators on $l^2(\mathcal{H})$.*

Proof. Suppose $T(a)$ and $T(\tilde{a})$ (and thus a and $\tilde{a}$) are invertible. One then may write down the identity $R_nT_n(a) = P_n + C_n$ with

$$\begin{aligned} R_n &= P_nT^{-1}(a)P_n + W_n(T^{-1}(\tilde{a}) - T(\tilde{a}^{-1}))W_n, \\ C_n &= -P_nT^{-1}(a)H(a)H(\tilde{a}^{-1})Q_nT(a)P_n - W_nT^{-1}(\tilde{a})H(\tilde{a})H(a^{-1})Q_nT(\tilde{a})W_n \end{aligned}$$

(see [5, p. 61]). Clearly, $\sup \|R_n\| < \infty$, and since $H(\tilde{a}^{-1})$ and $H(a^{-1})$ are in $\mathcal{Q}$, it follows that $H(\tilde{a}^{-1})Q_n \rightrightarrows 0$, $H(a^{-1})Q_n \rightrightarrows 0$, implying that $\|C_n\| \to 0$ as $n \to \infty$. In a similar way, one can show that $T_n(a)R_n = P_n + C'_n$ with $\|C'_n\| \to 0$ as $n \to \infty$. □

The preceding proof makes use of a curious identity. A perhaps more natural approach is based on the following fact, which has been known for a long time and has been employed by various authors in several contexts (see, e.g., [10], [4], [2], [21]).

LEMMA 2.2. *Let X be a linear space, P and Q be complementary projections on X (i.e., $P^2 = P$, $Q^2 = Q$, $P + Q = I$), and A be an invertible operator on X. Then the compression $PAP|\mathrm{Im}P$ is invertible on* ImP *if and only if the compression $QA^{-1}Q|\mathrm{Im}Q$ is invertible on* ImQ. *In that case,*

$$(PAP)^{-1}P = PA^{-1}P - PA^{-1}Q(QA^{-1}Q)^{-1}QA^{-1}P. \tag{6}$$

The simple proof, which merely amounts to verifying (6), is omitted (see, e.g., [5, p. 61]).

Now denote by $l^2_\#(\mathcal{H})$ the Hilbert space of all square-summable $\mathcal{H}$-valued doubly infinite sequences $(f_n)_{n=-\infty}^{\infty}$ and identify $l^2(\mathcal{H})$ as a subspace of $l^2_\#(\mathcal{H})$ in the natural way. Let P stand for the orthogonal projection of $l^2_\#(\mathcal{H})$ onto $l^2(\mathcal{H})$, put $Q = I - P$, and define the "flip operator" J on $l^2_\#(\mathcal{H})$ by $(Jf)_n = f_{-n}$ $(n = 0, \pm 1, \pm 2, \ldots)$. The Laurent operator $L(a)$ induced by a function $a \in L^\infty(\mathcal{L}(\mathcal{H}))$ is given by the matrix $(a_{j-k})_{j,k=-\infty}^{\infty}$. Traditionally, one simply writes a instead of $L(a)$. With these notations, we may write Toeplitz and Hankel operators in the following form:

$$\begin{aligned} T(a) &= PaP|\mathrm{Im}P, \quad T(\tilde{a}) = JQaQJ|\mathrm{Im}P, \\ H(a) &= PaQJ|\mathrm{Im}P, \quad H(\tilde{a}) = JQaP|\mathrm{Im}P. \end{aligned}$$

Second proof of Theorem 2.1. Suppose the operators $T(a) = PaP|\mathrm{Im}P$ and $T(\tilde{a}) = JQaQJ|\mathrm{Im}P$ are invertible. Lemma 2.2 then tells us that $T(\tilde{a}^{-1}) = JQa^{-1}QJ|\mathrm{Im}P$ and $T(a^{-1}) = Pa^{-1}P|\mathrm{Im}P$ are also invertible. Using formula (6), we get

$$\begin{aligned} Q_nT^{-1}(a)Q_n &= Q_n(PaP)^{-1}PQ_n \\ &= Q_nPa^{-1}PQ_n - Q_nPa^{-1}Q(Qa^{-1}Q)^{-1}Qa^{-1}PQ_n \\ &= Q_nPa^{-1}PQ_n - Q_nPa^{-1}QJ(JQa^{-1}QJ)^{-1}JQa^{-1}PQ_n \\ &= Q_nT(a^{-1})Q_n - Q_nH(a^{-1})T^{-1}(\tilde{a}^{-1})H(\tilde{a}^{-1})Q_n. \end{aligned} \tag{7}$$

The second term of (7) goes uniformly to zero because $H(a^{-1})$ is Q_n-compact. The first term of (7), the operator $Q_nT(a^{-1})Q_n|\mathrm{Im}Q_n$, has the same matrix as $T(a^{-1})$. It follows that the left-hand side of (7), the operator $Q_nT^{-1}(a)Q_n|\mathrm{Im}Q_n$, is invertible for all sufficiently large n and that $\|(Q_nT^{-1}(a)Q_n)^{-1}Q_n\| < 2\|T^{-1}(a^{-1})\|$ for all n large enough. For these n, we deduce from Lemma 2.2 and formula (6) that

$$(8)\quad (P_nT(a)P_n)^{-1}P_n = P_nT^{-1}(a)P_n - P_nT^{-1}(a)Q_n(Q_nT^{-1}(a)Q_n)^{-1}Q_nT^{-1}(a)P_n,$$

and since the norm of the right-hand side of (8) does not exceed

$$\|T^{-1}(a)\| + 2\|T^{-1}(a)\|\,\|T^{-1}(a^{-1})\|\,\|T^{-1}(a)\|,$$

we obtain that $T(a) \in \Pi\{P_n\}$. □

3. Q_n-compact Hankel operators and $C + \overline{H^\infty}$ symbols. It is easy to see that the collection $\mathcal{Q}$ of all Q_n-compact operators is a closed subset of $\mathcal{L}(\mathcal{H})$. If $\dim\mathcal{H} = \infty$, then $\mathcal{Q} \neq \mathcal{K}$ and hence $\mathcal{Q}$ cannot be a two-sided ideal of $\mathcal{L}(l^2(\mathcal{H}))$. However, $\mathcal{Q}$ is a two-sided ideal of certain C^*-subalgebras of $\mathcal{L}(l^2(\mathcal{H}))$. Let $\mathcal{A}$ denote the smallest closed subalgebra of $\mathcal{L}(l^2(\mathcal{H}))$ containing the set

$$\{T(a):\ a \in L^\infty(\mathcal{L}(\mathcal{H}))\} \cup \{H(a):\ a \in L^\infty(\mathcal{L}(\mathcal{H}))\} \cup \mathcal{Q}.$$

PROPOSITION 3.1. *$\mathcal{Q}$ is a closed two-sided ideal of $\mathcal{A}$.*

Proof. It suffices to show that $Q_nT(a)K \rightrightarrows 0$ and $Q_nH(a)K \rightrightarrows 0$ whenever a is in $L^\infty(\mathcal{L}(\mathcal{H}))$ and $K \in \mathcal{Q}$. Given any $\varepsilon > 0$, choose n_0 so that $\|K - P_{n_0}K\| = \|Q_{n_0}K\| < \varepsilon$. We have

$$\begin{aligned}\|Q_nT(a)K\| &\leq \|Q_nT(a)(K - P_{n_0}K)\| + \|Q_nT(a)P_{n_0}K\|\\ &\leq \varepsilon\|T(a)\| + \|Q_nT(a)P_{n_0}\|\|K\|,\end{aligned}$$

and because

$$Q_nT(a)P_{n_0} = \begin{pmatrix} a_{n+1} & \cdots & a_{n-n_0+1}\\ a_{n+2} & \cdots & a_{n-n_0+2}\\ \cdots & \cdots & \cdots \end{pmatrix},$$

it follows that

$$\|Q_nT(a)P_{n_0}\|^2 \leq (n_0+1)\sum_{k>n}\|a_k\|^2 < \varepsilon$$

if only n is large enough (note that $L^\infty(\mathcal{L}(\mathcal{H})) \subset L^2(\mathcal{L}(\mathcal{H}))$ and hence $\sum\|a_k\|^2 < \infty$). This proves that $Q_nT(a)K \rightrightarrows 0$. It can be shown similary that $Q_nH(a)K \rightrightarrows 0$. □

The classical Nehari and Hartman theorems were extended by Page [16] to the operator-valued case: if $a \in L^\infty(\mathcal{L}(\mathcal{H}))$, then

$$(9)\qquad \|H(a)\| = \mathrm{dist}\left(a,\ \overline{H^\infty(\mathcal{L}(\mathcal{H}))}\right),$$

where dist refers to the distance in $L^\infty(\mathcal{L}(\mathcal{H}))$, and we have

$$(10)\qquad H(a) \in \mathcal{K}(\mathcal{H}) \iff a \in C(\mathcal{K}(\mathcal{H})) + \overline{H^\infty(\mathcal{L}(\mathcal{H}))}.$$

PROPOSITION 3.2. *Let $a \in L^\infty(\mathcal{L}(\mathcal{H}))$. Then*

$$(11)\qquad H(a) \in \mathcal{Q} \iff a \in C(\mathcal{L}(\mathcal{H})) + \overline{H^\infty(\mathcal{L}(\mathcal{H}))}.$$

In what follows, we abbreviate the set on the right-hand side of (11) to $C+\overline{H^\infty}$.

Proof. Since $H(g)=0$ for $g\in\overline{H^\infty}$ and $H(f)\in\mathcal{Q}$ for $f\in C$, we see that $H(a)\in\mathcal{Q}$ for every $a=f+g\in C+\overline{H^\infty}$. To show the reverse implication, note first that $\|Q_nH(a)\|=\|H(\chi_{-n-1}a)\|$, where $\chi_k(t)=t^k$ $(t\in\mathbf{T})$, and then use (9) to obtain

$$\|H(\chi_{-n-1}a)\|=\operatorname{dist}(\chi_{-n-1}a,\,\overline{H^\infty})=\operatorname{dist}(a,\chi_{n+1}\overline{H^\infty})\geq\operatorname{dist}(a,\,C+\overline{H^\infty}).$$

Thus, if $\|Q_nH(a)\|\to 0$ then $\operatorname{dist}(a,\,C+\overline{H^\infty})=0$, and we are left with showing that $C+\overline{H^\infty}$ is closed.

This can be done most easily with the help of the Zalcman–Rudin lemma (see, e.g, [6, p. 75]), which says the following: if E and F are closed subspaces of a Banach space X and if there exists a sequence $\{S_n\}_{n=0}^\infty$ of operators $S_n\in\mathcal{L}(X)$ such that $\sup_n\|S_n\|<\infty$, $S_n(X)\subset E$ for all n, $S_n(F)\subset F$ for all n, and $\|S_nu-u\|\to 0$ as $n\to\infty$ for all $u\in E$, then $E+F$ is a closed subspace of X. The closedness of $C+\overline{H^\infty}$ follows from this lemma with $X=L^\infty(\mathcal{L}(\mathcal{H}))$, $E=C(\mathcal{L}(\mathcal{H}))$, $F=\overline{H^\infty(\mathcal{L}(\mathcal{H}))}$ and $S_na=\sigma_na$, where σ_na is the nth Fejer–Cesaro mean of a. □

PROPOSITION 3.3. *$C+\overline{H^\infty}$ is a closed subalgebra of $L^\infty(\mathcal{L}(\mathcal{H}))$.*

Proof (in the spirit of L. Coburn (see [18, p. 102]) and M. G. Krein [15]). We already know that $C+\overline{H^\infty}$ is closed. To prove that $C+\overline{H^\infty}$ is an algebra, take $a,b\in C+\overline{H^\infty}$ and note that

$$\begin{aligned}Q_nH(ab)&=Q_nPabQJ=Q_nPaPbQJ+Q_nPaQJJQbJ\\&=Q_nT(a)H(b)+Q_nH(a)T(\tilde{b}).\end{aligned}$$

Clearly, $Q_nH(a)T(\tilde{b})\rightrightarrows 0$, while Proposition 3.1 gives that $Q_nT(a)H(b)\rightrightarrows 0$. Consequently, $Q_nH(ab)\rightrightarrows 0$ and Proposition 3.2 finally implies that $ab\in C+\overline{H^\infty}$. □

An operator $A\in\mathcal{A}$ will be called *$\mathcal{Q}$-Fredholm* if it is invertible modulo Q_n-compact operators, i.e., if $A+\mathcal{Q}$ is invertible in $\mathcal{A}/\mathcal{Q}$. The group of invertible elements of a unital Banach algebra $\mathfrak{A}$ will henceforth be denoted by $G\mathfrak{A}$.

PROPOSITION 3.4. *Let $a\in C+\overline{H^\infty}$. Then*

$$T(a)\text{ is }\mathcal{Q}\text{-Fredholm}\iff a\in G(C+\overline{H^\infty}).$$

Proof. Let $RT(a)=I+K$ with $R\in\mathcal{L}(l^2(\mathcal{H}))$ and $K\in\mathcal{Q}$. Then

$$\|R\|\,\|T(a)f\|+\|Kf\|\geq\|f\|\qquad\forall f\in l^2(\mathcal{H}),$$

and hence

$$\|R\|\,\|PaPg\|+\|KPg\|+\|Qg\|\geq\|g\|\qquad\forall g\in l^2_\#(\mathcal{H}).$$

Denote the bilateral shift on $l^2_\#(\mathcal{H})$ by $U=L(\chi_1)$. By replacing g with U^ng in the last inequality and taking into account that U^n is an isometry, we get

$$\|R\|\,\|U^{-n}PaPU^ng\|+\|KPU^ng\|+\|U^{-n}QU^ng\|\geq\|g\|\quad\forall g\in l^2_\#(\mathcal{H}).$$

Obviously, $U^{-n}PU^n\to I$, $U^{-n}QU^n\to 0$, and $U^{-n}PaPU^n\to L(a)$ strongly as $n\to\infty$. Since

$$\begin{aligned}\|KPU^ng\|&\leq\|(K-KP_{n_0})PU^ng\|+\|KP_{n_0}PU^ng\|\\&\leq\|KQ_{n_0}\|\,\|g\|+\|K\|\,\|P_{n_0}PU^ng\|\end{aligned}$$

and $\|P_{n_0}PU^n g\|^2 \le \|g_{-n}\|^2 + \cdots + \|g_{-n+n_0}\|^2 = o(1)$, it follows that $KPU^n \to 0$ strongly. Consequently, $\|R\|\,\|L(a)g\| \ge \|g\|$ for all $g \in l^2_{\#}(\mathcal{H})$, which implies that $a \in GL^\infty(\mathcal{H})$.

To show that even $a \in G(C + \overline{H^\infty})$, consider the identity

$$(12)\quad 0 = Paa^{-1}QJ = PaQJJQa^{-1}QJ + PaPa^{-1}QJ = H(a)T(\tilde{a}^{-1}) + T(a)H(a^{-1}).$$

Multiplication of (12) by R from the left and by Q_n from the right gives

$$0 = RH(a)T(\tilde{a}^{-1})Q_n + H(a^{-1})Q_n + KH(a^{-1})Q_n.$$

From Propositions 3.1 and 3.2, we infer that $H(a)T(\tilde{a}^{-1})Q_n \rightrightarrows 0$ and $KH(a^{-1})Q_n \rightrightarrows 0$, implying that $H(a^{-1})Q_n \rightrightarrows 0$. Starting with $T(a)S = I + K$ and $0 = JQa^{-1}aP$ gives $Q_nH(a^{-1}) \rightrightarrows 0$. Thus $H(a^{-1}) \in \mathcal{Q}$ and therefore $a^{-1} \in C + \overline{H^\infty}$ by Proposition 3.3.

The implication " $\Longleftarrow$ " of the present proposition follows immediately from the identities

$$T(a)T(a^{-1}) = I - H(a)H(\tilde{a}^{-1}), \quad T(a^{-1})T(a) = I - H(a^{-1})H(\tilde{a})$$

in conjunction with Proposition 3.1 and 3.2. □

The following result was established by Devinatz and Shinbrot [10] with the help of other methods.

THEOREM 3.5. *Let $a \in C + \overline{H^\infty}$. Then*

$$T(a) \in \Pi\{P_n\} \iff T(a),\ T(\tilde{a}) \in G\mathcal{L}(l^2(\mathcal{H})).$$

Proof. Since $a^{-1} \in C + \overline{H^\infty}$ whenever $T(a)$ is invertible (Proposition 3.4), the second proof of the Theorem 2.1 can be literally used in the situation considered here. □

We remark that the first proof of Theorem 2.1 also works in the present setting: the only modification is that we now may not conclude that $H(a)H(\tilde{a}^{-1})Q_n \rightrightarrows 0$ because $H(\tilde{a}^{-1}) \in \mathcal{Q}$; we have rather to say that $H(a) \in \mathcal{Q}$ and thus $H(a)H(\tilde{a}^{-1}) \in \mathcal{Q}$ due to Proposition 3.1.

4. Localization. We now consider operators of the form $T(a) + K$, where $a \in L^\infty(\mathcal{L}(\mathcal{H}))$ is a possibly discontinuous function and K is a Q_n-compact operator. Notice that adding a perturbation K is not an academic subject. Many of the operators currently emerging are not pure Toeplitz operators but turn out to be perturbed Toeplitz operators. Our main result in this section implies the following. Suppose $a \in L^\infty(\mathcal{L}(\mathcal{H}))$ and $T(a) + K$ and $T(\tilde{a})$ are invertible. If in a neighborhood of each point $\tau \in \mathbf{T}$ the function a coincides with a function a_τ for which $T(a_\tau) \in \Pi\{P_n\}$, then $T(a) \in \Pi\{P_n\}$. The proof of this result is based on the approach developed in [19].

Denote by $\mathbf{F}$ the linear space of all sequences $\{A_n\}_{n=0}^\infty$ of operators $A_n \in \mathcal{L}(\mathrm{Im}P_n)$ such that

$$(13)\qquad \|\{A_n\}\| := \sup_{n \ge 0} \|A_n\| < \infty.$$

With the operations $\{A_n\}\{B_n\} := \{A_nB_n\}$, $\{A_n\}^* := \{A_n^*\}$ and the norm (13), the space $\mathbf{F}$ is a C^*-algebra. Let $\mathbf{J}$ stand for the collection of all sequences in $\mathbf{F}$ of the form $\{P_nKP_n + W_nLW_n + C_n\}$ with $K, L \in \mathcal{Q}$ and $\|C_n\| \to 0$ as $n \to \infty$. Finally, define $\mathbf{S}$ as the smallest closed subalgebra of $\mathbf{F}$ containing $\mathbf{J}$ and all sequences $\{T_n(a)\}$ with $a \in L^\infty(\mathcal{L}(\mathcal{H}))$. Clearly, $\mathbf{S}$ is a C^*-algebra.

PROPOSITION 4.1. *$\mathbf{J}$ is a closed two-sided ideal of $\mathbf{S}$.*

Proof. That $\mathbf{J}$ is closed can be seen as in the $\dim \mathcal{H} = 1$ case, for which we refer to [5, p. 67].

To prove that $\mathbf{J}$ is an ideal, we first show the following implications:

$$L \in \mathcal{Q} \Longrightarrow W_n L W_n \to 0 \text{ strongly,} \tag{14}$$

$$L, K \in \mathcal{Q} \Longrightarrow W_n L W_n K \rightrightarrows 0. \tag{15}$$

Given any $\varepsilon > 0$, we can find an n_0 such that

$$\|W_n L W_n - W_n P_{n_0} L W_n\| < \varepsilon/3 \quad \forall n \geq n_0.$$

For $n \geq n_0$, the matrix of $W_n P_{n_0} L W_n$ is $\left(\begin{smallmatrix} 0 & 0 \\ A_n & B_n \end{smallmatrix}\right)$ with

$$A_n = W_{n_0} P_{n_0} L Q_{n-n_0} W_{n_0}, \quad B_n = W_{n_0} P_{n_0} L P_{n-n_0} W_{n-n_0}.$$

Clearly, $\|A_n\| < \varepsilon/3$ if n is large enough. For every $f \in l^2(\mathcal{H})$, we have

$$B_n f = W_{n_0} P_{n_0} L P_{m_0} W_{n-n_0} f + W_{n_0} P_{n_0} L Q_{m_0} W_{n-n_0} f$$

and $\|L Q_{m_0}\| < \varepsilon/6$ if m_0 is sufficiently large. Since

$$P_{m_0} W_{n-n_0} f = (f_{n-n_0}, \ldots, f_{n-n_0+m_0}, 0, 0, \ldots),$$

it follows that $\|P_{m_0} W_{n-n_0} f\| < \varepsilon/6$ for all sufficiently large n. Thus, we have shown that $\|W_n L W_n f\| < \varepsilon \|f\|$ for all $f \in l^2(\mathcal{H})$ and all sufficiently large n. This proves (14). To establish (15), choose n_0 so that $\|K - P_{n_0} K\| < \varepsilon$ and $\|L - P_{n_0} L\| < \varepsilon$ and note that $\|W_n L W_n K\|$ does not exceed

$$\begin{aligned} &\|W_n L W_n (K - P_{n_0} K)\| + \|W_n (L - P_{n_0} L) W_n P_{n_0} K\| + \|W_n P_{n_0} L W_n P_{n_0} K\| \\ &\quad \leq \|L\| \varepsilon + \|K\| \varepsilon + \|W_n P_{n_0} L W_n P_{n_0} K\|. \end{aligned}$$

The assertion now follows from the observation that $W_n P_{n_0} L W_n P_{n_0}$ has the matrix $\left(\begin{smallmatrix} 0 & 0 \\ A_n & 0 \end{smallmatrix}\right)$ with $A_n = W_{n_0} P_{n_0} L Q_{n-n_0} W_{n_0}$ as above.

Now we can easily check that $\mathbf{J}$ is an ideal of $\mathbf{S}$. What we must show is the following: if $a \in L^\infty(\mathcal{L}(\mathcal{H}))$ and $K, L \in \mathcal{Q}$, then each of the six sequences

$$\begin{gathered} \{P_n T(a) P_n L P_n\}, \ \{P_n T(a) W_n L W_n\}, \ \{P_n K P_n L P_n\}, \\ \{P_n K W_n L W_n\}, \ \{W_n K W_n L P_n\}, \ \{W_n K W_n L W_n\} \end{gathered}$$

belongs to $\mathbf{J}$. For the last three sequences, this is immediate from (15). We have

$$P_n T(a) P_n L P_n = P_n T(a) P_{n_0} L P_n + P_n T(a) P_n (L - P_{n_0} L) P_n,$$

and since $T(a) P_{n_0} \in \mathcal{Q}$ by Proposition 3.1 and $\|L - P_{n_0} L\|$ is as small as desired if only n_0 is large enough, we deduce that $\{P_n T(a) P_n L P_n\} \in \mathbf{J}$ from the closedness of $\mathbf{J}$. That the second and third sequences are in $\mathbf{J}$ follows similarly from the representations

$$\begin{aligned} P_n T(a) W_n L W_n &= W_n T(\tilde{a}) P_{n_0} L W_n + W_n T(\tilde{a}) P_n (L - P_{n_0} L) W_n, \\ P_n K P_n L P_n &= P_n K P_{n_0} L P_n + P_n K P_n (L - P_{n_0} L) P_n. \qquad \square \end{aligned}$$

Now let A be any operator on $l^2(\mathcal{H})$ and suppose we have a sequence $\{A_n\} \in \mathbf{S}$ which converges strongly to A. We write $A \in \Pi\{A_n\}$ if the operators $A_n : \mathrm{Im}P_n \to \mathrm{Im}P_n$ are invertible for all sufficiently large n and $f^{(n)} = A_n^{-1}P_n g$ converges in $l^2(\mathcal{H})$ to a solution f of the equation $Af = g$ for every $g \in l^2(\mathcal{H})$. In the case where $A_n = P_nAP_n|\mathrm{Im}P_n$, we write $A \in \Pi\{P_n\}$ in place of $A \in \Pi\{A_n\}$, which is in accordance with our previous notation.

Again let $A \in \mathcal{L}(l^2(\mathcal{H}))$, $\{A_n\} \in \mathbf{S}$ and suppose $A_n \to A$ strongly. Then $\{W_nA_nW_n\}$ is also a sequence in $\mathbf{S}$. We claim that every sequence in $\mathbf{S}$ has a strong limit. To see this, it suffices to verify that $T_n(a)\,(a \in L^\infty(\mathcal{L}(\mathcal{H}))$, $P_nKP_n\,(K \in \mathcal{Q})$, and $W_nLW_n\,(L \in \mathcal{Q})$ converge strongly. But the strong convergence of $T_n(a)$ and P_nKP_n to $T(a)$ and K, respectively, is obvious, while (14) tells us that W_nLW_n converges strongly to zero. It follows that, in particular, $W_nA_nW_n$ has a strong limit; this limit will be denoted by $\tilde{A}$ (although it depends not only on A but also on the sequence $\{A_n\}$).

Finally, for $\{A_n\} \in \mathbf{S}$, we denote by $\{A_n\}^\pi$ the coset $\{A_n\} + \mathbf{J}$ in the quotient algebra $\mathbf{S}/\mathbf{J}$.

THEOREM 4.2. *Let $A \in \mathcal{L}(l^2(\mathcal{H}))$, $\{A_n\} \in \mathbf{S}$, and suppose that $A_n \to A$ strongly. Then $A \in \Pi\{A_n\}$ if and only if A and $\tilde{A}$ are invertible in $\mathcal{L}(l^2(\mathcal{H}))$ and $\{A_n\}^\pi$ is invertible in $\mathbf{S}/\mathbf{J}$.*

Proof. The "only if " part can shown by standard arguments (see [5, p. 68]). So assume $A, \tilde{A}$, and $\{A_n\}^\pi$ are invertible. From the invertibility of $\{A_n\}^\pi$, we deduce the existence of a sequence $\{R_n\} \in \mathbf{S}$ such that

$$A_nR_n = P_n + P_nKP_n + W_nLW_n + C_n$$

with $K, L \in \mathcal{Q}$ and $\|C_n\| \to 0$. Put

$$R_n' = R_n - P_nA^{-1}KP_n - W_n\tilde{A}^{-1}LW_n.$$

Since $\{P_nA^{-1}KP_n + W_n\tilde{A}^{-1}LW_n\} \in \mathbf{J}$, it follows that $\{R_n'\} \in \mathbf{S}$. Clearly,

$$(16)\quad A_nR_n' = P_n + P_n(K - A_nP_nA^{-1}K)P_n + W_n(L - W_nA_nW_n\tilde{A}^{-1}L)W_n + C_n.$$

We claim that if $\{B_n\}$ is any sequence in $\mathbf{S}$ and K any Q_n-compact operator, then $B_nK \rightrightarrows BK$ as $n \to \infty$. To show this, we may restrict ourselves to the cases where B_n is $T_n(a)$, P_nLP_n or W_nLW_n with $L \in \mathcal{Q}$. We have

$$T_n(a)P_nK - T(a)K = -Q_nT(a)K - P_nT(a)Q_nK \rightrightarrows 0$$

since $T(a)K \in \mathcal{Q}$ by Proposition 3.1 and $K \in \mathcal{Q}$ by assumption. Since $P_nL \rightrightarrows L$ and $P_nK \rightrightarrows K$, we obtain that $P_nLP_nK \rightrightarrows LK$. Finally, from (14) and (15), we infer that $W_nLW_nK \rightrightarrows 0 = 0 \cdot K$. This proves our claim.

From what has just been shown, we obtain in particular that

$$\begin{aligned} K - A_nP_nA^{-1}K &\rightrightarrows K - AA^{-1}K = 0, \\ L - W_nA_nW_n\tilde{A}^{-1}L &\rightrightarrows L - \tilde{A}\tilde{A}^{-1}L \;= 0, \end{aligned}$$

and hence (16) may be written in the form $A_nR_n' = P_n + C_n'$ with $\|C_n'\| \to 0$. In a similar fashion, one can find a sequence $\{R_n''\} \in \mathbf{S}$ such that $R_n''A_n = P_n + C_n''$ with $\|C_n''\| \to 0$. This implies that $A \in \Pi\{A_n\}$. □

The following extension of Theorem 3.5 is an immediate consequence of the preceding theorem.

THEOREM 4.3. *Let $a \in C + \overline{H^\infty}$ and $K \in \mathcal{Q}$. Then*

$$T(a) + K \in \Pi\{P_n\} \iff T(a) + K,\ T(\tilde{a}) \in G\mathcal{L}(l^2(\mathcal{H})).$$

Proof. Because $W_n(T(a)+K)W_n = T_n(\tilde{a}) + W_nKW_n$ converges strongly to $T(\tilde{a})$ (recall (14)), we get the implication "$\Longrightarrow$." Conversely, suppose $T(a)+K$ and $T(\tilde{a})$ are invertible. We must show that $\{T_n(a)+K\}^\pi = \{T_n(a)\}^\pi$ is also invertible.

Since $T(\tilde{a})$ is invertible, we have $\tilde{a}^{-1} \in C + H^\infty$ and thus $a^{-1} \in C + \overline{H^\infty}$ by virtue of Proposition 3.4. Consequently, $H(a)H(\tilde{a}^{-1})$ and $H(\tilde{a})H(a^{-1})$ are in $\mathcal{Q}$ due to Propositions 3.1 and 3.2. It remains to write down Widom's identity

$$T_n(a)T_n(a^{-1}) = P_n - P_nH(a)H(\tilde{a}^{-1})P_n - W_nH(\tilde{a})H(a^{-1})W_n$$

and to observe that $\{P_nH(a)H(\tilde{a}^{-1})P_n + W_nH(\tilde{a})H(a^{-1})W_n\} \in \mathbf{J}$. □

Let $\mathcal{C}$ be the smallest closed subalgebra of $\mathcal{L}(l^2(\mathcal{H}))$ containing all Toeplitz operators $T(\varphi)$ with $\varphi \in C(\mathcal{L}(\mathcal{H}))$. Using Proposition 3.2, one can easily show that every operator $A \in \mathcal{C}$ is of the form $A = T(\varphi) + K$ with $A \in C(\mathcal{L}(\mathcal{H}))$ and $K \in \mathcal{Q}$ (note that $\mathcal{Q}$ is in fact the quasi-commutator ideal of $\mathcal{C}$, i.e., the smallest closed two-sided ideal of $\mathcal{C}$ containing all quasi-commutators $T(\varphi\psi) - T(\varphi)T(\psi)$ $(\varphi, \psi \in C(\mathcal{L}(\mathcal{H})))$. Thus, Theorem 4.3 implies that if $A \in \mathcal{C}$, then

$$A \in \Pi\{P_n\} \iff A \text{ and } \tilde{A} := \lim_{n\to\infty} W_nAW_n \text{ are invertible.}$$

This is also the right place to give an interlude on so-called *paired operators*, which may be viewed as singular integral operators in matrix disguise and are, in the case of continuous coefficients, a nice example of Q_n-compactly perturbed Toeplitz operators. Given $a, b \in L^\infty(\mathcal{L}(\mathcal{H}))$, the paired operator induced by a and b is the operator $aP + bQ$ $(= L(a)P + L(b)Q)$ on $l^2_\#(\mathcal{H})$. Define $\mathcal{P}_n$ on $l^2_\#(\mathcal{H})$ by

$$\mathcal{P}_n\left((x_j)_{j=-\infty}^{\infty}\right) = (\ldots, 0, 0, x_{-n-1}, \ldots, x_{-1}, x_0, \ldots, x_n, 0, 0, \ldots).$$

We write $aP+bQ \in \Pi\{\mathcal{P}_n\}$ if the operators $A_n = \mathcal{P}_n(aP+bQ)\mathcal{P}_n|\mathrm{Im}\mathcal{P}_n$ are invertible for all sufficiently large n and $f^{(n)} = A_n^{-1}\mathcal{P}_ng$ converges in $l^2_\#(\mathcal{H})$ to a solution f of $(aP+bQ)f = g$ for every $g \in l^2_\#(\mathcal{H})$. The mapping

$$C\ :\ l^2_\#(\mathcal{H}) \to l^2(\mathcal{H}) \oplus l^2(\mathcal{H}),\ f \mapsto JQf \oplus Pf$$

is a Hilbert-space isomorphism, and $C(aP+bQ)C^{-1}$ and $C\mathcal{P}_nC^{-1}$ are given on $l^2(\mathcal{H}) \oplus l^2(\mathcal{H})$ by

$$\begin{pmatrix} JQbQJ & JQaP \\ PbQJ & PaP \end{pmatrix} = \begin{pmatrix} T(\tilde{b}) & H(\tilde{a}) \\ H(b) & T(a) \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} P_n & 0 \\ 0 & P_n \end{pmatrix},$$

respectively. Hence,

$$aP + bQ \in \Pi\{\mathcal{P}_n\} \iff \begin{pmatrix} T(\tilde{b}) & H(\tilde{a}) \\ H(b) & T(a) \end{pmatrix} \in \Pi\left\{\begin{pmatrix} P_n & 0 \\ 0 & P_n \end{pmatrix}\right\}.$$

Further, the mapping $D\ :\ l^2(\mathcal{H}) \oplus l^2(\mathcal{H}) \to l^2(\mathcal{H} \oplus \mathcal{H})$ defined by

$$D\ :\ (f_0, f_1, \ldots) \oplus (g_0, g_1, \ldots) \mapsto (f_0, g_0, f_1, g_1, \ldots)$$

is also a Hilbert-space isomorphism, and we have

$$D\begin{pmatrix} T(\tilde{b}) & H(\tilde{a}) \\ H(b) & T(a) \end{pmatrix} D^{-1} = T\begin{pmatrix} \tilde{b} & 0 \\ 0 & a \end{pmatrix} + H\begin{pmatrix} 0 & \tilde{a} \\ b & 0 \end{pmatrix},$$

$$D\begin{pmatrix} T_n(\tilde{b}) & P_nH(\tilde{a})P_n \\ P_nH(b)P_n & T_n(a) \end{pmatrix} D^{-1} = T_n\begin{pmatrix} \tilde{b} & 0 \\ 0 & a \end{pmatrix} + P_nH\begin{pmatrix} 0 & \tilde{a} \\ b & 0 \end{pmatrix} P_n.$$

Consequently,

$$aP + bQ \in \Pi\{\mathcal{P}_n\} \iff T(c) + K \in \Pi\{P_n\},$$

where $c = \left(\begin{smallmatrix} \tilde{b} & 0 \\ 0 & a \end{smallmatrix}\right)$ and $K = H\left(\begin{smallmatrix} 0 & \tilde{a} \\ b & 0 \end{smallmatrix}\right)$. If $a, b \in C(\mathcal{L}(\mathcal{H}))$, we are in the situation covered by Theorem 4.3 (only with $\mathcal{H} \oplus \mathcal{H}$ in place of $\mathcal{H}$). Since $T(\tilde{c}) = T\left(\begin{smallmatrix} b & 0 \\ 0 & \tilde{a} \end{smallmatrix}\right)$ is invertible if and only if $T(b)$ and $T(\tilde{a})$ are as well, we arrive at the following result, which was stated in a different form and proved by other methods (and under the a priori assumption that a and b be invertible) in [14].

THEOREM 4.4. *Let $a, b \in C(\mathcal{L}(\mathcal{H}))$. Then*

$$aP + bQ \in \Pi\{\mathcal{P}_n\} \iff aP + bQ \text{ and } PbP + QaQ \text{ are invertible.} \qquad \square$$

A moment's thought reveals that the above argument is also applicable to operators of the form $aP + bQ + L$, where $a, b \in C(\mathcal{L}(\mathcal{H}))$ and L is a Q_n-compact operator, that is, an operator for which $\mathcal{P}_nL \rightrightarrows L$ and $L\mathcal{P}_n \rightrightarrows L$. In that case, we have

$$aP + bQ + L \in \Pi\{\mathcal{P}_n\} \iff aP + bQ + L \text{ and } PbP + QaQ \text{ are invertible.}$$

We now turn to Toeplitz operators with discontinuous symbols. Let $C(\mathbf{C})$ denote the C^*-algebra of all complex-valued continuous functions on $\mathbf{T}$. If $\varphi \in C(\mathbf{C})$ and I is the identity operator on $\mathcal{H}$, we define $\varphi I \in \mathcal{L}(\mathcal{H})$ in the natural way (as $\varphi \otimes I$). In case φI is followed by another operator $a \in \mathcal{L}(\mathcal{H})$, we abbreviate $\varphi I A$ to φA.

Two functions $a, b \in L^\infty(\mathcal{L}(\mathcal{H}))$ are said to be *locally equivalent* at a point $\tau \in \mathbf{T}$ if

$$\inf\{\|\varphi(a - b)\| : \varphi \in C(\mathbf{C}),\ \varphi(\tau) = 1\} = 0.$$

THEOREM 4.5. *Let $a \in L^\infty(\mathcal{L}(\mathcal{H}))$ and $K \in \mathcal{Q}$. Suppose $T(a) + K$ and $T(\tilde{a})$ are invertible and a is at every point $\tau \in \mathbf{T}$ locally equivalent to a function $a_\tau \in L^\infty(\mathcal{L}(\mathcal{H}))$ for which $\{T_n(a_\tau)\}^\pi$ is invertible. Then $T(a) + K \in \Pi\{P_n\}$.*

Proof. By virtue of Theorem 4.3, we are left with showing that $\{T_n(a)\}^\pi$ is invertible. This will be done by making use of the local principle of Allan and Douglas (see, e.g., [6, Thm. 1.34]) in the algebra $\mathbf{S}^\pi := \mathbf{S}/\mathbf{J}$.

The mapping $\gamma : C(\mathbf{C}) \to \mathbf{S}^\pi$, $\varphi \mapsto \{T_n(\varphi I)\}^\pi$ is readily seen to be a C^*-algebra homomorphism. Moreover, γ is injective: if $\{T_n(\varphi I)\} = \{P_nKP_n + W_nLW_n + C_n\}$ is in $\mathbf{J}$, then $T(\varphi I) = K$ is Q_n-compact, whence $\|T(\varphi I)\| = \|Q_nT(\varphi I)Q_n\| \to 0$, implying that $\varphi = 0$. It follows that $\gamma(C(\mathbf{C}))$ is a commutative C^*-algebra isomorphic to $C(\mathbf{C})$ and thus that the maximal ideal space of $\gamma(C(\mathbf{C}))$ may be naturally identified with $\mathbf{T}$.

For $\varphi \in C(\mathbf{C})$ and $b \in L^\infty(\mathcal{L}(\mathcal{H}))$, we have

$$\{T_n(\varphi I)\}^\pi\{T_n(b)\}^\pi = \{T_n(\varphi b)\}^\pi = \{T_n(b\varphi I)\}^\pi = \{T_n(b)\}^\pi\{T_n(\varphi I)\}^\pi,$$

and hence $\gamma(C(\mathbf{C}))$ is contained in the center of $\mathbf{S}^\pi$. For $\tau \in \mathbf{T}$, denote by $\mathbf{J}_\tau^\pi$ the smallest closed two-sided ideal of $\mathbf{S}^\pi$ containing the set $\{\{T_n(\varphi I)\}^\pi : \varphi \in C(\mathbf{C}), \varphi(\tau) = 0\}$.

The local principle of Allan and Douglas says that $\{T_n(a)\}^\pi$ is invertible in $\mathbf{S}^\pi$ if and only if for every $\tau \in \mathbf{T}$, the coset $\{T_n(a)\}^\pi + \mathbf{J}_\tau^\pi$ is invertible in $\mathbf{S}^\pi/\mathbf{J}_\tau^\pi$. But it is easy to see that $\{T_n(a)\}^\pi + \mathbf{J}_\tau^\pi = \{T_n(a_\tau)\}^\pi + \mathbf{J}_\tau^\pi$, and because $\{T_n(a_\tau)\}^\pi$ was assumed to be invertible, so all the more is $\{T_n(a_\tau)\}^\pi + \mathbf{J}_\tau^\pi$. □

Since, by Theorem 4.3, the element $\{T_n(a_\tau)\}^\pi$ is invertible whenever $T(a_\tau) \in \Pi\{P_n\}$, the preceding theorem is also true with the phrase "for which $\{T_n(a_\tau)\}^\pi$ is invertible" replaced by "for which $T(a_\tau) \in \Pi\{P_n\}$."

We also remark that Theorem 4.5 extends to localization over the algebra $QC(\mathbf{C})$ of all quasicontinuous functions, i.e., the operator-valued version of Theorem 7.32(i) in [6] is valid.

5. Locally sectorial symbols. The *essential range* $\mathfrak{R}(a)$ of a function $a \in L^\infty(\mathcal{L}(\mathcal{H}))$ may be defined as the spectrum of the Laurent operator $L(a) \in \mathcal{L}(l^2_\#(\mathcal{H}))$. A scalar-valued function $a \in L^\infty(\mathcal{L}(\mathbf{C})) = L^\infty(\mathbf{C})$ is called globally sectorial if $\mathfrak{R}(a)$ is contained in some open half-plane whose boundary passes through the origin. There are at least two possibilities of extending the concept of sectoriality to operator-valued functions.

A function $a \in L^\infty(\mathcal{L}(\mathcal{H}))$ is said to be *globally analytically sectorial* ($a \in GAS$) if there are $b, c \in G\mathcal{L}(\mathcal{H})$ and an $\varepsilon > 0$ such that

$$\operatorname{Re}(ba(t)ch, h) \geq \varepsilon \|h\|^2$$

for almost all $t \in \mathbf{T}$ and all $h \in \mathcal{H}$. Thus, if we denote by $\mathfrak{H}(a)$ the *Hausdorff range* of the function a,

$$\mathfrak{H}(a) := \{(\alpha h, h) : \alpha \in \mathfrak{R}(a), \|h\| = 1\},$$

then $a \in GAS$ if and only if there are $b, c \in G\mathcal{L}(\mathcal{H})$ such that $\mathfrak{H}(bac)$ is contained in the right open half-plane.

We call a function $a \in L^\infty(\mathcal{L}(\mathcal{H}))$ *globally geometrically sectorial* ($a \in GGS$) if the convex hull of its essential range, conv $\mathfrak{R}(a)$, consists entirely of invertible operators: conv $\mathfrak{R}$ $(a) \subset G\mathcal{L}(\mathcal{H})$.

In the scalar case ($\dim\mathcal{H} = 1$), we have $GAS = GGS$, whereas in the $\dim \mathcal{H} > 1$ case $GAS \subset GGS$ but $GAS \neq GGS$ (see [1]).

A function $a \in L^\infty(\mathcal{L}(\mathcal{H}))$ is said to be *locally analytically* or *geometrically sectorial* ($a \in LAS$ or $a \in LGS$) if a is at every point $\tau \in \mathbf{T}$ locally equivalent to some function a_τ in GAS or LGS, respectively.

Functions in GAS are fairly well understood. One can show (see, e.g., [6, p. 104]) that the following are equivalent:

(i) $a \in GAS$;

(ii) there is a $d \in G\mathcal{L}(\mathcal{H})$ such that $\mathfrak{H}(ad)$ is contained in the right open half-plane;

(iii) there exist both an operator $e \in G\mathcal{L}(\mathcal{H})$ and a number $q \in (0,1)$ such that $\|I - a(t)e\| \leq q < 1$ for almost all $t \in \mathbf{T}$.

PROPOSITION 5.1. *If* $a \in GAS$, *then* $T(a) \in \Pi\{P_n\}$.

Proof. Let e and q be as in condition (iii) above. Then

$$\|I - T(a)T(e)\| \leq \|I - ae\| \leq q < 1,$$

which gives the invertibility of $T(a)$, and

$$\|P_n - T_n(a)T_n(e)\| \leq \|I - ae\| \leq q < 1.$$

The latter inequality implies that $T_n(a)$ is invertible for all n large enough and that $\|T_n^{-1}(a)\| \leq \|e\|/(1-q)$. □

THEOREM 5.2. *Let $a \in LAS$ and $K \in \mathcal{Q}$. Then*

$$T(a) + K \in \Pi\{P_n\} \iff T(a) + K,\ T(\tilde{a}) \in G\mathcal{L}(l^2(\mathcal{H})).$$

Proof. This is immediate from Theorem 4.5 and Proposition 5.1. □

We remark that Theorem 5.2 also holds for functions a which are locally analytically sectorial over $QC(\mathbf{C})$ (see Theorem 7.32(i) of [6]).

6. Piecewise continuous symbols. Let $PC(\mathcal{L}(\mathcal{H}))$ denote the C^*-algebra of all functions a in $L^\infty(\mathcal{L}(\mathcal{H}))$ having one-sided limits $a(\tau \pm 0)$ at every point $\tau \in \mathbf{T}$. At $\tau \in \mathbf{T}$, a function $a \in PC(\mathcal{L}(\mathcal{H}))$ is locally equivalent to the function $\chi_- a(\tau - 0) + \chi_+ a(\tau + 0)$, where $\chi_\pm$ stand for the characteristic functions of the half-circles $\{\tau e^{\pm i\theta} : 0 < \theta < \pi\}$. Thus, $a \in LAS$ if and only if for every $\tau \in \mathbf{T}$ there are $b_\tau, c_\tau \in G\mathcal{L}(\mathcal{H})$ such that $\mathfrak{H}(b_\tau a(\tau \pm 0) c_\tau)$ is contained in the right open half-plane, and $a \in LGS$ if and only if for every $\tau \in \mathbf{T}$ the line segment

$$[a(\tau-0), a(\tau+0)] = \{(1-\mu)a(\tau-0) + \mu a(\tau+0) \ : \ \mu \in [0,1]\}$$

is contained in $G\mathcal{L}(\mathcal{H})$.

Theorem 5.2 tells us that if $a \in PC(\mathcal{L}(\mathcal{H}))$ is in LAS, then $T(a) \in \Pi\{P_n\}$ if and only if $T(a)$ and $T(\tilde{a})$ are invertible. For matrix-valued functions $a \in PC(\mathcal{L}(\mathcal{H}))$ ($\dim \mathcal{H} < \infty$) it was shown by Clancey [9] that

$$T(a) \text{ is Fredholm} \iff a \in LAS \iff a \in LGS \tag{17}$$

(also see [6, Thm. 4.70]). Consequently, in the $\dim \mathcal{H} < \infty$ case we have the equivalence

$$T(a) \in \Pi\{P_n\} \iff T(a), T(\tilde{a}) \in G\mathcal{L}(l^2(\mathcal{H})) \tag{18}$$

for every $a \in PC(\mathcal{L}(\mathcal{H}))$. We have not been able to carry over (17) and (18) to the $\dim \mathcal{H} = \infty$ case for general $a \in PC(\mathcal{L}(\mathcal{H}))$. However, we will extend (18) and part of (17) to symbols a in $PC(\mathbf{C} + \mathcal{K}(\mathcal{H}))$ (recall the last paragraph of the introduction for the definition of $PC(\mathbf{C}+\mathcal{K}(\mathcal{H}))$). To do this, we present two independent approaches. The first is based on the theory of quarter-plane Toeplitz operators, while the second makes use of an approximation argument.

Here is the first approach. We now identify $l^2(\mathcal{H})$ with the Hilbert-space tensor product $l^2 \otimes \mathcal{H}$, where $l^2 := l^2(\mathbf{C})$. Then if $\chi_\pm$ are as above and $a, b \in \mathcal{L}(\mathcal{H})$ are any operators, we have

$$T(\chi_- a + \chi_+ b) = T(\chi_-) \otimes a + T(\chi_+) \otimes b \tag{19}$$

with $T(\chi_\pm) \in \mathcal{L}(l^2)$. Furthermore, if $\dim \mathcal{H} = \infty$ (and only this case will be considered in the following), we may without loss of generality assume that $\mathcal{H} = l^2$. Thus, we may think of (19) as an operator on $l^2 \otimes l^2$ (the l^2 space of a quarter-plane).

Let B denote the smallest closed subalgebra of $\mathcal{L}(l^2)$ containing the set $\{T(\varphi) : \varphi \in PC(\mathbf{C})\}$ and let $\mathsf{B} \otimes \mathsf{B}$ be the closure in $\mathcal{L}(l^2 \otimes l^2)$ of the set of all operators that are representable as finite sums of the form $\sum_j U_j \otimes V_j$ ($U_j, V_j \in \mathsf{B}$). In other words, $\mathsf{B} \otimes \mathsf{B}$ is the C^*-algebra tensor product of two copies of B. Hence, if in (19) the operators a and b lie in B , then $T(\chi_-) \otimes a + T(\chi_+) \otimes b$ belongs to $\mathsf{B} \otimes \mathsf{B}$, and operators in $\mathsf{B} \otimes \mathsf{B}$ are fairly well understood (see [11] or [6, pp. 363–369]).

THEOREM 6.1. *Let $a \in PC(\mathsf{B})$. Then*

$$T(a) \textit{ is } \mathcal{Q}\textit{-Fredholm} \iff a \in LGS.$$

Proof. For $\tau \in \mathbf{T}$, put $A_\tau = T(\chi_-) \otimes a(\tau-0) + T(\chi_+) \otimes a(\tau+0)$. Standard application of localization techniques gives that $T(a) + \mathcal{Q}$ is in $G(\mathcal{A}/\mathcal{Q})$ if and only if $A_\tau + \mathcal{Q}$ is in $G(\mathcal{A}/\mathcal{Q})$ for every $\tau \in \mathbf{T}$. Since $A_\tau \in \mathsf{B}\otimes\mathsf{B}$ and $\mathsf{B}\otimes\mathsf{B}$ is a C^*-subalgebra of $\mathcal{A}$, we obtain

$$\begin{aligned}&A_\tau + \mathcal{Q} \in G(\mathcal{A}/\mathcal{Q})\\ &\iff A_\tau + \mathcal{Q} \in G((\mathsf{B}\otimes\mathsf{B} + \mathcal{Q})/\mathcal{Q})\\ &\iff A_\tau + (\mathsf{B}\otimes\mathsf{B}\cap\mathcal{Q}) \in G(\mathsf{B}\otimes\mathsf{B}/(\mathsf{B}\otimes\mathsf{B}\cap\mathcal{Q})).\end{aligned}$$

Denote the compact operators on l^2 by $\mathcal{K} := \mathcal{K}(l^2)$. It is well known that $\mathcal{K}$ is a subset of B . When interpreting $l^2(\mathcal{H})$ as $l^2 \otimes \mathcal{H}$, the operators Q_n we have been working with so far may be identified as $Q_n \otimes I$. Using this, it is easy to show that $\mathsf{B}\otimes\mathsf{B}\cap\mathcal{Q}$ is nothing else than $\mathcal{K}\otimes\mathsf{B}$. Hence,

$$A_\tau + \mathcal{Q} \in G(\mathcal{A}/\mathcal{Q}) \iff A_\tau + \mathcal{K}\otimes\mathsf{B} \in G(\mathsf{B}\otimes\mathsf{B}/\mathcal{K}\otimes\mathsf{B}).$$

Invertibility criteria in $\mathsf{B}\otimes\mathsf{B}/\mathcal{K}\otimes\mathsf{B}$ were established in [11] (see also [6, pp. 363–369] for an alternative approach). All we need is the following. Suppose we are given a finite sum $\sum B_j \otimes C_j \in \mathsf{B}\otimes\mathsf{B}$ in which each B_j is a Toeplitz operator, say $T(b_j)$. Then

$$\begin{aligned}&\sum T(b_j)\otimes C_j + \mathcal{K}\otimes\mathsf{B} \in G(\mathsf{B}\otimes\mathsf{B}/\mathcal{K}\otimes\mathsf{B})\\ &\iff \sum((1-\mu)b_j(t-0) + \mu b_j(t+0))C_j \in G\mathsf{B} \quad \forall(t,\mu) \in \mathbf{T}\times[0,1].\end{aligned}$$

Application of this criterion to A_τ produces

$$\begin{aligned}&A_\tau + \mathcal{K}\otimes\mathsf{B} \in G(\mathsf{B}\otimes\mathsf{B}/\mathcal{K}\otimes\mathsf{B})\\ &\iff (1-\mu)a(\tau-0) + \mu a(\tau+0) \in G\mathsf{B} \quad \forall\mu \in [0,1]\\ &\iff [a(\tau-0), a(\tau+0)] \subset G\mathcal{L}(\mathcal{H})\end{aligned}$$

(for the last equivalence, note that B is a C^*-subalgebra of $\mathcal{L}(\mathcal{H})$). □

THEOREM 6.2. *Let $a \in PC(\mathsf{B})$ and $K \in \mathcal{Q}$. Then*

$$T(a) + K \in \Pi\{P_n\} \iff T(a) + K,\ T(\tilde{a}) \in G\mathcal{L}(l^2(\mathcal{H})).$$

Proof. Suppose $T(a) + K$ and $T(\tilde{a})$ are invertible. By Theorem 4.5, we are left with showing that

$$\{T_n(a_\tau)\} + \mathbf{J} \in G(\mathbf{S}/\mathbf{J}) \quad \forall\tau \in \mathbf{T},$$

where $a_\tau = \chi_- a(\tau-0) + \chi_+ a(\tau+0)$. We have

$$T_n(a_\tau) = T_n(\chi_-)\otimes a(\tau-0) + T_n(\chi_+)\otimes a(\tau+0). \tag{20}$$

The algebra $\mathbf{F}$ defined in §4 for $l^2(\mathcal{H})$ will be denoted by F in case $\mathcal{H} = \mathbf{C}$. Let S stand for the smallest closed subalgebra of F containing all sequences $\{T_n(\varphi)\}$ with $\varphi \in PC(\mathbf{C})$ and let J be the collection of all sequences of the form $\{P_nKP_n + W_nLW_n + C_n\}$ with $K, L \in \mathcal{K} := \mathcal{K}(l^2)$ and $\|C_n\| \to 0$. One can show (see [6, Prop. 7.27]) that J is a closed two-sided ideal of S.

Now recall that $l^2 \otimes l^2$ may be identified with $l^2(\mathbf{Z}_+ \times \mathbf{Z}_+)$, where $\mathbf{Z}_+ \times \mathbf{Z}_+$ is the discrete quarter-plane. Interpreting $l^2 \otimes l^2$ in this way allows us to think of the projections $P_n : l^2(l^2) \to l^2(l^2)$ defined by (3) as acting by the rule

$$\{g_{jk}\}_{j,k\geq 0} \mapsto \{h_{jk}\}_{j,k\geq 0}, \quad h_{j,k} = \begin{cases} g_{jk} & \text{if } j \leq n, \\ 0 & \text{if } j > n. \end{cases}$$

It is therefore convenient and also fits in with the notation of the preceding paragraph to replace P_n by $P_n \otimes I$ in this context.

Denote by $\mathbf{Y}$ the C^*-algebra of all sequences $\{A_n\}_{n=0}^\infty$ of operators $A_n \in \mathcal{L}(\mathrm{Im}(P_n \otimes I))$ such that $\|\{A_n\}\| := \sup_{n\geq 0} \|A_n\| < \infty$. For example, if $T_n(a_\tau)$ is given by (20), then $\{T_n(a_\tau)\}_{n=0}^\infty \in \mathbf{Y}$. Moreover, given a finite collection of sequences $\{U_n^{(j)}\} \in \mathsf{S}$ and a finite collection of operators $V^{(j)} \in \mathsf{B}$, it is clear that the sequence

$$\left\{\sum_j U_n^{(j)} \otimes V^{(j)}\right\}_{n=0}^\infty \tag{21}$$

belongs to $\mathbf{Y}$. The closure in $\mathbf{Y}$ of all sequences of the form (21) with $\{U_n^{(j)}\} \in \mathsf{S}$ and $V^{(j)} \in \mathsf{B}$ is denoted by $\mathsf{S}\otimes\mathsf{B}$, while $\mathsf{J}\otimes\mathsf{B}$ will stand for the closure in $\mathbf{Y}$ of the set of all sequences of the form (21) with $\{U_n^{(j)}\} \in \mathsf{J}$ and $V^{(j)} \in \mathsf{B}$. The sequence $\{T_n(a_\tau)\}_{n=0}^\infty$ defined by (20) obviously belongs not only to $\mathbf{Y}$ but even to $\mathsf{S} \otimes \mathsf{B}$. An invertibility criterion in $\mathsf{S}\otimes\mathsf{B}/\mathsf{J}\otimes\mathsf{B}$ can be obtained by the method of [6, pp. 381–389]. We merely quote the result for sequences of the form $\{\sum T_n(b_j) \otimes C_j\}_{n=0}^\infty$. One has

$$\begin{aligned} &\left\{\sum T_n(b_j) \otimes C_j\right\} + \mathsf{J} \otimes \mathsf{B} \in G(\mathsf{S} \otimes \mathsf{B}/\mathsf{J} \otimes \mathsf{B}) \\ &\iff \sum((1-\mu)b_j(t-0) + \mu b_j(t+0))C_j \in G\mathsf{B} \quad \forall(t,\mu) \in \mathbf{T} \times [0,1]. \end{aligned}$$

In the special case (19), we get, as in the proof of Theorem 6.1,

$$\{T_n(a_\tau)\} + \mathsf{J} \otimes \mathsf{B} \in G(\mathsf{S} \otimes \mathsf{B}/\mathsf{J} \otimes \mathsf{B}) \iff [a(\tau-0), a(\tau+0)] \subset G\mathcal{L}(\mathcal{H}).$$

Since $T(a) + K$ is invertible, we deduce from Theorem 6.1 that $[a(\tau-0), a(\tau+0)]$ is contained in $G\mathcal{L}(\mathcal{H})$ for every $\tau \in \mathbf{T}$. Hence, $\{T_n(a_\tau)\} + \mathsf{J} \otimes \mathsf{B}$ is invertible in $\mathsf{S}\otimes\mathsf{B}/\mathsf{J}\otimes\mathsf{B}$, and because clearly $\mathsf{J}\otimes\mathsf{B} \subset \mathbf{J}$ and $\mathsf{S}\otimes\mathsf{B} \subset \mathbf{S}$, it follows that all the more $\{T_n(a_\tau)\} + \mathbf{J}$ is invertible in $\mathbf{S}/\mathbf{J}$. □

We emphasize once again that $\mathbf{C} + \mathcal{K}(\mathcal{H}) \subset \mathsf{B}$ and therefore Theorems 6.1 and 6.2 hold in particular for $a \in PC(\mathbf{C} + \mathcal{K}(\mathcal{H}))$.

Let us finally present another approach to prove (18) for a in $PC(\mathbf{C}+\mathcal{K}(\mathcal{H}))$. We denote by $\mathbf{B}$ the smallest C^*-subalgebra of $\mathbf{S}$ containing all sequences $\{T_n(a)\}$ with $a \in PC(\mathbf{C} + \mathcal{K}(\mathcal{H}))$ and we let $\mathbf{G}$ stand for the collection of all sequences $\{C_n\} \in \mathbf{B}$ such that $C_n \rightrightarrows 0$ as $n \to \infty$. Clearly, $\mathbf{G}$ is a closed two-sided ideal of $\mathbf{B}$. Furthermore, let $\mathcal{B}$ denote the smallest C^*-subalgebra of $\mathcal{L}(l^2(\mathcal{H}))$ containing all operators $T(a)$ with $a \in PC(\mathbf{C} + \mathcal{K}(\mathcal{H}))$ and make $\mathcal{B} \times \mathcal{B}$ become a C^*-algebra by defining the algebraic operations on the ordered pairs $(T_1, T_2) \in \mathcal{B} \times \mathcal{B}$ componentwise and the norm as $\|(T_1, T_2)\| = \max\{\|T_1\|, \|T_2\|\}$. The mapping $\sigma : \mathbf{B} \to \mathcal{B} \times \mathcal{B}$ which associates with every sequence $\{A_n\}$ the pair of the strong limits $(\lim_{n\to\infty} A_n, \lim_{n\to\infty} W_n A_n W_n)$ is a C^*-algebra homomorphism. Since $\mathbf{G}$ is contained in the kernel of σ, we have a well-defined C^*-algebra homomorphism

$$\sigma^\pi : \mathbf{B}/\mathbf{G} \to \mathcal{B} \times \mathcal{B}, \quad \{A_n\} + \mathbf{G} \mapsto \sigma(\{A_n\}).$$

A dense subset of $\mathbf{B}$ is the set $\mathbf{B}_0$ of all sequences $\{A_n\}$ which are finite sums of finite products of the form

$$A_n = \sum_i \prod_j T_n(a_{ij}), \tag{22}$$

where $a_{ij} \in PC(\mathbf{C} + \mathcal{K}(\mathcal{H}))$ have at most finitely many jumps. Notice that

$$\sigma^\pi(\{A_n\} + \mathbf{G}) = \left(\sum_i \prod_j T(a_{ij}), \sum_i \prod_j T(\tilde{a}_{ij})\right) \tag{23}$$

in case A_n is given by (22).

THEOREM 6.3. *The mapping $\sigma^\pi : \mathbf{B}/\mathbf{G} \to \mathcal{B} \times \mathcal{B}$ is an isometrical C^*-algebra isomorphism of $\mathbf{B}/\mathbf{G}$ onto $\sigma^\pi(\mathbf{B}/\mathbf{G})$.*

Proof. Let $\{e_1, e_2, \ldots\}$ be any orthonormal basis of $\mathcal{H}$ and denote by S_k the orthonormal projection of $\mathcal{H}$ onto the linear hull of $\{e_1, \ldots, e_k\}$. Further, let $\mathbf{B}_k$ denote the smallest C^*-subalgebra of $\mathbf{B}$ containing all sequences of the form $\{T_n(fI + S_k h S_k)\}_{n=0}^{\infty}$, where $f \in PC(\mathbf{C})$ and $h \in PC(\mathcal{K}(\mathcal{H}))$. From [20, remark on p. 39], one can easily derive that

$$\|\sigma^\pi(\{A_n\} + \mathbf{G})\| = \|\{A_n\} + \mathbf{G}\| \tag{24}$$

for every sequence $\{A_n\} \in \mathbf{B}_k$. Now consider A_n of the form (22) and write $a_{ij} = f_{ij} I + h_{ij}$ with $f_{ij} \in PC(\mathbf{C})$ and $h_{ij} \in PC(\mathcal{K}(\mathcal{H}))$. Since h_{ij} has at most finitely many jumps, the essential range of h_{ij} is a compact subset of $\mathcal{K}(\mathcal{H})$. So Lemma 4.1 of [7] implies that $a_{ij}^k := f_{ij} I + S_k h_{ij} S_k$ converges in $PC(\mathcal{K}(\mathcal{H}))$ to a_{ij} as $k \to \infty$. Using (24) for $\{A_n\} \in \mathbf{B}_k$, we obtain

$$\left\|\sigma^\pi\left(\left\{\sum_i \prod_j T_n(a_{ij}^k)\right\} + \mathbf{G}\right)\right\| = \left\|\left\{\sum_i \prod_j T_n(a_{ij}^k)\right\} + \mathbf{G}\right\|,$$

and passage to the limit $k \to \infty$ yields

$$\left\|\sigma^\pi\left(\left\{\sum_i \prod_j T_n(a_{ij})\right\} + \mathbf{G}\right)\right\| = \left\|\left\{\sum_i \prod_j T_n(a_{ij})\right\} + \mathbf{G}\right\|$$

and thus (24) for every sequence $\{A_n\} \in \mathbf{B}_0$. Since $\mathbf{B}_0$ is dense in $\mathbf{B}$, we arrive at the assertion. □

THEOREM 6.4. *Suppose $a_{ij} \in PC(\mathbf{C} + \mathcal{K}(\mathcal{H}))$. Then*

$$\sum_i \prod_j T(a_{ij}) \in \Pi\left\{\sum_i \prod_j T_n(a_{ij})\right\} \tag{25}$$

if and only if both $\sum_i \prod_j T(a_{ij})$ and $\sum_i \prod_j T(\tilde{a}_{ij})$ are invertible. In particular, if $a \in PC(\mathbf{C} + \mathcal{K}(\mathcal{H}))$ and $K \in \mathcal{Q}$, then

$$T(a) + K \in \Pi\{P_n\} \iff T(a) + K,\ T(\tilde{a}) \in G\mathcal{L}(l^2(\mathcal{H})).$$

Proof. Define A_n by (22). The inclusion (25) holds if and only if $\{A_n\} + \mathbf{G}$ is invertible in $\mathbf{B}/\mathbf{G}$, and from Theorem 6.3 we infer that $\{A_n\} + \mathbf{G}$ is invertible if and only if the two operators on the right of (23) are invertible. □

Acknowledgment. This paper was strongly inspired by the article [14]. We wish to thank I. Gohberg and M. A. Kaashoek for a fruitful discussion on the topic and for useful comments on a preliminary version of the present paper.

REFERENCES

[1] E. AZOFF AND K. F. CLANCEY, *Toeplitz operators with sectorial matrix-valued symbol*, Indiana Univ. Math. J., 26 (1977), pp. 933–938.

[2] H. BART, I. GOHBERG, AND M. A. KAASHOEK, *The coupling method for solving integral equations*, Oper. Theory Adv. Appl., 12 (1984), pp. 39–73; *Addendum*, Integral Equations Operator Theory, 8 (1985), pp. 890–891.

[3] A. BÖTTCHER, *Fredholmness and finite section method for Toeplitz operators on* $l^p(\mathbf{Z}^2_{++})$ *with piecewise continuous symbols*, Z. Anal. Anwendungen, 3 (1984), pp. 97–110, 191–202.

[4] A. BÖTTCHER AND B. SILBERMANN, *Notes on the asymptotic behavior of block Toeplitz matrices and determinants*, Math. Nachr., 98 (1980), pp. 183–210.

[5] ———, *Invertibility and Asymptotics of Toeplitz Matrices*, Akademie-Verlag, Berlin, 1983.

[6] ———, *Analysis of Toeplitz Operators*, Akademie-Verlag, Berlin, 1989 and Springer-Verlag, Berlin, Heidelberg, New York, 1990.

[7] ———, *Operator-valued Szegö–Widom limit theorems*, Oper. Theory Adv. Appl., 71 (1994), pp. 33–53.

[8] A. BÖTTCHER, B. SILBERMANN, AND H. WIDOM, *A continuous analogue of the Fisher–Hartwig formula for piecewise continuous symbols*, J. Funct. Anal., 122 (1994), pp. 222–246.

[9] K. F. CLANCEY, *A local result for systems of Riemann–Hilbert barrier problems*, Trans. Amer. Math. Soc., 200 (1974), pp. 315–325.

[10] A. DEVINATZ AND M. SHINBROT, *General Wiener–Hopf operators*, Trans. Amer. Math. Soc., 145 (1969), pp. 467–494.

[11] R. DUDUCHAVA, *Discrete convolution operators on the quarter-plane and their indices*, Math. USSR-Izv., 11 (1977), pp. 1072–1084.

[12] I. GOHBERG AND I. A. FELDMAN, *Convolution Equations and Projection Methods for Their Solution*, Nauka, Moscow, 1971 (in Russian); Transl. Math. Monographs 41, American Mathematical Society, Providence, RI, 1974 (in English).

[13] I. GOHBERG AND M. A. KAASHOEK, *Asymptotic formulas of Szegö–Kac–Achiezer type*, Asymptotic Anal., 5 (1992), pp. 187–220.

[14] ———, *Projection method for block Toeplitz operators with operator-valued symbols*, Oper. Theory Adv. Appl., 71 (1994), pp. 79–104.

[15] M. G. KREIN, *On some new Banach algebras and theorems of Wiener–Levy type for Fourier series and integrals*, Mat. Issled., 1 (1966), pp. 82–109 (in Russian); Amer. Math. Soc. Transl. Ser. 2, 93 (1970), pp. 82–109.

[16] L. B. PAGE, *Bounded and compact vectorial Hankel operators*, Trans. Amer. Math. Soc., 150 (1970), pp. 529–539.

[17] M. RABINDRANATHAN, *On the inversion of Toeplitz operators*, J. Math. Mech., 19 (1969), pp. 195–206.

[18] D. SARASON, *Function Theory on the Unit Circle*, conference lecture notes, Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1978.

[19] B. SILBERMANN, *Lokale Theorie des Reduktionsverfahrens für Toeplitzoperatoren*, Math. Nachr., 104 (1981), pp. 137–146.

[20] ———, *On the limiting set of singular values of Toeplitz matrices*, Linear Algebra Appl., 182 (1993), pp. 35–43.

[21] F.-O. SPECK, *General Wiener–Hopf Factorization Methods*, Pitman Research Notes 119, Pitman, Boston, London, Melbourne, 1985.

[22] S. R. TREIL, *Geometric aspects of the theory of Hankel and Toeplitz operators*, Dissertation, Leningrad State University, St. Petersburg, Russia, 1985 (in Russian).

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 823–834, May 1996

012

SHARP ESTIMATES FOR COMPLETE ELLIPTIC INTEGRALS*

S.-L. QIU† AND M. K. VAMANAMURTHY‡

Abstract. Monotonicity and convexity properties of certain functions defined in terms of complete elliptic integrals are studied and sharp functional inequalities for these functions are obtained, thus answering some open questions.

Key words. concave, convex, elliptic integral

AMS subject classifications. 33A25

1. Introduction.

For $r \in (0,1)$, let $r' = (1-r^2)^{1/2}$. The functions

$$K(r) = \int_0^{\pi/2} (1 - r^2 \sin^2 t)^{-1/2} dt, \; K'(r) = K(r') \tag{1.1}$$

and

$$E(r) = \int_0^{\pi/2} (1 - r^2 \sin^2 t)^{1/2} dt, \; E'(r) = E(r') \tag{1.2}$$

are called the complete elliptic integrals of the first kind and second kind, respectively [BF], [BO], [BB]. Basic properties of these functions can be found in [WW].

It is well known that these elliptic integrals are indispensable tools for many applications in mathematics, physics, and engineering [C]. They play an important role in quasiconformal theory (see [LV], [AVV1]–[AVV4], [Vu], and [Q1]). Recently, many new properties were obtained for these functions and several conjectures and open problems were put forward [VV1], [VV2], [AVV5]–[AVV7]. Among them are the following three conjectures:

(i) The function $\sqrt[4]{1+r}K(r)/K(\sqrt{r})$ is increasing from [0,1) onto $[1, \sqrt[4]{2})$.

(ii) The function $r'e^{K(r)}$ is strictly concave on $(0,1)$.

(iii) For each $r \in (0,1)$,

$$1 + \frac{r'^2}{8} < \frac{K(r)}{\log(4/r')} < 1 + \frac{r'^2}{4}. \tag{1.3}$$

Conjecture (i) appears in [AVV6] and [AVV7], while (ii) and (iii) appear in [AVV7].

In this paper, we derive some monotonicity, concavity, and convexity properties of certain functions defined in terms of elliptic integrals, from which some sharp functional inequalities follow. In particular, we shall prove that conjectures (i)–(iii) are true. By presenting a double inequality stronger than (1.3), we shall derive lower and upper bounds for the function $(8+r^2)K(r)/\log(4/r')$ which improve the inequalities

$$9 < \frac{(8+r^2)K(r)}{\log(4/r')} < 9.1 \tag{1.4}$$

for $r \in (0,1)$. We observe that (1.4) is Conjecture (5) in [AVV6].

*Received by the editors December 29, 1993; accepted for publication (in revised form) September 23, 1994. This research was performed during the first author's visit to the Mathematics Department of the University of Helsinki with support from CIMO and the Finnish Ministry of Education.

†School of Science and Arts, Hangzhou Institute of Electronics Engineering, Hangzhou, Zhejiang 310037, People's Republic of China.

‡Department of Mathematics, School of Mathematical and Information Sciences, University of Auckland, P.O. Box 92019, Auckland, New Zealand.

We often write K, K' and E, E' instead of $K(r), K'(r)$ and $E(r), E'(r)$, respectively, when the argument of the function is clear from the context.

Some of the main results of this paper are as follows.

THEOREM 1.1. *The function* $f(r) = \sqrt[4]{1+r}\; K(r)/K(\sqrt{r})$ *is strictly increasing from* $[0,1)$ *onto* $[1, \sqrt[4]{2})$.

THEOREM 1.2. *The function* $g(r) = r'e^{K(r)}$ *is strictly decreasing and concave from* $(0,1)$ *onto* $(4, e^{\pi/2})$.

THEOREM 1.3. *For each* $r \in (0,1)$,

$$1 + \frac{1}{8}r'^2 < \frac{K(r)}{\log(4/r')} \tag{1.5}$$

$$< \min\left\{1.013872\left(1+\frac{1}{8}r'^2\right), 1+\frac{1}{4}r'^2\right\}.$$

These two inequalities are asymptotically sharp as r tends to 1.

THEOREM 1.4. *For each* $r \in (0,1)$,

$$9 < \frac{(8+r^2)(8+r'^2)}{8} < \frac{(8+r^2)K(r)}{\log(4/r')} < 9.096. \tag{1.6}$$

The lower bound is sharp, and the upper bound 9.096 *cannot be replaced by a constant less than* 9.09437.

Remark 1. The first inequality in (1.4) has been proven separately by R. Kühnau and S.-L. Qiu (see [K] and [Q2]). Our present proof, however, is very simple as it is an immediate consequence of the first inequality in (1.5).

2. Preliminary results. In this section, we obtain some elementary properties of K and E which are needed for proofs of the main theorems stated in §1.

THEOREM 2.1. (1) *The function* $f_1(r) = \frac{r^2}{E(r)-r'^2K(r)}$ *is strictly decreasing and concave from* $(0,1)$ *onto* $(1, \frac{4}{\pi})$. *In particular, for* $r \in (0,1)$,

$$1 + \left(\frac{4}{\pi} - 1\right)(1-r) < \frac{r^2}{E(r)-r'^2K(r)} < \frac{4}{\pi}.$$

(2) *The function* $f_2(r) = [E(r) - r'^2K(r)]/(1-r')$ *is strictly decreasing and concave from* $(0,1)$ *onto* $(1, \frac{\pi}{2})$. *In particular, for* $r \in (0,1)$,

$$1 + \left(\frac{\pi}{2} - 1\right)(1-r) < \frac{E(r)-r'^2K(r)}{1-r'} < \min\left\{\frac{\pi}{2}, 1+r'\right\}. \tag{2.1}$$

(3) *The function* $f_3(r) = \frac{1-(E(r)-r'^2K(r))}{r'^2\log(\frac{4}{r'})}$ *is strictly decreasing from* $(0,1)$ *onto* $(\frac{1}{2}, \frac{1}{\log 4})$.

(4) *The function* $f_4(r) = (a+r^2)E(r) - (a-r^2)K(r)$ *is strictly increasing (decreasing) on* $(0,1)$ *if and only if* $a \le 1$ $(a \ge 4)$. *Moreover,*

$$f_4((0,1)) = \begin{cases} (0,\infty) & \textit{if } a < 1, \\ (0,2) & \textit{if } a = 1, \\ (-\infty, 0) & \textit{if } a \ge 4. \end{cases}$$

(5) *For each* $a \in [4,\infty), f_5(r) = E(r)/(a-r^2)$ *is strictly decreasing from* $[0,1]$ *onto* $[\frac{1}{a-1}, \frac{\pi}{2a}]$.

(6) *The function* $f_6(r) = r'^2(K(r)-E(r))/r^2$ *is strictly decreasing and concave from* $(0,1)$ *onto* $(0, \frac{\pi}{4})$.

(7) *The function* $f_7(r) = (E(r)E'(r) - K(r)E'(r) + r^2K(r)K'(r))/r^2$ *is strictly decreasing from* $(0,1)$ *onto* $(\frac{\pi}{2}, \infty)$.

(8) *The function* $f_8(r) = (K(r)-E(r)+\log r')/r^2$ *is strictly increasing from* $(0,1)$ *onto* $(\frac{\pi-2}{4}, \log 4 - 1)$.

Proof. (1) By differentiation,

$$\frac{-f_1'(r)}{r} = \frac{(K-E)-(E-r'^2K)}{(E-r'^2K)^2},$$

which is positive and has the form $0/0$ at $r = 0$. The ratio of the derivatives of the numerator and the denominator is $1/(2r'^2K)$ so that the result (1) follows from [AVV5, Thm. 2.2(3)] and the l'Hôpital monotone rule [AVV7, Thm. 1.24].

The limiting values are clear.

(2) By differentiation,

$$f_2'(r) = -\frac{r}{r'} \cdot \frac{E-r'K}{(1-r')^2}.$$

Since $(E-r'K)/(1-r')^2$ is strictly increasing from $(0,1)$ onto $(\frac{\pi}{8}, 1)$ [AVV6, Thm. 3.7], $f_2'(r)$ is negative and strictly decreasing on $(0,1)$, yielding the monotonicity and concavity of f_2.

Next, from the equality

$$f_2(r) = (1+r') \int_0^{\pi/2} \frac{\cos^2 t}{\sqrt{1-r^2\sin^2 t}} dt,$$

we see that $f_2(0) = \frac{\pi}{2}$ and $f_2(1) = 1$.

(3) The function f_3 has the form $0/0$ at $r = 1$. The ratio of the derivatives of the numerator and the denominator $= \frac{K}{2\log(4/r')-1}$, which is strictly decreasing, by [AVV5, Thm. 2.2(5)], so that f_3 has the same property by the l'Hôpital monotone rule [AVV7, Thm. 1.24]. The value $f_3(0) = 1/\log 4$ is clear, while $f_3(1) = \frac{1}{2}$ by l'Hôpital's rule.

(4) By differentiation and simplification,

$$f_4'(r) = \frac{r}{r'^2}(4-a-3r^2)E,$$

so that the assertions follow.

(5) Since $f_4(0) = 0$ and

$$r(a-r^2)^2 f_5'(r) = f_4(r),$$

the monotonicity of f_5 follows by (4). The limiting values are clear.

(6) By differentiation and simplification,

$$-f_6'(r) = \frac{2(K-E)-r^2E}{r^3},$$

which has the form $0/0$ at $r = 0$. The ratio of derivatives of the numerator and the denominator is

$$\frac{2rE}{3r'^2} + \frac{K-E}{3r},$$

which is positive and increasing by [AVV6, Thms. 1.3 and 2.1(6)]. Hence the result follows by the l'Hôpital monotone rule [AVV7, Thm. 1.24].

(7) By differentiation and simplification,

$$f_7'(r) = -\frac{2}{r^3 r'^2}(E-r'^2K)(E'-r^2K') < 0,$$

from which the monotonicity of f_7 follows. Since

$$f_7(r) = E'\frac{E - r'^2K}{r^2} + K(K' - E'),$$

we see that $f_7(0) = \infty$ and $f_7(1) = \frac{\pi}{2}$.

(8) The function f_8 has the form $0/0$ at $r = 0$. The ratio of the derivatives of the numerator and the denominator is $(E - 1)/(2r'^2)$, which again has the form $0/0$ at $r = 1$. The ratio of the derivatives of the numerator and the denominator is $(K - E)/(4r^2)$, which is strictly increasing by definitions (1.1) and (1.2). Hence f_8 also has the same property by the l'Hôpital monotone rule [AVV7, Thm. 1.24]. The end values are clear. □

LEMMA 2.2. (1) *The function* $f_9(r) = (9 - r^2)E(r) - 3(3 - r^2)r'^2K(r)$ *is strictly increasing from* $[0, 1)$ *onto* $[0, 8)$.

(2) *On* $(0, 1)$, *the function*

$$f_{10}(r) = 3(-75 + 34r^2 - 15r^4)r'^2K(r) + (225 - 214r^2 + 21r^4)E(r)$$

has exactly two zeros r_1 *and* r_2, $r_1 < r_2$. *Moreover,* $f_{10}(r) > 0$ *for* $r \in (0, r_1) \cup (r_2, 1)$, *and* $f_{10}(r) < 0$ *for* $r \in (r_1, r_2)$.

(3) *On* $(0, 1)$, *the function*

$$f_{11}(r) = (162 - 153r^2 + 32r^4 - 9r^6)K(r) - 6(9 - r^2)(3 - r^2)E(r)$$

has exactly two zeros r_3, r_4, $r_3 < r_4$. *Moreover,* $f_{11}(r) > 0$ *for* $r \in (0, r_3) \cup (r_4, 1)$, *and* $f_{11}(r) < 0$ *for* $r \in (r_3, r_4)$.

(4) *The function* $f_{12}(r) = \log(4/r') - \{r^2(9 - r^2)K(r)/f_9(r)\}$ *is strictly increasing on* $(0, r_3]$ *and on* $(r_4, 1)$, *and strictly decreasing on* (r_3, r_4).

(5) *On* $(0, 1)$, *the function* $f_{12}(r)$ *has a unique zero* $r_0 \in (\sin 29°, 1/2)$ *such that* $f_{12}(r) > 0$ *for* $r \in (0, r_0)$ *and* $f_{12}(r) < 0$ *for* $r \in (r_0, 1)$.

Proof. (1) Since

$$f_9'(r) = r(13 - 9r^2)K(r) > 0, \tag{2.2}$$

the monotonicity of f_9 follows. Clearly, $f_9(0) = f_9(1) - 8 = 0$.

(2) By differentiation and simplification,

$$\frac{1}{r}f_{10}'(r) = g_1(r), \tag{2.3}$$

where $g_1(r) = (-540 + 60r^2)E + (541 - 462r^2 + 225r^4)K$, and

$$\frac{r'^2}{r}g_1'(r) = g_2(r), \tag{2.4}$$

where $g_2(r) = [(1 + 258r^2 + 45r^4)(E - r'^2K)/r^2] - (264 - 720r^2)r'^2K$. Clearly, $g_1(0) = \pi/2$, $g_1(1) = \infty$, $g_2(0) = -527\pi/4$, and $g_2(1) = 304$.

By [AVV5, Thm. 2.2(3)(7)] or Theorem 2.1(1), g_2 is strictly increasing on $[0, b]$, where $b = \sqrt{264/720} = \sqrt{11/30} > \sin 37°$. Since

$$g_2(\sin 37°) = 79.37\ldots > 0,$$

g_2 has a zero $r_5 \in (0, b)$ such that $g_2(r) < 0$ for $r \in [0, r_5)$ and $g_2(r) > 0$ for $r \in (r_5, b)$.

On the other hand, for $r \in (b,1)$, it is clear that

$$g_2(r) > (1 + 258r^2 + 45r^4)(E - r'^2 K)/r^2 > 0.$$

Hence, r_5 is the unique zero of g_2 in $[0,1]$, $g_2(r) < 0$ for $r \in [0,r_5)$, and $g_2(r) > 0$ for $r \in (r_5,1]$.

It now follows from (2.4) that g_1 is strictly decreasing on $(0,r_5]$ and strictly increasing on $[r_5,1)$. Since

$$g_1(\sin 32^\circ) = -30.09\ldots < 0,$$

g_1 has exactly two zeros $r_6, r_7 \in (0,1)$, $r_6 < r_7$, so that $g_1(r) > 0$ for $r \in [0,r_6) \bigcup (r_7,1]$ and $g_1(r) < 0$ for $r \in (r_6,r_7)$.

Consequently, it follows from (2.3) that f_{10} is strictly increasing on $(0,r_6]$ and on $[r_7,1)$ and strictly decreasing on $[r_6,r_7]$. By computation, we have

$$f_{10}(0) = 0,\ f_{10}\left(\frac{1}{2}\right) = -2.19\ldots \quad \text{and} \quad f_{10}(1) = 32.$$

Hence, the result (2) follows by the piecewise monotonicity of f_{10}.

(3) Clearly, $f_{11}(0) = 0$ and $f_{11}(1) = \infty$. Since

$$f_{11}\left(\frac{1}{2}\right) = -0.118\ldots < 0,$$

the result for f_{11} follows from the derivative

$$\frac{r'^2}{r} f_{11}'(r) = f_{10}(r)$$

by (2).

(4) Assertion (4) follows from the derivative

$$[f_9(r)]^2 f_{12}'(r) = rKf_{11}(r)$$

by (3).

(5) First, we have

$$\begin{aligned} f_{12}(0) &= 2\log 2 - \frac{9\pi}{2}\lim_{r\to 0} r^2 \left\{ \int_0^{\frac{\pi}{2}} \frac{11r^2 - 3r^4 - r^2(9-r^2)\sin^2 t}{\sqrt{1 - r^2\sin^2 t}}dt \right\}^{-1} \\ &= 2\log 2 - \frac{18}{13} > 0 \end{aligned}$$

and

$$f_{12}(1) = \lim_{r\to 1}\left\{ \left(\log\frac{4}{r'} - K\right) + r'K \cdot \frac{f_9(r) - (9r^2 - r^4)}{r' f_9(r)} \right\} = 0,$$

since

$$\lim_{r\to 1}\left(\log\frac{4}{r'} - K\right) = \lim_{r\to 1} r'K = 0 \tag{2.5}$$

and

$$\lim_{r\to 1}\frac{f_9(r)-r^2(9-r^2)}{r'f_9(r)}=0$$

by l'Hôpital's rule. Hence, by (4), f_{12} has a unique zero r_0 in $(0,1)$.

Next, by computation, we have

$$f_{12}\left(\frac{1}{2}\right)=-0.0002\ldots\quad\text{and}\quad f_{12}(\sin 29°)=0.0007\ldots.$$

Hence, $r_0\in(\sin 29°,\frac{1}{2})$. □

LEMMA 2.3. (1) *The function* $f_{13}(r)=(5-r^2)E(r)-(5-3r^2)r'^2K(r)$ *is strictly increasing from* $[0,1)$ *onto* $[0,4)$.

(2) *The function* $f_{14}(r)=(50-15r^2+9r^4)r'^2K(r)-(50-40r^2+6r^4)E(r)$ *is strictly decreasing from* $[\sqrt{3/23},1)$ *onto* $(-16,f_{14}(\sqrt{3/23})]$.

(3) *The function* $f_{15}(r)=\log(4/r')-[r^2(5-r^2)K(r)/f_{13}(r)]$ *is positive and strictly decreasing on* $[1/\sqrt{2},1)$.

Proof. (1) By differentiation,

$$f_{13}'(r)=9rr'^2K>0,$$

$0<r<1$. Clearly, $f_{13}(0)=0$ and $f_{13}(1)=4$.

(2) We have

$$\begin{aligned}f_{14}'(r)&=3r\{(35-7r^2)E+(-35+26r^2-15r^4)K\}\\&=3r^3\int_0^{\pi/2}\frac{19-15r^2-(35-7r^2)\sin^2 t}{\sqrt{1-r^2\sin^2 t}}dt\\&=21r^3(5-r^2)\int_0^{\pi/2}\frac{[(19-15r^2)/(35-7r^2)]-\sin^2 t}{\sqrt{1-r^2\sin^2 t}}dt.\end{aligned}$$

It is easy to verify that $(19-15r^2)/(35-7r^2)\le\frac{1}{2}$ for $r\in[\sqrt{3/23},1)$. Hence it follows from the above equalities that

$$\begin{aligned}f_{14}'(r)&\le\frac{21}{2}r^3(5-r^2)\int_0^{\pi/2}\frac{\cos 2t}{\sqrt{1-r^2\sin^2 t}}dt\\&=\frac{21}{2}r(5-r^2)[2E-(2-r^2)K]\end{aligned}$$

for $r\in[\sqrt{3/23},1)$, which is negative by Theorem 2.1(1).

Clearly, $f_{14}(1)=-16$.

(3) By differentiation,

$$[f_{13}(r)]^2f_{15}'(r)=rKf_{14}(r). \tag{2.6}$$

Since

$$\begin{aligned}f_{14}\left(\frac{\sqrt{2}}{2}\right)&=\left(50-\frac{21}{4}\right)\frac{1}{2}K\left(\frac{\sqrt{2}}{2}\right)-\left(30+\frac{3}{2}\right)E\left(\frac{\sqrt{2}}{2}\right)\\&=-1.05841\ldots,\end{aligned}$$

$f_{14}(r)<0$ for all $r\in[\frac{\sqrt{2}}{2},1)$ by (2). The result now follows from (2.6). □

3. Proofs of main theorems. In this section, we prove the theorems stated in §1.

Proof of Theorem 1.1. First, it is clear that $f(0) = 1$. By l'Hôpital's rule, we get $f(1) = \sqrt[4]{2}$.

Next, by differentiation,

$$\begin{aligned}(3.1)\qquad & 4r(1-r)(1+r)^{3/4}K^2(\sqrt{r})f'(r)\\ &= 4K(\sqrt{r})[E(r) - r'^2K(r)]\\ &+ K(r)\left\{r(1-r)K(\sqrt{r}) - 2(1+r)[E(\sqrt{r}) - (1-r)K(\sqrt{r})]\right\}.\end{aligned}$$

By Theorem 2.1(2) and [VV2, Lem. 2.2(5)], we have

$$rK(\sqrt{r}) > 2[E(\sqrt{r}) - (1-r)K(\sqrt{r})]$$

for $r \in (0,1)$. Hence, it follows from (3.1) that

$$\begin{aligned}&4r(1-r)(1+r)^{3/4}K^2(\sqrt{r})f'(r)\\ &> 4K(\sqrt{r})[E(r) - r'^2K(r)] - 4rK(r)[E(\sqrt{r}) - (1-r)K(\sqrt{r})]\\ &= 4r^2K(r)K(\sqrt{r})\left\{\frac{E(r) - r'^2K(r)}{r^2K(r)} - \frac{E(\sqrt{r}) - (1-r)K(\sqrt{r})}{rK(\sqrt{r})}\right\}.\end{aligned}$$

Hence $f'(r) > 0$ for each $r \in (0,1)$ by Theorem 2.1(2). □

By Theorem 1.1, one can derive lower and upper bounds for $\mu(r^2)/\mu(r)$, where

$$(3.2)\qquad \mu(r) = \frac{\pi}{2}\frac{K'(r)}{K(r)},$$

$0 < r < 1$, which is a very important special function in the theory of quasiconformal mappings [LV].

COROLLARY 3.1. *For each* $r \in (0,1)$,

$$(3.3)\qquad 1 < \sqrt[4]{\frac{2}{1+r^2}} < \frac{\mu(r^2)}{\mu(r)} < \frac{2}{(1+r^2)^{3/4}} < 2.$$

Proof. By [AVV7, Thm. 3.31], for each $r \in (0,1)$,

$$\sqrt{\frac{2}{1+r^2}} < \frac{K'(r^2)}{K'(r)} < \frac{2}{1+r^2}.$$

Hence, by Theorem 1.1, for each $r \in (0,1)$,

$$\begin{aligned}\frac{\mu(r^2)}{\mu(r)} &= (1+r^2)^{1/4}\frac{K'(r^2)}{f(r^2)K'(r)}\\ &< (1+r^2)^{1/4}\frac{K'(r^2)}{K'(r)} < \frac{2}{(1+r^2)^{3/4}} < 2\end{aligned}$$

and

$$\frac{\mu(r^2)}{\mu(r)} > \left(\frac{1+r^2}{2}\right)^{1/4}\frac{K'(r^2)}{K'(r)} > \left(\frac{2}{1+r^2}\right)^{1/4} > 1.$$

Here $f(r)$ is as in Theorem 1.1. □

Remark 2. In [QV], it was proved that the function $\mu(r^2)/\mu(r)$ is strictly decreasing from $(0,1)$ onto $(1,2)$. Hence

$$1 < \mu(r^2)/\mu(r) < 2,$$

so that (3.3) improves this result.

Proof of Theorem 1.2. First, by differentiation,

$$g'(r) = e^K \frac{E - r'^2 K - r^2}{rr'},$$

which is negative by (2.1), so that $g(r)$ is strictly decreasing on (0,1).

Clearly, $g(0) = e^{\pi/2}$. By (2.5), we have

$$g(1) = \exp\{\lim_{r \to 1}(\log r' + K)\} = 4.$$

Next, we have

$$r'^3 e^{-K} g''(r) = \frac{1}{r^2}\{(E - r'^2 K)^2 + r'^2(K - E)\} - 1. \tag{3.4}$$

Define $G(r) = \{(E - r'^2 K)^2 + r'^2(K - E)\}/r^2$. Now we want to estimate $G(r)$ by investigating two cases.

Case 1: $0 < r \le \sin 68°$. In this case, by Theorem 2.1 (1) and (6), we have

$$G(r) < \left(\frac{E(b) - b'^2 K(b)}{b}\right)^2 + \left(\frac{a'}{a}\right)^2 [K(a) - E(a)] = G_1(a, b)$$

for $r \in (a, b] \subset (0, 1)$. Making use of this inequality, we get the following estimates by computation:

$$\begin{aligned}
G(r) &< G_1(0, \sin 30°) = 0.950\ldots, & r &\in (0, \sin 30°],\\
G(r) &< G_1(\sin 30°, \sin 43°) = 0.984\ldots, & r &\in (\sin 30°, \sin 43°],\\
G(r) &< G_1(\sin 43°, \sin 52°) = 0.994\ldots, & r &\in (\sin 43°, \sin 52°],\\
G(r) &< G_1(\sin 52°, \sin 58°) = 0.986\ldots, & r &\in (\sin 52°, \sin 58°],\\
G(r) &< G_1(\sin 58°, \sin 62°) = 0.977\ldots, & r &\in (\sin 58°, \sin 62°],\\
G(r) &< G_1(\sin 62°, \sin 65°) = 0.976\ldots, & r &\in (\sin 62°, \sin 65°],\\
G(r) &< G_1(\sin 65°, \sin 68°) = 0.987\ldots, & r &\in (\sin 65°, \sin 68°].
\end{aligned}$$

From these inequalities, it follows that

$$G(r) < 1 \text{ for } r \in (0, \sin 68°]. \tag{3.5}$$

Case 2: $\sin 68° < r < 1$. In this case, we consider the function $G_2(r) = r^2 G(r) - r^2$. Since $E \ge 1$, we have

$$\begin{aligned}
\frac{1}{r} G_2'(r) &= 2K(E - r'^2 K - 1) + 3E - 2\\
&> 1 - 2r'K \cdot \frac{1 - (E - r'^2 K)}{r'} = G_3(r), \text{ say.}
\end{aligned} \tag{3.6}$$

Since $r'K$ is strictly decreasing on $(0,1), G_3(r)$ is strictly increasing on (0,1) by Theorem 2.1(3).

Hence, for $r \in (\sin 68^\circ, 1)$,

$$\begin{aligned} G_3(r) &> 1 + 2K(\sin 68^\circ)[E(\sin 68^\circ) - \cos^2 68^\circ K(\sin 68^\circ) - 1] \\ &= 0.015\ldots > 0, \end{aligned}$$

showing that $G_2(r)$ is strictly increasing on $(\sin 68^\circ, 1)$ by (3.6).

Since $G_2(1) = 0$, we get

$$G_2(r) < 0 \text{ for } r \in (\sin 68^\circ, 1),$$

or equivalently,

$$G(r) < 1 \text{ for } r \in (\sin 68^\circ, 1). \tag{3.7}$$

Now it follows from (3.4) that $g''(r) < 0$, for each $r \in [0,1)$, by (3.5) and (3.7). This yields the concavity of g and completes the proof. □

COROLLARY 3.2. *For each* $r \in (0,1)$,

$$1 - r < \frac{r' e^{K(r)} - 4}{e^{\pi/2} - 4} < 1.$$

The lower and upper bounds are asymptotically sharp as r *tends to* 1 *and* 0, *respectively.*

Proof of Theorem 1.3. First, we prove the following inequalities:

$$1 + \frac{1}{8} r'^2 < \frac{K(r)}{\log(4/r')} < 1.013872 \left(1 + \frac{1}{8} r'^2\right) \tag{3.8}$$

for $r \in (0,1)$. For this, consider the function

$$h(r) = \frac{1}{9 - r^2} \frac{K(r)}{\log(4/r')}, \quad 0 < r < 1,$$

with $h(0) = \pi/(36 \log 2)$ and $h(1) = \frac{1}{8} < h(0)$.

By differentiation,

$$h'(r) = f_9(r) f_{12}(r) / \left\{ r \left[r'(9 - r^2) \log \frac{4}{r'} \right]^2 \right\}, \tag{3.9}$$

where f_9 and f_{12} are as in Lemma 2.2. Hence, by Lemma 2.2(1),(4), and (5), h is increasing on $(0, r_0]$ and decreasing on $(r_0, 1)$, and consequently,

$$h(1) < h(r) \le h(r_0) \tag{3.10}$$

for each $r \in [0,1)$, with $r_0 \in (\sin 29^\circ, \frac{1}{2})$.

Next, we have

$$\begin{aligned} h(r_0) &= \frac{1}{9 - r_0^2} \frac{K(r_0)}{\log(4/r_0')} < \frac{K(\frac{1}{2})}{(9 - \frac{1}{4}) \log(4/\cos 29^\circ)} \\ &= 0.126733\ldots < 0.126734. \end{aligned} \tag{3.11}$$

Now, (3.8) follows from (3.10) and (3.11). Moreover, the lower bound in (3.8) is asymptotically sharp as r tends to 1.

Next, we prove that

$$\frac{K(r)}{\log(4/r')} < 1 + \frac{1}{4}r'^2, \quad 0 < r < 1. \tag{3.12}$$

It is easy to show that $1.013872(1+\frac{1}{8}r'^2) \le 1+\frac{1}{4}r'^2$ for $r \in [0, a]$, where $a = \sqrt{\frac{54697}{61633}} = 0.94205\ldots$. Hence,

$$\frac{K(r)}{\log(4/r')} < 1 + \frac{1}{4}r'^2 \quad \text{for } r \in (0, a]. \tag{3.13}$$

Next, define

$$H(r) = \frac{1}{5-r^2}\frac{K(r)}{\log(4/r')}, \quad 0 < r < 1.$$

Clearly, $H(0) = \dfrac{\pi}{20\log 2}$ and $H(1) = \frac{1}{4} > H(0)$. By differentiation,

$$rr'^2\left[(5-r^2)\log\frac{4}{r'^2}\right]^2 H'(r) = f_{13}(r)f_{15}(r), \tag{3.14}$$

where f_{13} and f_{15} are as in Lemma 2.3.

By Lemma 2.3(3) for $r \in [\sqrt{2}/2, 1)$,

$$\begin{aligned} f_{15}(r) &> f_{15}(1) \\ &= \lim_{r\to 1}\left\{\left(\log\frac{4}{r'} - K\right) + r'K\frac{f_{13}(r) - r^2(5-r^2)}{r'f_{13}(r)}\right\} = 0 \end{aligned}$$

because of (2.5) and the fact that

$$\lim_{r\to 1}\frac{f_{13}(r) - r^2(5-r^2)}{r'f_{13}(r)} = 0$$

by l'Hôpital's rule. Therefore, by Lemma 2.3(1), it follows from (3.14) that $H(r)$ is strictly increasing on $[\sqrt{2}/2, 1)$, and hence,

$$H(r) < H(1) = \frac{1}{4} \tag{3.15}$$

for $r \in [\sqrt{2}/2, 1)$. In particular, (3.15) holds for $r \in (a, 1)$. Hence,

$$\frac{K(r)}{\log(4/r')} < 1 + \frac{1}{4}r'^2 \text{ for } r \in (a, 1). \tag{3.16}$$

Now, (3.12) follows from (3.13) and (3.16). Finally, (1.5) follows from (3.8) and (3.12). □

Proof of Theorem 1.4. First, by (1.5), we have

$$\frac{(8+r^2)K(r)}{\log(4/r')} > \frac{(8+r^2)(9-r^2)}{8} > 9 \tag{3.17}$$

for $r \in (0,1)$. The two inequalities are sharp at $r = 1$. Thus, the first and second inequalities in (1.6) hold and the lower bound 9 is sharp.

Next, by [Q2, Proof of Thm.], we have

$$\frac{(8+r^2)K(r)}{\log(4/r')} \le \max_{\sin 41^\circ < r < \sin 42^\circ} \frac{(8+r^2)K(r)}{\log(4/r')} \tag{3.18}$$

for each $r \in [0, \sin 41^\circ] \cup [\sin 42^\circ, 1]$.

For $r \in (\sin 41^\circ, \sin 42^\circ), h(r)$ is strictly decreasing by Lemma 2.2(5). Here $h(r)$ is as defined in the proof of Theorem 1.3. On the other hand, it is easy to verify that $(8+r^2)(9-r^2)$ is strictly increasing on $(0, \frac{1}{\sqrt{2}})$ and strictly decreasing on $(\frac{1}{\sqrt{2}}, 1)$. Therefore, for $\sin 41^\circ < r < \sin 42^\circ$,

$$\begin{aligned}
&\frac{(8+r^2)K}{\log(4/r')} = h(r)(8+r^2)(9-r^2) \\
&< \ h(\sin 41^\circ)(8+\sin^2 42^\circ)[9-\sin^2 42^\circ) \\
&\le \frac{(9-0.66905^2)(8+0.66915^2)}{9-0.65615^2}\,\frac{1.79925}{1.6677} \\
&= 9.0959\ldots < 9.096.
\end{aligned}$$

Combining the above estimate with (3.18) yields the third inequality in (1.6).

Finally, the upper bound in (1.6) cannot be replaced with a constant less than 9.09437, since

$$\begin{aligned}
\left.\frac{(8+r^2)K}{\log\frac{4}{r'}}\right|_{r=\sin 41^\circ} &\ge \frac{(8+0.65605^2)\times 1.79915}{\log\frac{4}{0.75465}} \\
&= 9.09437\ldots \qquad \square
\end{aligned}$$

Acknowledgment. The authors wish to thank Dr. M. Vuorinen for hospitality and many helpful discussions.

REFERENCES

[AVV1] G. D. ANDERSON, M. K. VAMANAMURTHY, AND M. VUORINEN, *Dimension-free quasiconformal distortion in n-space*, Trans. Amer. Math. Soc., 297 (1986), pp. 687–706.

[AVV2] ———, *Sharp distortion theorems for quasiconformal mappings*, Trans. Amer. Math. Soc., 305 (1988), pp. 95–111.

[AVV3] ———, *Special functions of quasiconformal theory*, Exposition. Math., 7(1989), pp. 97–136.

[AVV4] ———, *Inequalities for extremal distortion function*, in Proc. 13th Rolf Nevanlinna Cologuium, Joensuu, Finland, 1987, Lecture Notes in Math. 1351, Springer-Verlag, Berlin, New York, (1988), pp. 1–11.

[AVV5] ———, *Functional inequalities for complete elliptic integrals and their ratios*, SIAM J. Math. Anal., 21 (1990), pp. 536–549.

[AVV6] ———, *Functional inequalities for hypergeometric functions and complete elliptic integrals*, SIAM J. Math. Anal., 23 (1992), pp. 512–524.

[AVV7] ———, *Conformal Invariants, Inequalities, and Quasiconformal Mappings*, John Wiley, New York, to appear.

[BB] J. M. BORWEIN AND P. B. BORWEIN, *Pi and the AGM*, John Wiley, New York, 1987.

[BF] P. F. BYRD AND M. D. FRIEDMAN, *Handbook of Elliptic Integrals for Enginers and Physicists*, Grundlehren Math. Wiss. 57, Springer-Verlag, Berlin, New York, 1954.

[BO] F. BOWMAN, *Introduction to Elliptic Functions with Applications*, Dover, New York, 1961.

[C] B. C. CARLSON, *Special Functions of Applied Mathematics*, Academic Press, New York, 1977.

[K] R. KÜHNAU, *Eine methode, die positivität einer funktion zu prüfen*, Z. Angew. Math. Mech., 74 (1994), pp. 140–142.

[LV] O. LEHTO AND K. I. VIRTANEN, *Quasiconformal Mappings in the Plane*, 2nd ed., Grundlehren Math. Wiss 126, Springer-Verlag, New York, Berlin, 1973.

[Q1] S.-L. Qiu, *Distortion properties of K-q.c. maps and a better estimate of Mori's constant*, Acta Math. Sinica, 35 (1992), pp. 492–504 (in Chinese).

[Q2] ———, *The proof of a conjecture on the first elliptic integrals*, J. Hangzhou Inst. Elec. Engrg., 3 (1993), pp. 29–36.

[QV] S.-L. Qiu and M. Vuorinen, *Submultiplicative properties of the φ_K-distortion function*, Studia Math., to appear.

[VV1] M. K. Vamanamurthy and M. Vuorinen, *Functional inequalities, Jacobi products, and quasiconformal maps*, Illinois J. Math. 38 (1994), pp. 394–419.

[VV2] ———, *Inequalities for means*, J. Math. Anal. Appl., 183 (1994), pp. 155–166.

[Vu] M. Vuorinen, *Conformal Geometry and Quasiregular Mappings*, Lecture Notes in Math. 1319, Springer-Verlag, Berlin, New York, 1988.

[WW] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th ed., Cambridge University Press, Cambridge, 1958.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 835–864, May 1996

013

ON SOME SHARP REGULARITY ESTIMATIONS OF L^2-SCALING FUNCTIONS*

KA-SING LAU[†], MANG-FAI MA[†], AND JIANRONG WANG[‡]

Abstract. Let f be a compactly supported L^2-solution of the two-scale dilation equation and α be the L^2-Lipschitz exponent of f. We prove, in addition to other results, that there exists an integer $k \geq 0$ such that (i) $\frac{1}{h^{2\alpha}|\ln h|^k} \int_{-\infty}^{\infty} |f(x+h) - f(x)|^2 dx \approx p(h)$ as $h \to 0^+$, where p is a nonzero bounded continuous function with $p(2h) = p(h)$, and (ii) for $s > \alpha$, there exists a nonzero bounded continuous q (depends on s) with $q(2T) = q(T)$ and $\frac{1}{T^{2(s-\alpha)}(\ln T)^k} \int_{-T}^{T} |\omega^s \hat{f}(\omega)|^2 d\omega \approx q(T)$ as $T \to \infty$. The above α and k can be calculated through a transition matrix. These improve the previous result of Cohen and Daubechies concerning the Besov space containing f and Villemoes's result on the Sobolev exponent of $\hat{f}$.

Key words. asymptotics, compactly supported L^2-solutions, dilation equation, Fourier transformation, Lipschitz exponent, spectral radius, regularity, Tauberian theorem, wavelet

AMS subject classifications. 26A15, 26A18, 39A10, 42A05

1. Introduction.

The existence, regularity, and orthogonality of the compactly supported L^2-solution (notation: L^2_c-solution) of the two-scale dilation equation

$$f(x) = \sum_{n=0}^{N} c_n f(2x - n) \tag{1.1}$$

have been studied in great detail (e.g., [CD], [CH], [D], [DL1], [DL2], [E], [H], [LW1], [V], [W]). In much of the literature, the techniques and emphases are on the frequency domain, i.e., the consideration of the Fourier transformation of (1.1),

$$\hat{f}(\omega) = m_0\left(\frac{\omega}{2}\right) \hat{f}\left(\frac{\omega}{2}\right),$$

where $m_0(\omega) = \frac{1}{2}\sum c_n e^{in\omega}$ and $\hat{f}(\omega) = \int_{-\infty}^{\infty} f(x)e^{-i\omega x}dx$. On the other hand, there are linear algebraic methods on the time domain which also yield many important results concerning continuous solutions ([DL1], [DL2], [CH], [W]) and L^p-solutions [LW1].

In this paper, we continue our study through the second method. For the L^2-case, the existence and regularity results in [CD] and [V] are largely derived from the $(2N-1) \times (2N-1)$ matrix $\mathbf{W}_N$ associated with the operator $\mathbf{A}$ on functions in the frequency domain defined by

$$\mathbf{A}g(\omega) = \left|m_0\left(\frac{\omega}{2}\right)\right|^2 g\left(\frac{\omega}{2}\right) + \left|m_0\left(\frac{\omega}{2} + \pi\right)\right|^2 g\left(\frac{\omega}{2} + \pi\right)$$

(which was introduced in [CR]). The matrix $\mathbf{W}_N$ actually comes out more naturally in the time-domain consideration. For $g \in L^2(\mathbb{R})$ supported in $[0, N]$, if we let $\mathbf{a}(g)$ denote the autocorrelation vector of $a_n(g) = \int g(x+n)\overline{g(x)}dx$, $|n| < N$, and $Sg(x) = \sum_{n=0}^{N} c_n g(2x-n)$, then

$$\mathbf{a}(Sg) = \frac{1}{2}\mathbf{W}_N \mathbf{a}(g) \tag{1.2}$$

*Received by the editors February 4, 1994; accepted for publication (in revised form) August 4, 1994.

[†]Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, PA 15260.
[‡]Houston Advanced Research Center, 4800 Research Forest Drive, The Woodlands, TX 77381.

(Proposition 3.1). This is the most basic and important relationship in the L^2-consideration. Note that if g is an L^2_c-solution of (1.1), then $Sg = g$, and it follows that $\mathbf{a}(g)$ is a 2-eigenvector of $\mathbf{W}_N$. Villemoes [V] essentially proved that (1.1) has an L^2_c-solution if and only if $\mathbf{W}_N$ has a 2-eigenvector which is positive definite. Here we will give another characterization of the existence of the L^2_c-solution based on $\mathbf{W}_N$ and two other associated matrices $\mathbf{T}_0$ and $\mathbf{T}_1$ used in [DL1], [DL2], [CH], [W], and [LW1]. We also simplify a theorem of Cohen and Daubechies [CD, Thm. 4.3] concerning the eigenvalues of $\mathbf{W}_N$ and the Riesz basis property.

Our main objective is to consider the regularity of the L^2_c-solutions. Assuming $\sum c_n = 2$, let

$$\Lambda_{\max} = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } \mathbf{W}_N^+ \text{ and } |\lambda| \neq 2\}$$

($\mathbf{W}_N^+$ is certain truncation of $\mathbf{W}_N$ to the positive coordinates) and let

$$\alpha = -\ln(\Lambda_{\max}/2)/(2\ln 2);$$

then $0 < \alpha \leq 1$. In [V], Villemoes proved that if f is an L^2_c-solution of (1.1) and if $r < \alpha$, then $\int_{-\infty}^{\infty} |\omega^r \hat{f}(\omega)|^2 d\omega < \infty$ so that f is in the Sobolev space $H^r(\mathbb{R})$ for $r < \alpha$, and the Sobolev exponent of f is α. By using the Littlewood–Paley method, Cohen and Daubechies [CD] showed that f is in the Besov space $B_2^{r,\infty}$ for all $r < \alpha$ (an equivalent definition of Besov space $B_2^{r,\infty}$ is $\sup_{h>0} \frac{1}{h^r}\|\Delta_h f\|_2 < \infty$, where $\Delta_h f = f(\cdot + h) - f(\cdot)$ [P]). They left out the critical case when the exponent $r = \alpha$. Here we obtain some sharp estimations of the regularity of f and the decaying rate of $\hat{f}$, which improve the previous results.

THEOREM 1.1. *Let f be an L^2_c-solution of* (1.1). *Let m be the highest order among those eigenvalues λ of $\mathbf{W}_N^+$ such that $|\lambda| = \Lambda_{\max}$; then*

$$\frac{1}{h^{2\alpha}|\ln h|^{m-1}} \int_{-\infty}^{\infty} |\Delta_h f|^2 = p(h) + o(h)$$

as $h \to 0^+$, where p is a nonzero bounded continuous multiplicative periodic function of period 2 (*i.e.*, $p(2h) = p(h), h > 0$). *(The order of an eigenvalue λ is the power of the factor $(x - \lambda)$ in the minimal polynomial.)*

If we define the L^2-Lipschitz exponent of $g \in L^2(\mathbb{R})$ by

$$L^2\text{-Lip}(g) = \inf\{s : 0 < \limsup_{h\to 0^+} \frac{1}{h^s}\|\Delta_h g\|_2\},$$

then it follows from Theorem 1.1 that the L^2-Lipschitz exponent of the L^2_c-solution f is α, which is also the Sobolev exponent of f for $0 < \alpha < 1$. To study higher-order regularity, the usual assumption is the l-sum rule, $l > 1$. Here we do not need such a hypothesis, we use the lth-order difference $\Delta_h^{(l)} f$ to define the L^2-Lipschitz order for $0 \leq \alpha \leq l$ ($\Lambda_{\max}$ has to be redefined). Theorem 1.1 can be extended accordingly with the exception that when α is an integer, then the logarithmic factor can be of order $m - 1$ or $m - 2$.

For the frequency domain, we have the following asymptotic result (including higher-order α).

THEOREM 1.2. *Under the above assumptions, for any $s > \alpha$*

$$\frac{1}{T^{2(s-\alpha)}(\ln T)^k} \int_{-T}^{T} |\omega^s \hat{f}(\omega)|^2 d\omega \approx q(T)$$

as $T \to \infty$, where q is a nonzero bounded continuous function with $q(2T) = q(T)$; if α is not *an integer, then $k = m - 1$; if α is an integer, then $k = m - 1$ or $m - 2$.*

Theorem 1.1 corresponds to Theorem 5.4 later in the text. The main idea of the proof is to extend the identity (1.2) to another autocorrelation vector $\mathbf{\Phi}(h) = [\Phi_0(h), \Phi_1(h), \cdots, \Phi_N(h)]$, where Φ_n is defined by

$$\Phi_n(h) = \int_{-\infty}^{\infty} \Delta_h f(x+n)\Delta_h f(x)dx,$$

and show that for any λ-eigenvector $\mathbf{u}$ of $\mathbf{W}_N^+, \lambda \neq 0, 2$, $\langle \mathbf{\Phi}(h), \mathbf{u}\rangle = h^{2\beta}p(h)$, where $\beta = -\ln(\lambda/2)/(2\ln 2)$ and p is a nonzero bounded continuous multiplicative periodic function (Lemma 5.1, Theorem 5.2). The most involved step is to show that $\langle \mathbf{\Phi}(h), \mathbf{u}\rangle \neq 0$ (Lemma 4.3), which makes use of a classical result of L. Schwartz on the *mean periodic functions* [Sch], [K], [RL]. Theorem 1.2 is contained in Theorem 5.7 and in §6, it is derived from Theorem 1.1 by using a new form of Tauberian theorem proved in [L3].

We remark that equation (1.1) actually describes a certain self-similarity of f. The self-similar measures in fractal theory are also defined by the same class of functional equation [Hu]. The genuine ideas of calculating the asymptotic properties in Theorems 1.1 and 1.2 are already contained in [L1], [LW2], [S1], [S2], and in particular in [L2].

The Daubechies four-coefficient scaling function $D_4 = f$ provides an interesting example for the above theorems (see §6 and the appendix). It follows from a direct calculation and Theorem 1.1 that $\Lambda_{\max} = \frac{1}{2}$, $\alpha = 1$, and the regularity is given by $\frac{1}{h^2|\ln h|}\int_{-\infty}^{\infty} |\Delta_h f|^2 \approx p(h)$ as $h \to 0^+$. It is also differentiable a.e. [D], [DL2], but the derivative is not in $L^2(\mathbb{R})$ in view of the asymptotic regularity behavior as $h \to 0^+$.

We organize the paper as follows. In §2, we introduce the transition matrix $\mathbf{W}_N$ as well as the two associated matrices $\mathbf{W}$ and $\mathbf{W}_N^+$. In §3, we consider some basic properties of the transition matrices in connection with the autocorrelation functions. For completeness, we simplify the existence characterization of the L_c^2-solutions proved in [LW1]. We also give a short proof of a theorem in [CD] concerning the eigenvalues of $\mathbf{W}_N$ when the solution has the Riesz basis property (Theorem 3.7). In §4 we set up the basic lemmas for the proof of Theorem 1.1, Lemma 4.3 being the most important one. Section 5 contains the proof of Theorems 1.1 and 1.2. Section 6 is concerned with the higher-order difference and the L^2-Lipschitz exponent $\alpha > 1$. At the end, we also include an appendix which contains some graphic implementations of the theorems where the functional equation (1.1) takes only four coefficients.

2. The transition matrices.

For any sequence $\{c_n\} \in \ell^1(\mathbb{Z})$, we let

$$\omega_n = \sum_{k\in\mathbb{Z}} c_k c_{k-n}, \quad n \in \mathbb{Z}.$$

Then ω_n is the convolution of the two sequences $\{c_n\}$ and $\{c_{-n}\}$; $\{\omega_n\} \in \ell^1(\mathbb{Z})$ and $\omega_{-n} = \omega_n$. We define the infinite matrix $\mathbf{W}$ by

$$\mathbf{W} = [\omega_{i-2j}] = \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \\ \cdots & \omega_1 & \omega_{-1} & \omega_{-3} & \cdots \\ \cdots & \omega_2 & \omega_0 & \omega_{-2} & \cdots \\ \cdots & \omega_3 & \omega_1 & \omega_{-1} & \cdots \\ & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

and $\mathbf{W}_N$ is the restriction of $\mathbf{W}$ on the entries $-N \le i, j \le N$. We also define

$$\mathbf{W}^+ = \begin{pmatrix} \omega_0 & \omega_{-2} & \omega_{-4} & \cdots \\ \omega_1 + \omega_{-1} & \omega_{-1} + \omega_{-3} & \omega_{-3} + \omega_{-5} & \cdots \\ \omega_2 + \omega_{-2} & \omega_0 + \omega_{-4} & \omega_{-2} + \omega_{-6} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

that is, each entry of $\mathbf{W}^+$ is given by

$$w_{ij}^+ = \begin{cases} \omega_{-2j} & \text{if } i = 0, \\ \omega_{i-2j} + \omega_{-i-2j} & \text{if } i > 0. \end{cases}$$

Geometrically, $\mathbf{W}^+$ is obtained by first deleting the left-half part of the columns of $\mathbf{W}$, then reflecting the upper half of this truncated matrix with respect to the zeroth row and adding it to the lower half. Similarly, we can truncate the matrix $\mathbf{W}_N$ to obtain $\mathbf{W}_N^+$.

When there is no confusion, we use $\mathbf{u}$ to denote the column vectors $[u_0, \dots, u_n]^t$, $[u_{-n}, \dots, u_0, \dots, u_n]^t$, and $[\dots, u_{-1}, u_0, u_1, \dots]^t$ ($[\cdot]^t$ denotes the transpose). We define $\mathbf{F} : \mathbb{C}^{N+1} \to \mathbb{C}^{2N+1}$ by

$$\mathbf{F}(\mathbf{u}) = \left[\frac{u_N}{2}, \dots, \frac{u_1}{2}, u_0, \frac{u_1}{2}, \dots, \frac{u_N}{2}\right]^t, \quad \mathbf{u} \in \mathbb{C}^{N+1},$$

and $\mathbf{G} : \mathbb{C}^{2N+1} \to \mathbb{C}^{N+1}$ by

$$\mathbf{G}(\mathbf{u}) = [u_0, u_1 + u_{-1}, \dots, u_N + u_{-N}]^t, \quad \mathbf{u} \in \mathbb{C}^{2N+1}.$$

It is clear that the adjoints of $\mathbf{F}$ and $\mathbf{G}$ are given by

$$\mathbf{F}^*(\mathbf{u}) = \left[u_0, \frac{1}{2}(u_1 + u_{-1}), \dots, \frac{1}{2}(u_N + u_{-N})\right]^t, \quad \mathbf{u} \in \mathbb{C}^{2N+1},$$

and

$$\mathbf{G}^*(\mathbf{u}) = [u_N, \dots, u_1, u_0, u_1, \dots, u_N]^t, \quad \mathbf{u} \in \mathbb{C}^{N+1}.$$

By a λ-eigenvector of a matrix $\mathbf{M}$, we mean a right eigenvector corresponding to the eigenvalue λ. The basic eigen properties of $\mathbf{W}_N$ and $\mathbf{W}_N^+$ are related as follows.

PROPOSITION 2.1. *If* $\mathbf{u} \in \mathbb{C}^{N+1}$ *is a* λ*-eigenvector of* $\mathbf{W}_N^+$ $((\mathbf{W}_N^+)^*$*, resp.), then* $\mathbf{F}(\mathbf{u})$ $(\mathbf{G}^*(\mathbf{u})$*, resp.)* $\in \mathbb{C}^{2N+1}$ *is a* λ*-eigenvector of* $\mathbf{W}_N$ $((\mathbf{W}_N)^*$*, resp.).*

Conversely, if $\mathbf{u} \in \mathbb{C}^{2N+1}$ *is a* λ*-eigenvector of* $\mathbf{W}_N$ $((\mathbf{W}_N)^*$*, resp.), then* $\mathbf{G}(\mathbf{u})$ $((\mathbf{F}^*)(\mathbf{u})$*, resp.) is either* 0 *or a* λ*-eigenvector of* $\mathbf{W}_N^+$ $((\mathbf{W}_N^+)^*$*, resp.).*

Proof. Using elementary linear algebra and the fact that $\omega_n = \omega_{-n}$ for all $n \in \mathbb{Z}$, we have

$$\mathbf{W}_N \circ \mathbf{F} = \mathbf{F} \circ \mathbf{W}_N^+ \quad \text{and} \quad \mathbf{G} \circ \mathbf{W}_N = \mathbf{W}_N^+ \circ \mathbf{G}. \tag{2.1}$$

Suppose $\mathbf{u} \in \mathbb{C}^{N+1}$ is a λ-eigenvector of $\mathbf{W}_N^+$; then $\mathbf{F}(\mathbf{u}) \ne 0$ and by (2.1),

$$\mathbf{W}_N(\mathbf{F}(\mathbf{u})) = \mathbf{F}(\mathbf{W}_N^+\mathbf{u}) = \mathbf{F}(\lambda\mathbf{u}) = \lambda\mathbf{F}(\mathbf{u}).$$

On the other hand, suppose $\mathbf{u} \in \mathbb{C}^{2N+1}$ is a nonzero λ-eigenvector of $\mathbf{W}_N$; then

$$\mathbf{W}_N^+(\mathbf{G}(\mathbf{u})) = \mathbf{G}(\mathbf{W}_N\mathbf{u}) = \mathbf{G}(\lambda\mathbf{u}) = \lambda\mathbf{G}(\mathbf{u}).$$

The statements for the adjoints follow from the dual relationship of (2.1):

$$\mathbf{F}^* \circ (\mathbf{W}_N)^* = (\mathbf{W}_N^+)^* \circ \mathbf{F}^* \quad \text{and} \quad (\mathbf{W}_N)^* \circ \mathbf{G}^* = \mathbf{G}^* \circ (\mathbf{W}_N^+)^*. \tag{2.2}$$

Remark. If $\mathbf{u}$ is a λ-eigenvector of $\mathbf{W}_N$, then the proposition implies that $\mathbf{v} = \mathbf{F}(\mathbf{G}(\mathbf{u}))$ and $\mathbf{w} = \mathbf{u} - \mathbf{v}$ are also eigenvectors of $\mathbf{W}_N$ provided that they are not zero. Note that $\mathbf{v}$ is a symmetric and $\mathbf{w}$ is antisymmetric. If all the λ-eigenvectors of $\mathbf{W}_N$ are antisymmetric, then λ is not an eigenvalue of $\mathbf{W}_N^+$.

If $c_n = 0$ for all $n \in \mathbb{Z} \setminus \{0, 1, \ldots, N\}$, then $\omega_n = 0$ for all $|n| > N$, and

$$\mathbf{W}_N = \begin{pmatrix} \omega_N & \omega_{N-2} & \cdots & \omega_{-N+2} & \omega_{-N} & 0 & \cdots & 0 & 0 \\ 0 & \omega_{N-1} & \cdots & \omega_{-N+3} & \omega_{-N+1} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \omega_1 & \omega_{-1} & \omega_{-3} & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \omega_2 & \omega_0 & \omega_{-2} & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \omega_3 & \omega_1 & \omega_{-1} & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \omega_{N-1} & \omega_{N-3} & \cdots & \omega_{-N+1} & 0 \\ 0 & 0 & \cdots & 0 & \omega_N & \omega_{N-2} & \cdots & \omega_{-N+2} & \omega_{-N} \end{pmatrix}.$$

PROPOSITION 2.2. *Suppose $\sum_n c_{2n} = \sum_n c_{2n+1} = 1$ and $c_n = 0$ for all $n \in \mathbb{Z} \setminus \{0, 1, \ldots, N\}$. Then 2 is an eigenvalue of the matrices $\mathbf{W}_N$, $\mathbf{W}_{N-1}$, and $\mathbf{W}_N^+$, the vector $[1, \ldots, 1]^t \in \mathbb{C}^{2N+1}$ (or $\mathbb{C}^{2N-1}$) is a 2-eigenvector of $\mathbf{W}_N$ ($\mathbf{W}_{N-1}$, respectively), and $[1, 2, \ldots, 2]^t \in \mathbb{C}^{N+1}$ is a 2-eigenvector for $\mathbf{W}_N^+$.*

Proof. Note that the sum of each row of $\mathbf{W}_N$ is 2. Hence 2 is an eigenvalue; the corresponding eigenvector is $[1, 1, \ldots, 1]^t$ and Proposition 2.1 implies that 2 is also an eigenvalue of $\mathbf{W}_N^+$ with eigenvector $[1, 2, \ldots, 2]^t$.

3. The autocorrelation function. Let L_c^2 denote the set of all L^2-functions with compact supports. We call the solution of (1.1) a *scaling function*. It is well known that if $f \in L^1(\mathbb{R})$, then $\operatorname{supp} f \subseteq [0, N]$. For convenience, we assume that the c_n's are real, where $c_n = 0$ for all $n \in \mathbb{Z} \setminus \{0, 1, \ldots, N\}$, so that the solution is also real. Note that $\sum c_n = 2^m$, where $m \geq 1$ is a necessity condition for the existence of an L^1-solution f; if $m > 1$, then f is the $(m-1)$th derivative of another L^1-scaling function corresponding to the coefficients $\{2^{-(m-1)} c_n\}$ [DL1]. We will assume, without loss of generality, that $\sum c_n = 2$ throughout the paper.

For $g : \mathbb{R} \to \mathbb{R}$, we define

$$(Sg)(x) = \sum_{n=0}^{N} c_n g(2x - n).$$

It is easy to show that if $\operatorname{supp} g \subseteq [0, N]$, then $\operatorname{supp}(Sg) \subseteq [0, N]$ also. For each such g, we let

$$a_n(g) = \int_{-\infty}^{\infty} g(t + n) g(t) dt, \quad n \in \mathbb{Z},$$

be the nth autocorrelation of g defined on $\mathbb{Z}$. It is clear that $a_n(g) = a_{-n}(g)$ and $a_n(g) = 0$ for all $|n| \geq N$. By slightly abusing notations, we use $\mathbf{a}(g)$ to denote the autocorrelation vectors of g:

$$\mathbf{a}(g) = [\ldots, a_{-1}(g), a_0(g), a_1(g), \ldots]^t \quad \text{or} \quad [a_{-k}(g), \ldots, a_0(g), \ldots, a_k(g)]^t,$$

depending on the situation. Let $\mathbf{e}_n$ be the vector (finitely or infinitely many entries) with 1 on the nth entry and 0 otherwise. The major property of the transition matrix $\mathbf{W}$ defined in §2 is given in the following proposition.

PROPOSITION 3.1. *Let $g \in L^2(\mathbb{R})$ be supported in $[0, N]$. Then*

$$\mathbf{a}(Sg) = \frac{1}{2}\mathbf{W}^*\mathbf{a}(g), \tag{3.1}$$

where $\mathbf{W}^*$ is the adjoint of $\mathbf{W}$. In particular,

$$\int_{-\infty}^{\infty} |(Sg)(t)|^2 dt = \frac{1}{2}\langle \mathbf{a}(g), \mathbf{W}\mathbf{e}_0\rangle = \frac{1}{2}\langle \mathbf{a}(g), \mathbf{W}_{N-1}\,\mathbf{e}_0\rangle.$$

Proof. The proof is based on the following observation: for $n \in \mathbb{Z}$,

$$\begin{aligned}
a_n(Sg) &= \int_{-\infty}^{\infty} (Sg)(t+n)(Sg)(t)\,dt \\
&= \sum_{i,j\in\mathbb{Z}} c_j c_i \int_{-\infty}^{\infty} g(2t+2n-j)g(2t-i)\,dt \\
&= \frac{1}{2}\sum_{i,j\in\mathbb{Z}} c_j c_i \int_{-\infty}^{\infty} g(t+2n+i-j)g(t)\,dt \\
&= \frac{1}{2}\sum_{k\in\mathbb{Z}}\left(\sum_{i\in\mathbb{Z}} c_i c_{i-(k-2n)}\right) a_k(g) \\
&= \frac{1}{2}\sum_{k\in\mathbb{Z}} \omega_{k-2n} a_k(g) \\
&= \frac{1}{2}[\mathbf{W}^*\,\mathbf{a}(g)]_n
\end{aligned}$$

(the second equality holds since we assume that $c_n = 0$ for all $n \in \mathbb{Z}\setminus\{0, 1, \ldots, N\}$). The last identity in the proposition holds due to the fact that $a_n(g) = 0$ for all $|n| \geq N$.

Remark. For $y \in \mathbb{R}$, if we let $a_n^{(y)}(g) = \int_{-\infty}^{\infty} g(t+n-y)g(t)dt$, then the same calculation yields

$$\mathbf{a}^{(y)}(Sg) = \frac{1}{2}\mathbf{W}^*\mathbf{a}^{(2y)}(g). \tag{3.1$'$}$$

We will use this fact in Theorem 3.7.

Recall that a sequence $\{u_n\}_{n=-\infty}^{\infty}$ is called *positive definite* if for any finite sequence $\{\xi_n\}$, $\sum u_{m-n}\xi_m\bar{\xi}_n \geq 0$. It is well known that the autocorrelation sequence $\{a_n(g)\}$ (letting $a_n(g) = 0$ for all $|n| \geq N$) is positive definite.

PROPOSITION 3.2. *Suppose f is a nonzero L_c^2-solution of* (1.1); *then* $\mathbf{a}(f)$ *is a 2-eigenvector of* $(\mathbf{W}_{N-1})^*$, $\sum a_n(f) \neq 0$, *and* $\{a_n(f)\}_{n=-\infty}^{\infty}$ *is a positive-definite sequence.*

Proof. In view of Proposition 3.1 and the remark above, we need only show that $\sum a_n(f) \neq 0$. This follows from the well-known Poisson formula $\sum a_n(f)e^{in\omega} = \sum |\hat{f}(\omega + 2\pi n)|^2$ and the sum is strictly positive for $\omega = 0$.

The existence of a vector satisfying the above conditions also implies the existence of an L^2_c-solution of (1.1), which has been observed by Villemoes in [V] (where he uses $\sum a_n(f)e^{inx} \geq 0$ instead of using the fact that $\{a_n(f)\}_{n=-\infty}^{\infty}$ is positive definite).

In order to construct the L^2_c-solution f of the dilation equation (1.1), we can formally proceed as follows: take a function g with supp$g \subseteq [0, N]$ and consider $\{S^k(g)\}_{k=1}^{\infty}$. If this sequence converges in L^2 to a function f, then f will be a solution to (1.1). Equivalently, we can write

$$S^k(g) = g + \sum_{l=0}^{k-1} S^l(Sg - g) \tag{3.2}$$

and consider the convergence of the series $\sum_{l=0}^{k-1} S^l(Sg - g)$. Let

$$\mathbf{T}_0 = [c_{2i-j-1}]_{1\leq i,j\leq N} = \begin{pmatrix} c_0 & 0 & 0 & \dots & 0 \\ c_2 & c_1 & c_0 & \dots & 0 \\ c_4 & c_3 & c_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \dots & c_{N-1} \end{pmatrix},$$

$$\mathbf{T}_1 = [c_{2i-j}]_{1\leq i,j\leq N} = \begin{pmatrix} c_1 & c_0 & 0 & \dots & 0 \\ c_3 & c_2 & c_1 & \dots & 0 \\ c_5 & c_4 & c_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \dots & c_N \end{pmatrix}.$$

These matrices were used in [DL1], [DL2], [CH], and [W] to study the continuous scaling solutions. In [LW1, Thm. 4.3], the authors also use such matrices to give a necessary and sufficient condition for the existence of the L^p_c-solutions; for the L^2-case, the criterion is reduced to consider the eigenvalues of $\mathbf{W}$. The proof is simplified later in Theorem 3.4. First, we state a very useful lemma concerning $\mathbf{T}_0 + \mathbf{T}_1$ which is proven in [LW1].

LEMMA 3.3. *Suppose* $\sum_{n=0}^{N} c_n = 2$. *Then the following hold:*

(i) 2 *is an eigenvalue of* $\mathbf{T}_0 + \mathbf{T}_1$.

(ii) *If* $\mathbf{v}$ *is a* 2*-eigenvector of* $\mathbf{T}_0 + \mathbf{T}_1$, *let* $g = \sum_{n=0}^{N-1} v_n \chi_{[n-1,n)}$; *then*

$$\left[\int_0^1 S^k g, \dots, \int_{N-1}^{N} S^k g\right]^t = \mathbf{v}.$$

(iii) *For* $1 \leq p < \infty$, *let* $f \in L^p_c(\mathbb{R})$ *be the* L^p_c*-solution of* (1.1) *and let* $\mathbf{v} = [\int_0^1 f, \dots, \int_{N-1}^{N} f]^t$; *then* $\mathbf{v}$ *is a* 2*-eigenvector of* $\mathbf{T}_0 + \mathbf{T}_1$. *For such* $\mathbf{v}$, *if we let* g *be defined as in* (ii), *then* $\{S^k(g)\}$ *converges back to* f *in the* L^p*-norm.*

THEOREM 3.4. *Suppose* $\sum_{n=0}^{N} c_n = 2$. *Let* $\mathbf{v}$ *be a* 2*-eigenvector of* $\mathbf{T}_0 + \mathbf{T}_1$ *and let* $g = \sum_{n=0}^{N-1} v_n \chi_{[n,n+1)}$. *Let* $H_{\mathbf{v}}$ *be the smallest invariant subspace of* $\mathbf{W}_{N-1}$ *containing* *the autocorrelation vector* $\mathbf{a}(Sg-g)$. *Then* (1.1) *has a nonzero* L_c^2*-solution if and only if all the eigenvalues of* $\mathbf{W}_{N-1}$ *restriced to* $H_{\mathbf{v}}$ *have modulus less than* 2.

Proof. Let $\tilde{g} = Sg - g$. Note that by Proposition 3.1, we have

$$\begin{aligned}\|S^l\tilde{g}\|^2 &= \int_0^N |S^l\tilde{g}(t)|^2\,dt \\ &= \frac{1}{2^l}\langle \mathbf{a}(\tilde{g}),\ \mathbf{W}_{N-1}^l\mathbf{e}_0\rangle \\ &= \langle \frac{1}{2^l}(\mathbf{W}_{N-1}^l)^*\,\mathbf{a}(\tilde{g}),\ \mathbf{e}_0\rangle.\end{aligned}$$

The assumption that $\frac{1}{2}(\mathbf{W}_{N-1})^*$ restricted on $H_{\mathbf{v}}$ has spectral radius less than 1 implies that $\{\frac{1}{2^l}(\mathbf{W}_{N-1}^l)^*\mathbf{a}(\tilde{g})\}$ converges to zero geometrically as $l \to \infty$; so does $\{\|S^l\tilde{g}\|^2\}$. Consequently, $S^k(g) = g + \sum_{l=0}^{k-1} S^l\tilde{g}$ converges in L^2. Let f be the limit. Then $f \in L_c^2(\mathbb{R})$; $f \neq 0$ because by Lemma 3.3(ii),

$$\left[\int_0^1 S^l\tilde{g}, \ldots, \int_{N-1}^N S^l\tilde{g}\right] = 0$$

so that

$$\left[\int_0^1 f, \ldots, \int_{N-1}^N f\right] = \left[\int_0^1 g, \ldots, \int_{N-1}^N g\right] = \mathbf{v} \neq \mathbf{0}.$$

To prove the converse, we observe that (3.2) and Proposition 3.1 imply that

$$\frac{1}{2^l}(\mathbf{W}^l)^*\mathbf{a}(Sg-g) = \mathbf{a}(S^l(Sg-g)) \longrightarrow 0 \quad \text{as} \quad l \longrightarrow \infty.$$

It follows that all the eigenvalues of $\mathbf{W}_{N-1}$ restricted to $H_{\mathbf{v}}$ have modulus less than 2.

For the special case where $\sum_n c_{2n} = \sum_n c_{2n+1} = 1$, Theorem 3.4 yields a simple criterion for the existence of the L_c^2-solution (see also [CD, Thm. 3.3]). We need to make use of the following simple facts.

LEMMA 3.5. *Suppose* $\sum_n c_{2n} = \sum_n c_{2n+1} = 1$.

(i) *Let* $\mathbf{v} = [v_0, v_1, \ldots, v_{N-1}]^t$ *and let* $g = \sum_{n=0}^{N-1} v_n\chi_{[n,n+1)}$. *Then for any* $k \in \mathbb{N}$ *and for almost all* $x \in [0,1)$,

$$\sum_{n=0}^{N-1} S^k g(x+n) = \sum_{n=0}^{N-1} v_n.$$

(ii) *Let* $H = \{\mathbf{u} \in \mathbb{C}^{2N-1} : \sum_{n=-(N-1)}^{N-1} u_n = 0\}$. *Then* $(\mathbf{W}_{N-1})^*$ *is invariant on* H.

Proof. The proofs of (i) and (ii) are quite similar. For (i), we make use of the fact that $[1, \ldots, 1]$ is a left 1-eigenvector of $\mathbf{T}_0$ and $\mathbf{T}_1$ (see,e.g., [H]). To prove (ii), note that $[1, 1, \ldots, 1]^t$ is a 2-eigenvector of $\mathbf{W}_{N-1}$ (Proposition 2.2); hence for $\mathbf{u} \in H$,

$$[1,1,\ldots,1]\,(\mathbf{W}_{N-1})^*\,\mathbf{u} = 2\,[1,1,\ldots,1]\,\mathbf{u} = 2\sum_{n=-(N-1)}^{N-1} u_n = 0.$$

This implies that the sum of the coordinates of $(\mathbf{W}_{N-1})^* \mathbf{u}$ is zero so that $(\mathbf{W}_{N-1})^*$ is invariant on H.

COROLLARY 3.6. *Suppose $\sum c_{2n} = \sum c_{2n+1} = 1$. If the eigenvalues of $(\mathbf{W}_{N-1})^*$ restricted on H have moduli less than 2, then* (1.1) *has an L^2_c-solution.*

Proof. Let $\mathbf{v} = [v_0, v_1, \ldots, v_{n-1}]^t$ be a 2-eigenvector of $\mathbf{T}_0 + \mathbf{T}_1$ as in Theorem 3.4. Then Lemma 3.5 implies that $\sum_{n=0}^{N-1} \tilde{g}(x+n) = 0$ for almost all $x \in [0,1)$. Since supp$g \subseteq [0,N]$, we actually have $\sum_{n=-\infty}^{\infty} \tilde{g}(x+n) = 0$ for almost all $x \in [0,1)$ and hence for almost all $x \in \mathbb{R}$. Therefore,

$$\begin{aligned}\sum_{|n|\le N-1} \mathbf{a}_n(\tilde{g}) &= \sum_{n=-\infty}^{\infty} \mathbf{a}_n(\tilde{g}) \\ &= \sum_{n=-\infty}^{\infty} \int_0^N \tilde{g}(t+n)\tilde{g}(t)\,dt \\ &= \int_0^N \left(\sum_{n=-\infty}^{\infty} \tilde{g}(t+n) \right) \tilde{g}(t)\,dt \\ &= 0,\end{aligned}$$

and $\mathbf{a}(\tilde{g}) \in H$. This implies that the subspace $H_{\mathbf{v}}$ in Theorem 3.4 is contained in H. By assumption, $\frac{1}{2}(\mathbf{W}_{N-1})^*$ restricted on H has spectral radius less than 1, and Theorem 3.4 applies.

In [LW1, Prop. 4.6], it is proven that the converse of the above corollary is also true if we assume that 2 is a simple eigenvalue of $\mathbf{W}_{N-1}$ and $\{\mathbf{W}^l_{N-1}\mathbf{e}_1\}$ generates $\mathbb{C}^{2N-1}$; for the four-coefficient case (N=3), the above additional assumptions are always true except for the case $c_0 = c_3 = 1$. By using a long and rather complicated argument Cohen and Daubechies [CD, Thm. 4.3] also showed that the converse is true if f has the Riesz-basis property. In the following, we will give a short proof of their theorem.

Recall that a function $f \in L^2(\mathbb{R})$ is said to satisfy the *Riesz-basis property* if the sequence of functions $f_n = f(\cdot - n), n \in \mathbb{Z}$ forms a Riesz-basis for the closure of its linear span in $L^2(\mathbb{R})$, i.e., there exist $C_1, C_2 > 0$ such that

$$C_1 \sum |\alpha_n|^2 \le \|\sum \alpha_n f_n\|^2 \le C_2 \sum |\alpha_n|^2.$$

Cohen [C], Lawton [La], and Villemoes [V] have given different criteria for such a property in terms of the Fourier transformation. In particular, Villemoes showed that if an L^2_c-solution f has the Riesz-basis property, then $\sum c_{2n} = \sum c_{2n+1} = 1$. Also, assuming such a summing condition, f has the Riesz-basis property if and only if $\sum a_n(f)e^{in\omega}$ is strictly positive.

THEOREM 3.7. *Suppose f is a solution of* (1.1) *and has the Riesz-basis property. Then $(\mathbf{W}_{N-1})^*$ restricted on H has spectral radius less than* 2.

Proof. Since f has the Riesz-basis property, then $\sum c_{2n} = \sum c_{2n+1} = 1$ [V] so that $(\mathbf{W}_{N-1})^*$ is invariant on H. All eigenvalues of $(\mathbf{W}_{N-1})^*$ have moduli less than or equal to 2 (see Propositon 5.3 in §5). The proof will be complete if we show that $(\mathbf{W}_{N-1})^*$ does not have another 2-eigenvector other than $\mathbf{a}(f)$, which is not in H.

Note that $\sum a_n(f)e^{in\omega} = \sum |\hat{f}(\omega + 2\pi k)|^2 > 0$ by the Riesz-basis property. Suppose $\mathbf{u}$ is another 2-eigenvector of $(\mathbf{W}_{N-1})^*$. By letting $u_n = 0$ for all $|n| \ge N$, $\mathbf{u}$ is a 2-eigenvector of $\mathbf{W}^*$. By Wiener's theorem, there exists $\{r_n\}_{n=-\infty}^{\infty} \in \ell^1$ such that

$$\sum r_n e^{in\omega} = \frac{\sum u_n e^{in\omega}}{\sum a_n(f)e^{in\omega}}.$$

It follows that $\mathbf{u} = \mathbf{r} * (\mathbf{a}(f))$ and

$$\begin{aligned} u_n &= \lim_{l\to\infty} \frac{1}{2^l}\langle(\mathbf{W}^*)^l\mathbf{u},\, \mathbf{e}_n\rangle \\ &= \lim_{l\to\infty} \frac{1}{2^l}\sum_k \langle r_k(\mathbf{W}^*)^l\mathbf{a}^{(k)}(f),\, \mathbf{e}_n\rangle \\ &= \lim_{l\to\infty} \sum_k \langle r_k\mathbf{a}^{(2^{-l}k)}(f),\, \mathbf{e}_n\rangle & \text{(use (3.1)')} \\ &= \sum_k r_k \,\langle \lim_{l\to\infty} \mathbf{a}^{(2^{-l}k)}(f),\, \mathbf{e}_n\rangle \\ &= C\,\langle\mathbf{a}(f),\, \mathbf{e}_n\rangle, \end{aligned}$$

where $C = \sum_k r_k$. This implies that $\mathbf{u}$ is a scalar multiple of $\mathbf{a}(f)$ and the proof is complete.

Remark. The above discussion gives a simple criterion for computer to check for the Riesz-basis property of the solution f given $\{c_n\}_{n=0}^{N}$ with $\sum c_{2n} = \sum c_{2n+1} = 1$: first show that the 2-eigenvalue of $(\mathbf{W}_{N-1})^*$ is simple and all other eigenvalues are less than 2 in modulus (this implies the existence of the solution by Corollary 3.6 and $\sum a_n(f)e^{in\omega} \geq 0$ by Proposition 3.2), and then show that the polynomial $\frac{1}{2}\sum a_n(f)z^n$ has no root on the unit circle.

4. Some lemmas. In the rest of the paper, we will use the difference quotient $\frac{1}{h^{2\beta}}\int_{-\infty}^{\infty}|\Delta_h f(t)|^2dt$ to study the regularity properties of the scaling function f. We prefer to use the matrix $\mathbf{W}_N^+$ rather than $\mathbf{W}_N$ because using the latter, we have to discard those eigenvalues which only give antisymmetric eigenvectors (see the remark for Proposition 2.1 and Lemma 4.3). As before, we assume that $\sum c_n = 2$. For $h \in \mathbb{R}$ and $n \in \mathbb{Z}$, we also define

$$\Phi_n(h) = \int_{-\infty}^{\infty} \Delta_h f(t+n)\Delta_h f(t)dt.$$

Since f is supported by $[0, N]$,

$$\Phi_n(h) = 0 \quad \forall\, 0 < h < 1,\ |n| \geq N+1. \tag{4.1}$$

We use $\mathbf{\Phi}(h)$ to denote

$$[\Phi_0(h), \Phi_1(h), \ldots, \Phi_N(h)]^t \quad\text{and}\quad [\Phi_0(h), \Phi_1(h), \ldots]^t.$$

If necessary, we will add the superscript N or ∞ to $\mathbf{\Phi}(h)$ to make the distinction. It is clear that $\Delta_h f$ satisfies

$$\Delta_h f(x) = \sum_{n=0}^{N} c_n \Delta_{2h} f(2x-n). \tag{4.2}$$

PROPOSITION 4.1. *Let* $\mathbf{W}$ *be the transition matrix corresponding to the scaling function* f *satisfying* (1.1). *Then for* $\mathbf{u} = [u_0, u_1, \ldots, u_N]^t \in \mathbb{C}^{N+1}$,

$$\langle \mathbf{\Phi}(h), \mathbf{u}\rangle = \frac{1}{2}\langle \mathbf{\Phi}(2h), \mathbf{W}_N^+ \mathbf{u}\rangle, \quad 0 < h < \frac{1}{2}. \tag{4.3}$$

Proof. The proof is basically the same as that of Proposition 3.1, using (4.2) instead of (1.1). For $0 \le n \le N$,

$$\begin{aligned}\Phi_n(h) &= \int_{-\infty}^{\infty} \Delta_h f(t+n)\Delta_h f(t)\, dt \\ &= \sum_{i,j\in\mathbb{Z}} c_j c_i \int_{-\infty}^{\infty} \Delta_{2h} f(2t+2n-j)\Delta_{2h} f(2t-i)\, dt \\ &\ \vdots \\ &= \frac{1}{2}[\mathbf{W}^* \mathbf{G}^*(\mathbf{\Phi}(2h))]_n \\ &= \frac{1}{2}[(\mathbf{W}_N^+)^*(\mathbf{\Phi}(2h))]_n.\end{aligned}$$

LEMMA 4.2. *If* $\mathbf{u}$ *is a* 2*-eigenvector or a* 0*-eigenvector of* $\mathbf{W}_N^+$, *then*

$$\langle \mathbf{\Phi}(h), \mathbf{u}\rangle = 0 \qquad \forall\ 0 < h < 1.$$

Proof. Let $\mathbf{u}$ be a 2-eigenvector of $\mathbf{W}_N^+$. We have, by (4.3), $\langle \mathbf{\Phi}(\frac{h}{2}), \mathbf{u}\rangle = \langle \mathbf{\Phi}(h), \mathbf{u}\rangle$. for all $0 < h < 1$. Hence, inductively,

$$\langle \mathbf{\Phi}(h), \mathbf{u}\rangle = \left\langle \mathbf{\Phi}\left(\frac{h}{2^2}\right), \mathbf{u}\right\rangle = \cdots = \left\langle \mathbf{\Phi}\left(\frac{h}{2^m}\right), \mathbf{u}\right\rangle$$

for all $0 < h < 1$, $m \ge 0$. But $\langle \mathbf{\Phi}(\frac{h}{2^m}), \mathbf{u}\rangle \to 0$ as $m \to \infty$, so it follows that $\langle \mathbf{\Phi}(h), \mathbf{u}\rangle = 0$ for all $0 < h < 1$. The same conclusion also holds if $\mathbf{u}$ is a 0-eigenvector since in such a case $\mathbf{W}_N^+ \mathbf{u} = 0$.

Our main lemma is the following.

LEMMA 4.3. *If* $\mathbf{u}$ *is a* λ*-eigenvector of* $\mathbf{W}_N^+$ *with* $\lambda \ne 0$ *or* 2, *then*

$$\langle \mathbf{\Phi}(h), \mathbf{u}\rangle \ne 0 \qquad \textit{for some } 0 < h < \frac{1}{2}.$$

The proof of this lemma is rather long. The basic idea is to prove by contradiction. Suppose otherwise, i.e., $\psi(h) = \langle \mathbf{\Phi}(h), \mathbf{u}\rangle = 0$ for all $0 < h < \frac{1}{2}$. We show that when $\mathbf{u}$ is replaced by the corresponding λ-eigenvector $\tilde{\mathbf{u}}$ for $\mathbf{W}^+$ (Lemma 4.4) and the inner product in $\psi(h)$ is acting on all positive coordinates, then the identity holds for all $h \in \mathbb{R}$. From this we deduce that the corresponding sequence $\{u_n\}$ is a linear combination of certain exponential sequences of the form $\{e^{ian}\}$. However, the eigenproperty of $\{u_n\}$ implies that this is impossible.

For this purpose, we first observe that for $h \in \mathbb{R}$,

$$\begin{aligned}\Phi_n(h) &= \int_{-\infty}^{\infty} \Delta_h f(t+n)\Delta_h f(t)\, dt \\ &= \int_{-\infty}^{\infty} (f(t+n+h) - f(t+n))(f(t+h) - f(t))\, dt.\end{aligned}$$

Multiplying out the integrand and changing the variables, we obtain

$$\Phi_n(h) = 2a_n(f) - \Psi_n(h), \tag{4.4}$$

where

$$\Psi_n(h) = \int_{-\infty}^{\infty} [f(t+h-n) + f(t+h+n)]f(t)\,dt$$

and $a_n(f)$ is the nth autocorrelation. We let

$$\mathbf{\Psi}^N(h) = [\Psi_0(h), \Psi_1(h), \ldots, \Psi_N(h)]^t \quad \text{and} \quad \mathbf{\Psi}^\infty(h) = [\Psi_0(h), \Psi_1(h), \ldots]^t.$$

Note that $[a_0(f), a_1(f), \ldots, a_N(f)]^t$ is a 2-eigenvector of $(\mathbf{W}_N^+)^*$ (use Proposition 3.2 and (2.2)). It is orthogonal to any λ-eigenvector $\mathbf{u}$ of $\mathbf{W}_N^+$ with $\lambda \neq 2$. For such $\mathbf{u}$, (4.4) implies that

$$\langle \mathbf{\Phi}^N(h), \mathbf{u}\rangle = -\langle \mathbf{\Psi}^N(h), \mathbf{u}\rangle, \quad h \in \mathbb{R}. \tag{4.5}$$

Let $\mathcal{S}$ be the class of all one-sided infinite sequences.

LEMMA 4.4. *If* $\mathbf{u}$ *is a* λ*-eigenvector of* $\mathbf{W}_N^+$ *with* $\lambda \neq 0$, *then there exists* $\tilde{\mathbf{u}} \in \mathcal{S}$, *a* λ*-eigenvector of* $\mathbf{W}^+$ *such that* $\tilde{u}_n = u_n$ *for all* $0 \le n \le N$. *Furthermore, for such a* $\tilde{\mathbf{u}}$,

$$\langle \mathbf{\Psi}^\infty(h), \tilde{\mathbf{u}}\rangle = \langle \mathbf{\Psi}^N(h), \mathbf{u}\rangle, \quad 0 < h < \frac{1}{2}.$$

Proof. Let $\mathbf{u} = [u_0, u_1, \ldots, u_N]^t$ be a λ-eigenvector of $\mathbf{W}_N^+$ with $\lambda \neq 0$. Note that for $i > 0$,

$$w_{ij}^+ = \omega_{i-2j} + \omega_{-i-2j}.$$

Since $\omega_n = 0$ for all $|n| > N$ and $N < i \le j$, $-i-2j < i-2j < -N$, we hence have $w_{ij}^+ = 0$ for all $N < i \le j$. We now construct $\tilde{\mathbf{u}}$ as follows: let $\tilde{u}_n = u_n$ for all $0 \le n \le N$ and define $\tilde{u}_{n+1}$ inductively as

$$\tilde{u}_{n+1} = \frac{1}{\lambda}\sum_{k=0}^{n} w_{n+1,k}^+ \tilde{u}_k, \quad n \ge N. \tag{4.6}$$

Then $\tilde{\mathbf{u}}$ is the required vector. The last assertion follows from the fact that for $0 < h < \frac{1}{2}$, $\Psi_n(h) = 0$ for all $n > N$.

In [Sch] (see also [K], [RL]), L. Schwartz proved the following classical result on *mean periodic functions*: Let μ be a bounded regular Borel measure on $\mathbb{R}$ with compact support. Let $\mathcal{C}$ be the class of continuous functions on $\mathbb{R}$ equipped with the compact open topology. Suppose there exists a nonzero $g \in \mathcal{C}$ that satisfies the convolution equation

$$\int_{-\infty}^{\infty} g(x-y)d\mu(y) = 0 \qquad \forall\, x \in \mathbb{R}.$$

Then g belongs to the closed linear subspace spanned by

$$\{e^{ia(\cdot)} : a \in \mathbb{C}, \int_{-\infty}^{\infty} e^{-iay}d\mu(y) = 0\}.$$

Heuristically, the convolution equation implies that $\hat{g}(z)\hat{\mu}(z) = 0$ (in the distribution sense). Since μ has compact support, $\hat{\mu}$ is an entire function and has only countably many discrete zeros. It follows that the support of $\hat{g}$ must be contained in the zeros of $\hat{\mu}$, and g is of the form asserted. We need the discrete version, which is an easy corollary of the above theorem: if $\{w_n\}_{n=-\infty}^{\infty}$ is a given sequence with only finitely many nonzero terms and if $\{x_n\}_{n=-\infty}^{\infty}$ is any sequence satisfying

$$\sum_{k=-\infty}^{\infty} x_{n-k}w_k = 0 \quad \forall\, n \in \mathbb{Z},$$

then $\{x_n\}_{n=-\infty}^{\infty}$ belongs to the closed (with respect to the product topology) linear subspace spanned by

$$\{\{e^{ian}\} : a \in \mathbb{C},\ \sum_{n=-\infty}^{\infty} e^{-ian}w_n = 0\}.$$

LEMMA 4.5. *Let* $\tilde{\mathbf{u}}$ *be a* λ*-eigenvector of* $\mathbf{W}^+$ *with* $\lambda \neq 0$ *or* 2. *Then*

$$\langle \mathbf{\Psi}^\infty(h), \tilde{\mathbf{u}}\rangle \neq 0 \qquad \textit{for some } 0 < h < \frac{1}{2}.$$

Proof. By Proposition 4.1, Lemma 4.4, and (4.5), we have for any $\mathbf{v} \in \mathcal{S}$,

$$\langle \mathbf{\Psi}^\infty(h), \mathbf{v}\rangle = \frac{1}{2}\langle \mathbf{\Psi}^\infty(2h), \mathbf{W}^+\mathbf{v}\rangle \quad \forall\, h \in \mathbb{R}. \tag{4.7}$$

For any fixed h, $\Psi_n(h) = 0$ for all large n; hence $\langle \mathbf{\Psi}^\infty(h), \tilde{\mathbf{u}}\rangle$ is well defined and is continuous on h. Suppose the lemma is false, i.e.,

$$\langle \mathbf{\Psi}^\infty(h), \tilde{\mathbf{u}}\rangle = 0 \qquad \forall\, 0 < h < \frac{1}{2}. \tag{4.8}$$

By (4.7), we have

$$0 = \langle \mathbf{\Psi}^\infty(h), \tilde{\mathbf{u}}\rangle = \frac{1}{2}\langle \mathbf{\Psi}^\infty(2h), \mathbf{W}^+\tilde{\mathbf{u}}\rangle = \frac{\lambda}{2}\langle \mathbf{\Psi}^\infty(2h), \tilde{\mathbf{u}}\rangle.$$

The assumption that $\lambda \neq 0$ implies that $\langle \mathbf{\Psi}^\infty(2h), \tilde{\mathbf{u}}\rangle = 0$ for all $0 < h < \frac{1}{2}$, i.e., (4.8) holds for all $0 < h < 1$. Repeating the same argument, we have that (4.8) holds for all $h > 0$ and hence for all $h \in \mathbb{R} \setminus \{0\}$ since $\mathbf{\Psi}^\infty(-h) = \mathbf{\Psi}^\infty(h)$. By continuity, we also have $\langle \mathbf{\Psi}^\infty(0), \tilde{\mathbf{u}}\rangle = 0$. We hence conclude that

$$\sum_{k=0}^{\infty} \tilde{u}_k \left(f * \tilde{f}(h-k) + f * \tilde{f}(h+k)\right) = 0 \qquad \forall\, h \in \mathbb{R},$$

where $\tilde{f}(x) = f(-x)$. By letting $x_0 = 2u_0$, $x_n = x_{-n} = \tilde{u}_n$ for $n > 0$, and by replacing h with $h + n$, $0 \leq h < 1$, we can rewrite the above as

$$\sum_{k=-\infty}^{\infty} x_{n-k} f * \tilde{f}(h+k) = 0 \qquad \forall\, n \in \mathbb{Z},\ h \in [0, 1).$$

Note that the autocorrelation function $f * \tilde{f}$ is continuous and has compact support. For each fixed $h \in [0,1)$, if we regard the sequence $\{f * \tilde{f}(h+n)\}$ as the $\{w_n\}$ in the above digression, then $\{x_n\}_{n=-\infty}^{\infty}$ must be in the closed linear subspace spanned by

$$A_h = \{\{e^{ian}\} : a \in \mathbb{C}, \sum_{n=-\infty}^{\infty} f * \tilde{f}(h+n)e^{-ian} = 0\}.$$

Since this is true for all $h \in [0,1)$, $\{x_n\}_{n=-\infty}^{\infty}$ must be in the closed linear subspace spanned by $\bigcap_{h\in[0,1)} A_h$. By using the Poisson summation formula ([Ch, p. 47]), we have

$$\begin{aligned} 0 &= \sum_{n=-\infty}^{\infty} f * \tilde{f}(h+n)e^{-ina} \\ &= \sum_{n=-\infty}^{\infty} \hat{f}(a+2\pi n)\hat{f}(-(a+2\pi n))e^{ih(a+2\pi n)} \\ &= e^{iha} \sum_{n=-\infty}^{\infty} \hat{f}(a+2\pi n)\hat{f}(-(a+2\pi n))e^{i2\pi nh} \end{aligned}$$

for all $h \in [0,1)$. This implies that

$$\hat{f}(a+2\pi n)\hat{f}(-(a+2\pi n)) = 0 \quad \forall n \in \mathbb{Z}.$$

Observe that the Fourier transformation of (1.1) is $\hat{f}(z) = \hat{f}(\frac{z}{2})m_0(\frac{z}{2})$, where $m_0(z) = \frac{1}{2}\sum c_n e^{inz}$ is a trigonometric polynomial of degree N. Let $F(z) = \hat{f}(z)\hat{f}(-z)$, $Q(e^{iz}) = m_0(z)m_0(-z)$. Since $F \neq 0$ in a neighborhood of 0, we conclude from $0 = F(a) = F(\frac{a}{2})Q(e^{ia/2})$ that for some l, $e^{ia/2^l}$ must be a root of Q. Hence the sequence $\{x_n\}_{n=-\infty}^{\infty}$ is in the close linear span of all the sequences of the form

$$\{\{e^{ian}\} : e^{ia/2^l} \text{ is a root of } Q(z) \text{ for some } l\}. \tag{4.9}$$

Now, by a direct calculation,

$$[\mathbf{W}(e^{ia(\cdot)})]_n = \sum_{k=-\infty}^{\infty} w_{n-2k}e^{iak} = 2^2 e^{ian/2}Q(e^{ia/2}).$$

This implies that for some l,

$$\mathbf{W}^l(e^{ia(\cdot)}) = 2^{2l}e^{ia(\cdot)/2^l}Q(e^{ia/2^l}) = 0. \tag{4.10}$$

On the other hand, in view of Lemma 4.4, the vector $\mathbf{x} = [\ldots, x_{-1}, x_0, x_1, \ldots]$ satisfies $\mathbf{W}\mathbf{x} = \lambda\mathbf{x}$. This is a contradiction since $\{x_n\}_{n=-\infty}^{\infty}$ is a combination of the sequences $\{e^{ian}\}_{n=-\infty}^{\infty}$ in (4.9), and (4.10) implies that $\mathbf{x}$ can not be an eigenvector.

Proof of Lemma 4.3. Suppose that $\mathbf{u}$ is an λ-eigenvector of $\mathbf{W}_N^+$ with $\lambda \neq 0$ or 2. By Lemma 4.4, there exists $\tilde{\mathbf{u}} \in \mathcal{S}$ such that $\mathbf{W}^+\tilde{\mathbf{u}} = \lambda\tilde{\mathbf{u}}$ and $\tilde{u}_n = u_n$ for all $0 \le n \le N$. By Lemma 4.5, we have $\langle \mathbf{\Psi}^\infty(h), \tilde{\mathbf{u}}\rangle \neq 0$ for some $0 < h < \frac{1}{2}$. For such h,

$$\begin{aligned} \langle \mathbf{\Phi}^N(h), \mathbf{u}\rangle &= -\langle \mathbf{\Psi}^N(h), \mathbf{u}\rangle && \text{(by (4.5))} \\ &= -\langle \mathbf{\Psi}^\infty(h), \tilde{\mathbf{u}}\rangle \\ &\neq 0. \end{aligned}$$

5. The L^2-Lipschitz exponent and asymptotics. For any $\beta \in \mathbb{C}$, we let

$$\mathbf{\Phi}^{(\beta)}(h) = \frac{1}{h^{2\beta}}\mathbf{\Phi}(h).$$

LEMMA 5.1. *Let $\lambda \neq 0, 2$ be an eigenvalue of $\mathbf{W}_N^+$, let $\beta = -\ln(\lambda/2)/(2\ln 2)$ (i.e., $\lambda/2^{1-2\beta} = 1$ and β takes the principal branch when λ is complex), and let $\hat{\mathbf{u}}$ be the corresponding eigenvector. Then*

$$\langle \mathbf{\Phi}^{(\beta)}(2h), \hat{\mathbf{u}}\rangle = \langle \mathbf{\Phi}^{(\beta)}(h), \hat{\mathbf{u}}\rangle \quad \forall\, 0 < h < \frac{1}{2}.$$

Proof. Let $\phi(h) = \langle \mathbf{\Phi}^{(\beta)}(h), \hat{\mathbf{u}}\rangle$. By Proposition 4.1, we have for $0 < h < \frac{1}{2}$,

$$\begin{aligned}\langle \mathbf{\Phi}^{(\beta)}(h), \hat{\mathbf{u}}\rangle &= \frac{1}{h^{2\beta}}\langle \mathbf{\Phi}(h), \hat{\mathbf{u}}\rangle \\ &= \frac{1}{2^{1-2\beta}} \cdot \frac{1}{(2h)^{2\beta}}\langle \mathbf{\Phi}(2h), \mathbf{W}_N^+\hat{\mathbf{u}}\rangle \\ &= \frac{\lambda}{2^{1-2\beta}}\langle \mathbf{\Phi}^{(\beta)}(2h), \hat{\mathbf{u}}\rangle.\end{aligned}$$

By the choice of β, we have $\phi(h) = \phi(2h)$ for all $0 < h < \frac{1}{2}$.

Recall that if $\mathbf{M}$ is a matrix on a vector space V with characteristic polynomial $p(x) = (x - \lambda_1)^{\ell_1} \cdots (x - \lambda_k)^{\ell_k}$ and minimal polynomial $q(x) = (x - \lambda_1)^{m_1} \cdots (x - \lambda_k)^{m_k}$, then $V = V_1 \oplus \cdots \oplus V_k$, each V_i has dimension ℓ_i, $\mathbf{M}$ is invariant on V_i, and $(\mathbf{M} - \lambda_i\mathbf{I})^{m_i}V_i = 0$ (m_i is called the *order* of λ_i). Moreover, according to the Jordan decomposition theorem,

$$V_i = U_{i1} \oplus \cdots \oplus U_{ir_i},$$

where each $s_{ij} := \dim U_{ij} \leq m_i$, with at least one of the $s_{ij} = m_i$; each U_{ij} is generated by

$$\mathbf{u}_1 = \mathbf{u},\ \mathbf{u}_2 = (\mathbf{M} - \lambda_i\mathbf{I})\mathbf{u}, \ldots,\ \mathbf{u}_{s_{ij}} = (\mathbf{M} - \lambda_i\mathbf{I})^{s_{ij}-1}\mathbf{u} \tag{5.1}$$

and $(\mathbf{M}-\lambda_i\mathbf{I})^{s_{ij}}\mathbf{u} = 0$ for some $\mathbf{u}$. Note that the last vector in (5.1) is a λ_i-eigenvector of $\mathbf{M}$.

Lemma 5.1 can be strengthened as follows.

THEOREM 5.2. *Let $\lambda \neq 0, 2$ be an eigenvalue of $\mathbf{W}_N^+$ and let $\beta = -\ln(\lambda/2)/(2\ln 2)$. Suppose there exists an m such that $(\mathbf{W}_N^+ - \lambda\mathbf{I})^{m-1}\mathbf{u} \neq \mathbf{0}, (\mathbf{W}_N^+ - \lambda\mathbf{I})^m\mathbf{u} = \mathbf{0}$. Then*

$$\langle \mathbf{\Phi}^{(\beta)}(h), \mathbf{u}\rangle = \sum_{k=1}^{m}(\ln h)^{k-1}p_k(h), \quad 0 < h < \frac{1}{2}, \tag{5.2}$$

where $p_k(h) = p_k(2h)$ for all $h > 0$ and $p_m \neq 0$. In particular, if $m = 1$, then $\langle \mathbf{\Phi}^{(\beta)}(h), \mathbf{u}\rangle = p_1(h)$.

Proof. Let

$$\mathbf{u}_m = \mathbf{u}, \ldots,\ \mathbf{u}_1 = (\mathbf{W}_N^+ - \lambda\mathbf{I})^{m-1}\mathbf{u}$$

and let $\phi_k(h) = \langle \mathbf{\Phi}^{(\beta)}(h), \mathbf{u}_k\rangle$. Note that $\mathbf{u}_1$ is a λ-eigenvector of $\mathbf{W}_N^+$. Hence by Lemma 4.3, $\phi_1 \neq 0$, and Lemma 5.1,

$$\phi_1(h) = \phi_1(2h) \qquad \forall\, 0 < h < \frac{1}{2}.$$

Let $g_1(h) = \phi_1(h)$. For $\mathbf{W}_N^+\mathbf{u}_2 = \lambda\mathbf{u}_2 + \mathbf{u}_1$, by the same argument as in Lemma 5.1, we have

$$\phi_2(h) = \phi_2(2h) + \frac{1}{\lambda}\phi_1(2h).$$

Let $g_2(h) = \phi_2(h) + \frac{\ln h}{\lambda \ln 2} g_1(h)$. Then

$$\phi_2(h) = g_2(h) - \frac{\ln h}{\lambda \ln 2} g_1(h),$$

and for $0 < h < \frac{1}{2}$,

$$\begin{aligned} g_2(h) &= \phi_2(2h) + \frac{1}{\lambda}\phi_1(h) + \frac{\ln h}{\lambda \ln 2}\, g_1(h) \\ &= \phi_2(2h) + \frac{\ln(2h)}{\lambda \ln 2} g_1(2h) \\ &= g_2(2h). \end{aligned}$$

Let $g_3(h) = \phi_3(h) + \frac{\ln h}{\lambda \ln 2} g_2(h) - \frac{(\ln h)(\ln(2h))}{2(\lambda \ln 2)^2} g_1(h)$. Then by a similar argument as above, we have

$$g_3(h) = g_3(2h) \qquad \forall\, 0 < h < \frac{1}{2}.$$

Inductively, we can find g_j, $1 \le j \le m$, such that $g_j(h) = g_j(2h)$ for $0 < h < \frac{1}{2}$ and

$$\phi_j(h) = g_j(h) + \sum_{k=1}^{j-1} \frac{(-1)^{j-k}}{(j-k)!\,(\lambda \ln 2)^{j-k}} \left(\prod_{l=1}^{j-k} \ln(2^{l-1}h) \right) g_k(h).$$

For $j = m$, we group those terms with $(\ln h)^k$ together and denote the corresponding coefficient by $p_k(h)$. Then p_k satisfies the periodic condition, and $p_m(h) = c\phi_1(h) \neq 0$. If we extend p_k by $p_k(h) = p_k(2h)$ to all h, the theorem follows.

Now, we define

$$\Lambda_{\max} = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } \mathbf{W}_N^+ \text{ and } |\lambda| \neq 2\}.$$

PROPOSITION 5.3. *Suppose f is an L_c^2-solution of* (1.1). *Then* $\frac{1}{2} \le \Lambda_{\max} < 2$.

Proof. We first claim that $\Lambda_{\max} < 2$. Otherwise, let λ be an eigenvalue with $|\lambda| > 2$ and let $\mathbf{u}$ be a corresponding eigenvector. By using Proposition 4.1, we have

$$\left\langle \mathbf{\Phi}\left(\frac{h}{2^m}\right), \mathbf{u} \right\rangle = \left(\frac{\lambda}{2}\right)^m \langle \mathbf{\Phi}(h), \mathbf{u} \rangle \quad \forall\, 0 < h < \frac{1}{2}.$$

By Lemma 4.3, there exists $0 < h < \frac{1}{2}$ such that $|\langle \mathbf{\Phi}(h), \mathbf{u} \rangle| \neq 0$. Hence $|\langle \mathbf{\Phi}(\frac{h}{2^m}), \mathbf{u} \rangle|$ does not tend to zero as $m \to \infty$. This is a contradiction, and the claim follows.

If $\Lambda_{\max} < \frac{1}{2}$, then for any $\mathbf{u}$ in Theorem 5.2, the corresponding β satisfies $\operatorname{Re}\beta = -\ln(|\lambda|/2)/(2\ln 2) > 1$ and hence $\limsup_{h\to 0^+} \frac{1}{h^2}|\langle \mathbf{\Phi}(h), \mathbf{u} \rangle| = 0$. Since all such $\mathbf{u}$ form a Jordan basis, it follows that

$$\limsup_{h\to 0^+} \frac{1}{h^2} \int_{-\infty}^{\infty} |\Delta_h f(t)|^2\, dt = \limsup_{h\to 0^+} \frac{1}{h^2} \langle \mathbf{\Phi}(h), \mathbf{e}_0 \rangle = 0.$$

This implies that $\sup_{h>0} \frac{1}{h^2} \int_{-\infty}^{\infty} |\Delta_h f|^2 < \infty$ so that $f' \in L^2(\mathbb{R})$ and $\int |f'|^2 = \lim_{h\to 0} \frac{1}{h^2} \int_{-\infty}^{\infty} |\Delta_h f|^2 = 0$. This implies that $f = 0$ a.e., a contradiction.

THEOREM 5.4. *Suppose f is an L^2_c-solution of the dilation equation (1.1). Let $\alpha = -\ln(\Lambda_{\max}/2)/(2\ln 2)$ and let m be the highest order of the eigenvalues λ of $\mathbf{W}_N^+$ such that $|\lambda| = \Lambda_{\max}$. Then*

$$\lim_{h\to 0^+}\left(\frac{1}{h^{2\alpha}|\ln h|^{m-1}}\int_{-\infty}^{\infty}|\Delta_h f(x)|^2dx - p(h)\right) = 0, \tag{5.3}$$

where $p(h)$ is a nonzero bounded continuous multiplicative periodic function of period 2, i.e., $p(h) = p(2h), h > 0$.

Proof. Write $\mathbf{e}_0 = \sum_i b_i\mathbf{u}_i$, where $\{\mathbf{u}_i\}$ is a Jordan basis corresponding to the matrix $\mathbf{W}_N^+$. Let λ be the eigenvalues (there may be more than one) such that $|\lambda| = \Lambda_{\max}$ and has highest order m. By Theorem 5.2 and the choice of α, the terms $|\langle\mathbf{\Phi}(h), \mathbf{u}_i\rangle|$ of the form $h^{2\alpha}|\ln h|^{m-1}p(h)$ dominate $\langle\mathbf{\Phi}(h), \mathbf{e}_0\rangle$ as $h \to 0^+$; the corresponding coefficient b_i's are not all zero since $|\langle\mathbf{\Phi}(h), \mathbf{e}_0\rangle| \geq c|\langle\mathbf{\Phi}(h), \mathbf{u}_i\rangle|$ for some $c > 0$. We hence have

$$\int_{-\infty}^{\infty}|\Delta_h f(x)|^2dx = \langle\mathbf{\Phi}(h), \mathbf{e}_0\rangle = h^{2\alpha}|\ln h|^{m-1}p(h) + \delta(h),$$

where $p(h) = p(2h)$ and $\lim_{h\to 0^+}\delta(h)/(h^{2\alpha}|\ln h|^{m-1}) = 0$, and the theorem follows.

COROLLARY 5.5. *Suppose λ is an eigenvalue of $\mathbf{W}_N^+$ such that $|\lambda| = \Lambda_{\max} = \frac{1}{2}$ and λ has order 1. Then $\lambda = \frac{1}{2}$, f is differentiable a.e., and $f \in L^2(\mathbb{R})$.*

Proof. Let $\beta = -\ln(\lambda/2)/(2\ln 2) = 1 + i\theta$ and let $\mathbf{u}$ be the λ-eigenvector. Then

$$\begin{aligned}\frac{1}{h^2}\int_{-\infty}^{\infty}|\Delta_h f(x)|^2dx &= h^{i2\theta}\langle\mathbf{\Phi}^{\beta}(h), \mathbf{u}\rangle + o(h)\\ &= h^{i2\theta}p(h) + o(h),\end{aligned}$$

where $o(h) \to 0$ as $h \to 0$ and p is bounded and $p(h) = p(2h)$. It is well known that if $\sup_{h>0}\frac{1}{h^2}\int_{-\infty}^{\infty}|\Delta_h f(x)|^2dx$ is bounded, then f' exists a.e., $f' \in L^2(\mathbb{R})$, and

$$\lim_{h\to 0^+}\frac{1}{h^2}\int_{-\infty}^{\infty}|\Delta_h f(x)|^2dx = \int_{-\infty}^{\infty}|f'(x)|^2dx.$$

This implies that $\lim_{h\to 0^+}h^{i2\theta}p(h)$ exists. In view of periodicity, we have $\theta = 0$ and $p(h) = C$, and the corollary follows.

COROLLARY 5.6. *Let f, α, and m be as in Theorem 5.4. Then*

$$\sup_{n>0}\frac{1}{n^{m-1}2^{-2n\alpha}}\int_{2^{n-1}\pi\leq|\omega|<2^n\pi}|\hat{f}(\omega)|^2d\omega < \infty.$$

Proof. By using the Plancherel Theorem, we have

$$\begin{aligned}\varphi(h) &:= \frac{1}{h^{2\alpha}|\ln h|^{m-1}}\int_{-\infty}^{\infty}|\Delta_h f(x)|^2dx\\ &= \frac{C_1}{h^{2\alpha}|\ln h|^{m-1}}\int_{-\infty}^{\infty}|\hat{f}(\omega)|^2\sin^2(h\omega/2)d\omega\\ &\geq \frac{C_2}{h^{2\alpha}|\ln h|^{m-1}}\int_{\frac{\pi}{2h}\leq|\omega|<\frac{\pi}{h}}|\hat{f}(\omega)|^2d\omega.\end{aligned}$$

Since $\varphi(h)$ is bounded by Theorem 5.4, the result follows by taking $h = 2^{-n}$.

We remark that the above corollary implies that for any $r < \alpha$, f is in the Bosov space $B_2^{r,\infty}$ in the sense used in [CD, Thm. 3.3]. It also implies that $\int_{-\infty}^{\infty} |\omega^r \hat{f}(\omega)|^2 d\omega < \infty$ for $r < \alpha$, which is proven in [V]. By using the same technique as in [CD, Thm. 5.1], we can improve a pointwise estimate of $\hat{f}(\omega)$ presented there.

COROLLARY 5.7. *Let f, α, and m be as in Theorem* 5.4 *and assume that* $\sum c_{2n} = \sum c_{2n+1} = 1$. *Then*

$$|\hat{f}(\omega)| \le \frac{C(\ln(1+|\omega|))^{(m-1)/2}}{(1+|\omega|)^\alpha}.$$

Proof. For $\omega \in [2^{n-1}\pi, 2^n\pi], n \ge 1$, the assumption on the coefficients implies that $\hat{f}(2k\pi) = 0$ [Ch]. Hence

$$\begin{aligned}|\hat{f}(\omega)|^2 &\le \int_{2^{n-1}\pi \le |\xi| \le 2^n\pi} \frac{d|\hat{f}(\xi)|^2}{d\xi} d\xi \\ &\le C_1 \left(\int |\frac{d\hat{f}(\xi)}{d\xi}|^2 d\xi\right)^{1/2} \left(\int_{2^{n-1}\pi \le |\xi| \le 2^n\pi} |\hat{f}(\xi)|^2 d\xi\right)^{1/2} \\ &\le C_2 n^{(m-1)/2} 2^{-n\alpha} \\ &\le \frac{C_3(\ln(1+|\omega|))^{(m-1)/2}}{(1+|\omega|)^\alpha}.\end{aligned}$$

We define the *L^2-Lipschitz exponent* of a function $g \in L^2(\mathbb{R})$ as

$$\alpha := L^2\text{-Lip}(g) = \inf\{\beta > 0 : 0 < \limsup_{h\to 0^+} \frac{1}{h^{2\beta}} \int_{-\infty}^{\infty} |\Delta_h g(t)|^2 \, dt\}. \tag{5.4}$$

Note that $0 < \limsup_{h\to 0^+} \frac{1}{h^2} \int_{-\infty}^{\infty} |\Delta_h g(t)|^2 \, dt$ (otherwise, we can derive a contradiction by using the argument in the last paragraph of Proposition 5.3 to show that $g = 0$ a.e.). Hence $0 \le \alpha \le 1$. Also,

$$\limsup_{h\to 0^+} \frac{1}{h^{2\beta}} \int_{-\infty}^{\infty} |\Delta_h g(t)|^2 \, dt = \begin{cases} 0 & \text{if } \beta < \alpha, \\ \infty & \text{if } \beta > \alpha. \end{cases}$$

The next corollary follows directly from Theorem 5.4.

COROLLARY 5.8. *Suppose f is an L_c^2-solution of the dilation equation* (1.1). *Let* $\alpha = -\ln(\Lambda_{\max}/2)/(2\ln 2)$. *Then* $0 < \alpha \le 1$ *is the $L^2(\mathbb{R})$-Lipschitz exponent of f.*

Corollaries 5.6 and 5.7 give certain estimates of the Fourier transform of f. In the following we consider yet another sharper estimate on the average of the Fourier transformation of the L^2-scaling function. We make use of a special form of Tauberian theorem to convert the asymptotic result in Theorem 5.4 into the frequency domain. For $\beta, \gamma \in \mathbb{R}$, let

$$\mathcal{W}_{\beta,\gamma} = \left\{ g : g \text{ loc. Riem. integ. on } \mathbb{R}^+, \sum_{k=-\infty}^{\infty} \sup_{2^k \le t < 2^{k+1}} t^\beta |\ln t|^\gamma |g(t)| < \infty \right\}.$$

The following theorem is proven in [L3, Cor. 4.5].

THEOREM 5.9. *Suppose $F \geq 0$ is measurable on $\mathbb{R}^+$ and is bounded on $[0, a)$ for some $a > 0$. Let $g \in \mathcal{W}_{\beta,\gamma}(\mathbb{R}^+), \beta > 0, \gamma \geq 0$ be such that $G(\xi) = \int_0^\infty g(t)t^{\beta-1+i\xi}dt \neq 0$ for all ξ. Then*

$$\lim_{T\to\infty}\left(\frac{1}{T^\beta(\ln T)^\gamma}\int_0^\infty F(t)g\left(\frac{t}{T}\right)dt - P(T)\right) = 0$$

if and only if

$$\lim_{T\to\infty}\left(\frac{1}{T^\beta(\ln T)^\gamma}\int_0^T F(t)dt - Q(T)\right) = 0,$$

where P and Q are bounded multiplicative periodic functions of the same period and $P \not\equiv 0$ if and only if $Q \not\equiv 0$.

THEOREM 5.10. *Suppose f is the L^2_c-solution of* (1.1) *with L^2-Lipschitz exponent $\alpha \neq 1$. Let m be the highest order of the eigenvalues λ such that $|\lambda| = \Lambda_{\max}$. Then for any s such that $\alpha < s$, there exists a bounded continuous multiplicative periodic function q such that $q(T) = q(2T)$ and*

$$\lim_{T\to\infty}\left(\frac{1}{T^{2(s-\alpha)}(\ln T)^{m-1}}\int_{-T}^T |\xi^s\hat{f}(\xi)|^2 d\xi - q(T)\right) = 0.$$

Proof. By using the Phancherel Theorem as in Corollary 5.6,

$$\begin{aligned}\varphi(h) &= \frac{1}{h^{2\alpha}|\ln h|^{m-1}}\int_{-\infty}^\infty |\hat{f}(\omega)|^2\sin^2(h\omega/2)d\omega \\ &= \frac{1}{h^{2(\alpha-s)}|\ln h|^{m-1}}\int_{-\infty}^\infty |\omega^s\hat{f}(\omega)|^2\frac{\sin^2(h\omega/2)}{|h\omega|^{2s}}d\omega.\end{aligned}$$

By letting

$$F(\omega) = |\omega^s\hat{f}(\omega)|^2 + |\omega^s\hat{f}(-\omega)|^2, \quad g(\omega) = \frac{\sin^2(\omega/2)}{|\omega|^{2s}}, \quad h = \frac{1}{T},$$

the above reduces to

$$\varphi\left(\frac{1}{T}\right) = \frac{1}{T^{2(s-\alpha)}|\ln T|^{m-1}}\int_0^\infty F(\omega)g(\frac{\omega}{T})d\omega.$$

Let $\beta = 2(s-\alpha)$. Then $g \in \mathcal{W}_{\beta,m-1}$. Indeed, for $0 < \alpha < 1$,

$$\begin{aligned}&\sum_{k=-\infty}^\infty \sup_{2^k\leq\omega<2^{k+1}} \omega^\beta|\ln\omega|^{m-1}g(\omega) \\ &= \sum_{k=-\infty}^\infty \sup_{2^k\leq\omega<2^{k+1}} |\ln\omega|^{m-1}\frac{\sin^2(\omega/2)}{\omega^{2\alpha}} \\ &\leq C\left(\sum_{k=0}^\infty |k|^{m-1}2^{-2k(1-\alpha)} + \sum_{k=0}^\infty |k|^{m-1}2^{-2\alpha k}\right) < \infty.\end{aligned}$$

Also note that for $\alpha < s$, g is integrable and

$$\begin{aligned}G(\xi) &= \int_0^\infty g(\omega)\omega^{2(s-\alpha)-1+i\xi}d\omega = \int_0^\infty \sin^2(\omega/2)\omega^{i\xi-2\alpha-1}d\omega \\ &= -\frac{\sqrt{\pi}\Gamma(-\alpha+\frac{i\xi}{2})}{4^{(\alpha+1-\frac{i\xi}{2})}\Gamma(\alpha+\frac{1}{2}-\frac{i\xi}{2})} \neq 0\end{aligned} \tag{5.5}$$

for all ξ (where $\Gamma(\xi)$ is the gamma function which has no zero and has simple poles at $\xi = 0, -1, -2, \ldots$; the calculation is by Mathematica). Hence the conditions of g in Theorem 5.9 are fulfilled. Together with Theorem 5.4, there exists a nonzero q such that $q(2T) = q(T)$ for all T, and

$$\lim_{T\to\infty}\left(\frac{1}{T^{2(s-\alpha)}(\ln T)^{m-1}}\int_{-T}^{T}|\omega^s\hat{f}(\omega)|^2 d\omega - q(T)\right) = 0.$$

6. Higher-order L^2-Lipschitz exponent. In this section, we consider the higher-order difference so that the Lipschitz exponent is allowed to be greater than 1. For any interger $l > 0$, we define the lth-order difference of a function $g \in L^2(\mathbb{R})$ by

$$\Delta_h^{(l)}f(x) = \sum_{k=0}^{l}(-1)^k\binom{l}{k}f(x-kh)$$

and the L^2-Lipschitz exponent of g by

$$\alpha := L^2\text{-Lip}(g) = \inf\left\{\beta > 0 : 0 < \limsup_{h\to 0^+}\frac{1}{h^{2\beta}}\int_{-\infty}^{\infty}|\Delta_h^{(l)}g(t)|^2\,dt\right\}. \tag{6.1}$$

It is well known that $0 \le \alpha \le l$. For $0 < \alpha < 1$, the definition used in (5.4) coincides with new definition here, but for $\alpha = 1$, the two definitions may or may not be the same. We will clarify this situation in the following. Furthermore, we show that the asymptotic properties in the last section are also preserved for higher-order cases.

For simplicity, we only consider the case $l = 2$. Let f be the L_c^2-solution of (1.1), let

$$\tilde{\Phi}_n(h) = \int_{-\infty}^{\infty}\Delta_h^{(2)}f(x+n)\Delta_h^{(2)}f(x)dx$$

and let $\tilde{\mathbf{\Phi}}(h) = [\tilde{\Phi}_0(h), \tilde{\Phi}_1(h), \ldots, \tilde{\Phi}_N(h)]$.

LEMMA 6.1. *Suppose f is the L_c^2-solution of* (1.1). *Then for any* $\mathbf{u} \in \mathbb{C}^{N+1}$,

$$\langle\tilde{\mathbf{\Phi}}(h), \mathbf{u}\rangle = C\left\langle \mathbf{\Phi}(h) - \frac{1}{4}\mathbf{\Phi}(2h), \mathbf{u}\right\rangle. \tag{6.2}$$

Proof. Let $\mathbf{u} \in \mathbb{C}^{N+1}$. Then

$$\begin{aligned}\langle\tilde{\mathbf{\Phi}}(h), \mathbf{u}\rangle &= \sum_{n=0}^{N}u_n\int_{-\infty}^{\infty}\Delta_h^{(2)}f(x+n)\Delta_h^{(2)}f(x)dx \\ &= C\sum_{n=0}^{N}u_n\int_{-\infty}^{\infty}|\hat{f}(\omega)|^2e^{in\omega}\sin^4(h\omega)d\omega\end{aligned}$$

$$= C\sum_{n=0}^{N} u_n \int_{-\infty}^{\infty} |\hat{f}(\omega)|^2 e^{in\omega} \left(\sin^2(h\omega) - \frac{1}{4}\sin^2(2h\omega)\right) d\omega$$

$$= C\langle \mathbf{\Phi}(h) - \frac{1}{4}\mathbf{\Phi}(2h), \mathbf{u}\rangle.$$

LEMMA 6.2. *Let λ be an eigenvalue of $\mathbf{W}_N^+$, $\lambda \neq 0$ or 2.*
(i) *Let* $\mathbf{u}$ *be a λ-eigenvector. Then*

$$\langle \tilde{\mathbf{\Phi}}(h), \mathbf{u}\rangle \begin{cases} \not\equiv 0 & \text{if } \lambda \neq \frac{1}{2} \\ \equiv 0 & \text{if } \lambda = \frac{1}{2}. \end{cases}$$

(ii) *Suppose $\lambda = \frac{1}{2}$ has order $m > 1$, and let* $\mathbf{u}$ *be such that $(\mathbf{W}_N^+ - \frac{1}{2}\mathbf{I})^{m-1}\mathbf{u} \neq 0$. Then*

$$\langle \tilde{\mathbf{\Phi}}(h), \mathbf{u}\rangle = \tilde{p}(h)(\ln h)^{m-2}h^2 + \delta(h), \tag{6.3}$$

where $\tilde{p}$ is a nonzero bounded continuous function with $\tilde{p}(h) = \tilde{p}(2h)$ and $\delta(h)$ has order smaller than $(\ln h)^{m-2}h^2$.

Proof. (i) Note that for $\beta = -\ln(\lambda/2)/(2\ln 2)$, Lemma 5.1 implies that

$$\langle \mathbf{\Phi}(h), \mathbf{u}\rangle = p(h)h^{2\beta}$$

for some nonzero bounded continuous p such that $p(h) = p(2h)$. Hence by (6.2),

$$\langle \tilde{\mathbf{\Phi}}(h), \mathbf{u}\rangle = C\left(1 - \frac{2^{2\beta}}{4}\right)p(h)h^{2\beta}, \tag{6.4}$$

and the result follows.

(ii) Theorem 5.2 implies that $\langle \mathbf{\Phi}(h), \mathbf{u}\rangle$ has order $(\ln h)^{m-1}p(h)h^2$ as $h \to 0$, and $\frac{1}{4}\langle \mathbf{\Phi}(2h), \mathbf{u}\rangle$ has order $\frac{1}{4}(\ln 2h)^{m-1}p(2h)(2h)^2 = (\ln 2h)^{m-1}p(h)h^2$. Note that $(\ln h)^{m-1} - (\ln 2h)^{m-1}$ is of order $(\ln h)^{m-2}$. Consequently,

$$\langle \tilde{\mathbf{\Phi}}(h), \mathbf{u}\rangle = \left\langle \mathbf{\Phi}(h) - \frac{1}{4}\mathbf{\Phi}(2h), \mathbf{u}\right\rangle = \tilde{p}(h)(\ln h)^{m-2}h^2 + \delta(h),$$

where $\tilde{p}$ and $\delta(h)$ is as asserted (the order of $\delta(h)$ follows from the same argument and Theorem 5.2).

By using Lemma 6.2(i), we can extend Corollary 5.8 as follows.

PROPOSITION 6.3. *Suppose f is an L^2_c-solution of* (1.1). *Then the L^2-Lipschitz exponent of f is given by*

$$0 < \alpha = -\ln(\tilde{\Lambda}_{\max}/2)/(2\ln 2) \leq 2, \tag{6.5}$$

where

$$\tilde{\Lambda}_{\max} := \max\left\{|\lambda'| : \lambda' \textit{ eigenvalue of } \mathbf{W}_N^+ \textit{ and } |\lambda'| < \frac{1}{2}\right\}$$

if $\frac{1}{2}$ is an eigenvalue of order 1 *and is the only eigenvalue of modulus $\frac{1}{2}$*;

$$\tilde{\Lambda}_{\max} := \Lambda_{\max}$$

otherwise.

In the first case, f is differentiable a.e. and $f' \in L^2(\mathbb{R})$ (see the proof of Corollary 5.5), and f has Lipschitz exponent > 1. In the other case, the new and old definitions coincide.

The corresponding extension for Theorem 5.4 is the following.

THEOREM 6.4. *For the above α, let m be the highest order among those λ such that $|\lambda| = \tilde{\Lambda}_{\max}$. Then*

$$\lim_{h\to 0^+}\left(\frac{1}{h^{2\alpha}|\ln h|^{m-1}}\int_{-\infty}^{\infty}|\Delta_h^{(2)}f(x)|^2dx - p(h)\right) = 0 \tag{6.6}$$

except for the special case where $\lambda = \frac{1}{2}$ and the order m is strictly greater then the other eigenvalues of moduli $\frac{1}{2}$; in such a case, $\alpha = 1$ and

$$\lim_{h\to 0^+}\left(\frac{1}{h^{2}|\ln h|^{m-2}}\int_{-\infty}^{\infty}|\Delta_h^{(2)}f(x)|^2dx - p(h)\right) = 0. \tag{6.6$'$}$$

For the Fourier asymptotic result corresponding to Theorem 5.10, we note that

$$\begin{aligned}\tilde{\varphi}(h) &= \frac{1}{h^{2\alpha}|\ln h|^{m-1}}\int_{-\infty}^{\infty}|\Delta_h^{(2)}f(x)|^2dx\\ &= \frac{1}{h^{2\alpha}|\ln h|^{m-1}}\int_{-\infty}^{\infty}|\hat{f}(\omega)|^2\sin^4(\frac{h\omega}{2})d\omega.\end{aligned}$$

By taking

$$g(\omega) = \frac{\sin^4(\omega/2)}{|\omega|^{2s}}$$

and observing that for $\alpha < s$, g is integrable and

$$\begin{aligned}G(\xi) &= \int_0^{\infty} g(\omega)\omega^{2(s-\alpha)-1+i\xi}d\omega = \int_0^{\infty}\sin^4(\omega/2)\omega^{i\xi-2\alpha-1}d\omega\\ &= \frac{\sqrt{\pi}(4^{(\alpha-1-\frac{i\xi}{2})}-1)\Gamma(-\alpha+\frac{i\xi}{2})}{4^{(\alpha+1-\frac{i\xi}{2})}\Gamma(\alpha+\frac{1}{2}-\frac{i\xi}{2})} \neq 0\end{aligned}$$

for all ξ (compare this with (5.5)), we have the following.

THEOREM 6.5. *Under the same hypotheses as in Theorem* 6.4, *for $\alpha < s$, except for the special case,*

$$\lim_{T\to\infty}\left(\frac{1}{T^{2(s-\alpha)}|\ln T|^{m-1}}\int_{-T}^{T}|\omega^s\hat{f}(\omega)|^2d\omega - q(T)\right) = 0 \tag{6.7}$$

for some nonzero bounded continuous q (depending on s) such that $q(2T) = q(T)$. For the special case, we have

$$\lim_{T\to\infty}\left(\frac{1}{T^{2(s-1)}|\ln T|^{m-2}}\int_{-T}^{T}|\omega^s\hat{f}(\omega)|^2d\omega - q(T)\right) = 0. \tag{6.7$'$}$$

For the third-order difference $\Delta_h^{(3)} f$, we can use

$$\sin^6 x = \frac{15}{16}\sin^2 x - \frac{3}{8}\sin^2 2x + \frac{1}{16}\sin^2 3x$$

to replace the relationship in (6.2). For eigenvalues $\lambda = \frac{1}{2}$ or $\frac{1}{8}$ and the order m is as in the special case in Theorem 6.4 (the corresponding α are 1 and 2, respectively), the logarithmic terms in the asymptotic formulas are of order $m-2$. For the other cases, they are $m-1$. The higher-order difference $\Delta_h^{(l)} f$ behaves the same way.

As an example, we consider Daubechies's well-known scaling function D_4. Let f be the solution of (1.1) with coefficients

$$c_0 = \frac{1+\sqrt{3}}{4}, \quad c_3 = \frac{1-\sqrt{3}}{4}, \quad c_1 = 1 - c_3, \quad c_2 = 1 - c_0.$$

It follows from a direct calculation that

$$\mathbf{W}_3^+ = \begin{pmatrix} 2 & 0 & 0 & 0 \\ \frac{9}{4} & 1 & -\frac{1}{8} & 0 \\ 0 & 2 & 0 & 0 \\ -\frac{1}{4} & \frac{9}{8} & \frac{9}{8} & -\frac{1}{8} \end{pmatrix}$$

and the eigenvalues are 2, $\frac{1}{2}$, and $-\frac{1}{8}$, where $\frac{1}{2}$ has order 2. It fits into the above special case. By Theorem 5.4, $\alpha = 1$ and

$$\frac{1}{h^2|\ln h|}\int_{-\infty}^{\infty} |\Delta_h f(x)|^2 dx \approx p(h)$$

as $h \to 0^+$. It is known that f is differentiable a.e. [D], [DL2], but the asymptotic property implies that $f' \notin L^2(\mathbb{R})$. If we consider the second-order difference, then by a direct calculation and making use of the expressions in Theorem 5.2 and (6.4), we have

$$\frac{1}{h^2}\int_{-\infty}^{\infty} |\Delta_h^{(2)} f(x)|^2 dx \approx 2p(h)$$

as $h \to 0$.

For the Fourier transformation, we cannot apply Theorem 5.10 since $\alpha = 1$, but we can use (6.7)$'$ derived from the higher-order difference. It implies that for $1 < s$, there exists a bounded continuous q (depends on s) satisfying $q(T) = q(2T)$ and

$$\psi(T) = \frac{1}{T^{2(s-1)}}\int_{-T}^{T} |\omega^s \hat{f}(\omega)|^2 d\omega \approx q(T) \tag{6.8}$$

as $T \to \infty$. We include some graphic illustrations of this in the appendix.

Appendix. For the four-coefficient dilation equation

$$f(x) = c_0 f(2x) + c_1 f(2x-1) + c_2 f(2x-2) + c_3 f(2x-3)$$

with $c_0 + c_2 = 1$, $c_1 + c_3 = 1$, we use c_0 and c_3 as independent parameters to plot the various regions and functions.

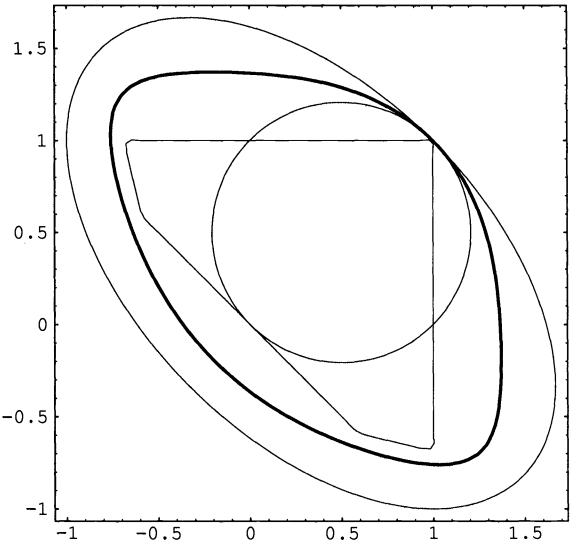


FIG. 1.

In Figure 1, the region bounded by the thicker curve corresponding to those (c_0, c_3) where all eigenvalues of $\mathbf{W}_3^+$ are less than or equal to 2 (2 is simple except at $(c_0, c_3) = (1, 1)$). This is the exact region where the L_c^2-solution exists [LW1].

The circular curve

$$\left(c_0 - \frac{1}{2}\right)^2 + \left(c_3 - \frac{1}{2}\right)^2 = \frac{1}{2}$$

is the circle of orthogonality [La]. The wavelet generated by the corresponding scaling function is orthonormal.

The triangular region is an approximation where the joint spectral radius of T_0 and T_1 restricted on H are less than 1 and the continuous solutions exist [DL1], [CH].

The ellipse is given by

$$c_0^2 + c_3^2 - c_0 - c_3 + c_0 c_3 = 0,$$

outside which no L^1-solution exists [H].

Figure 2(a) is the graph of the L^2-Lipschitz exponent $\alpha = -\ln(\Lambda_{\max}/2)/(2\ln 2)$.

Figure 2(b) is the graph of α on the circle of orthogonality, plotted in terms of the angles. Note that D_4 is the smoothest one on the circle.

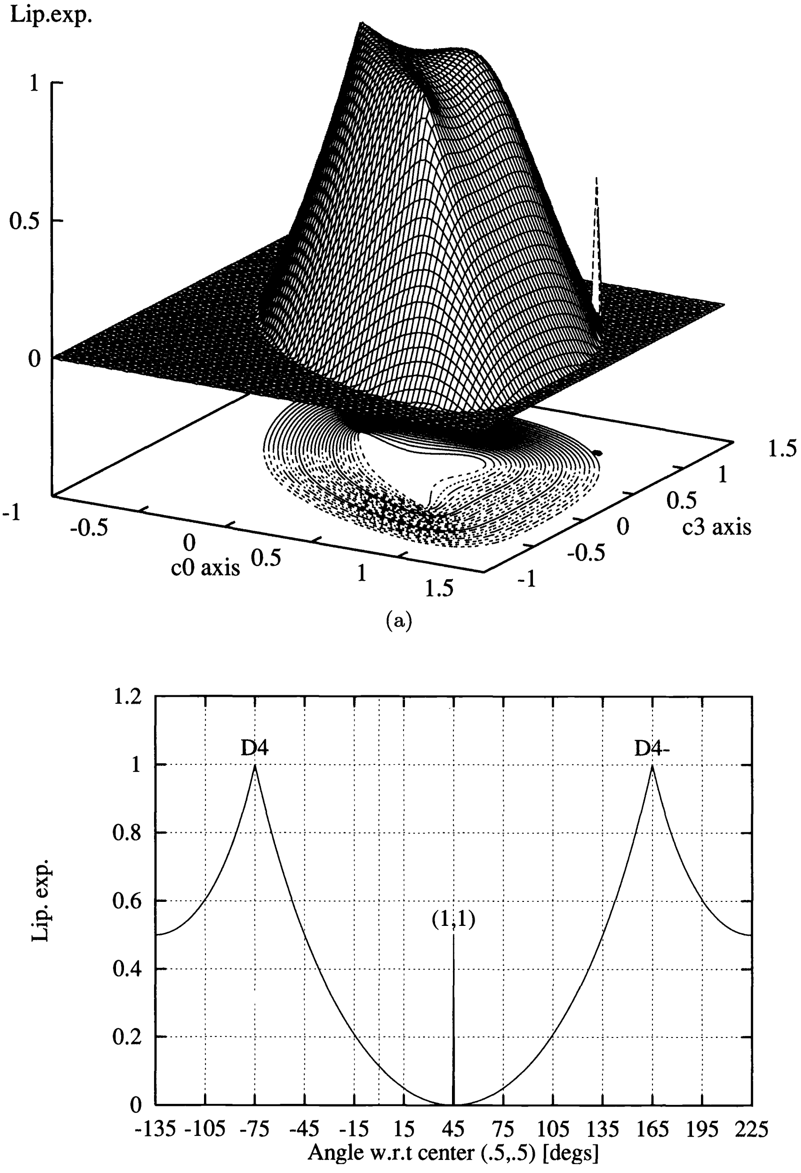
Figure 3(a) is the graph of the L^2-Lipschitz exponent $\alpha = -\ln(\tilde{\Lambda}_{\max}/2)/(2\ln 2)$, using the second-order difference.

Figure 3(b) is the cross-section of $c_0 + c_3 = 1$; Figure 3(c) is the cross-section of $c_0 = c_3$.

Figure 4(a) is the Daubechies scaling function $f = D_4$.

Figure 4(b) is its Fourier transformation $\hat{f}(\omega)$.

Figures 4(c), 4(e), and 4(g) are $\omega^s \hat{f}(\omega)$ with $s = 1.5, 1.25, 1.00$, respectively, and Figures 4(d) and 4(f) are the corresponding averages $\psi(T) = \frac{1}{T^{2(s-1)}} \int_{-T}^{T} |\omega^s \hat{f}(\omega)|^2 d\omega$.



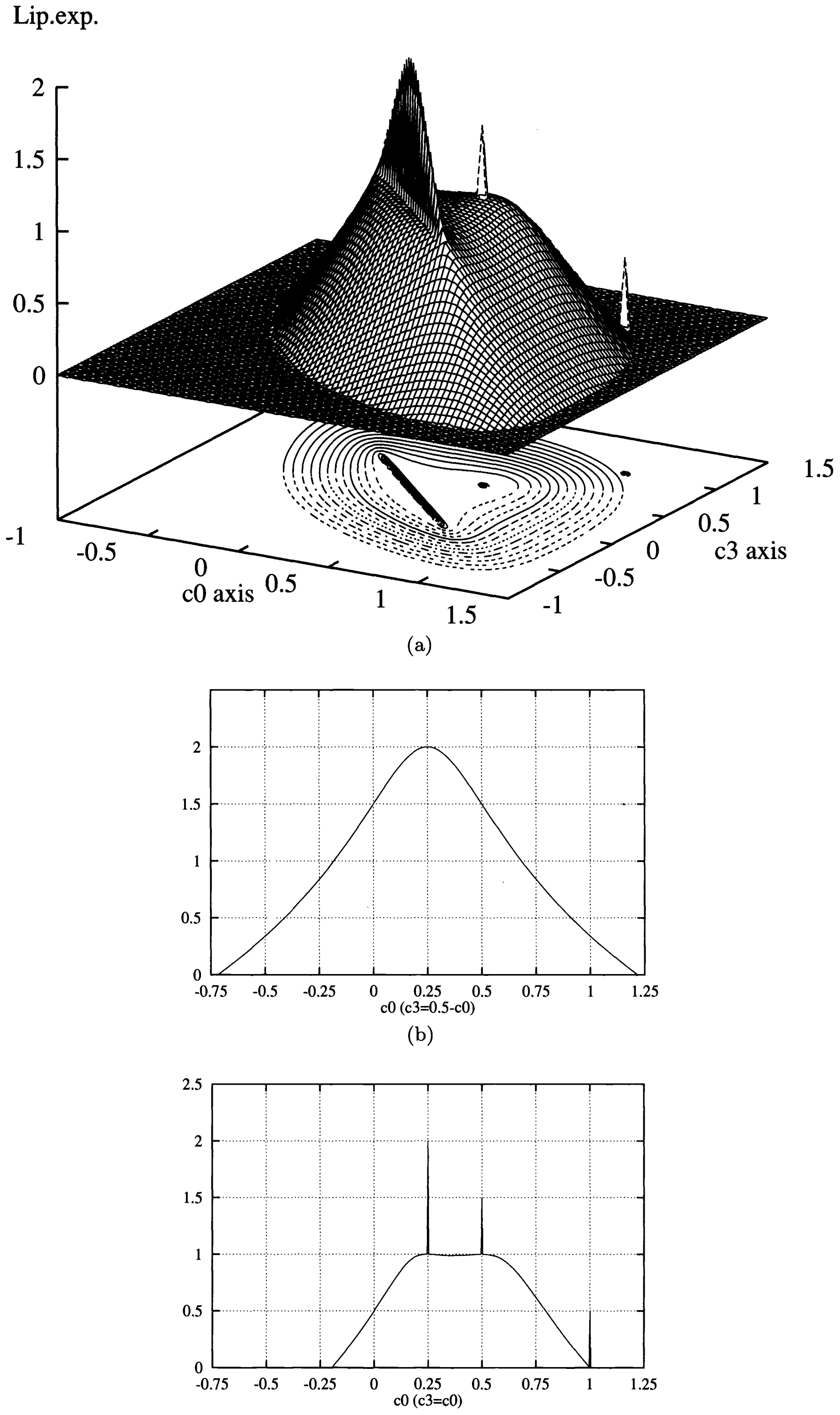
(a)

(b)

FIG. 2.

Note that the $\psi(T)$ are approximately multiplicative periodic as $T \to \infty$. They follows from (6.5). We cannot draw such conclusions from the theorem for $s = \alpha = 1$ (see Figure 4(h)). However, if we take $\psi(T) = \frac{1}{\ln T} \int_{-T}^{T} |\omega \hat{f}(\omega)|^2 d\omega$, then it looks multiplicative periodic as in Figure 4(i); we have no proof for that yet.

Acknowledgments. The graphs in the appendix are due to Mr. Wonkoo Kim, to whom we express our deep gratitude.

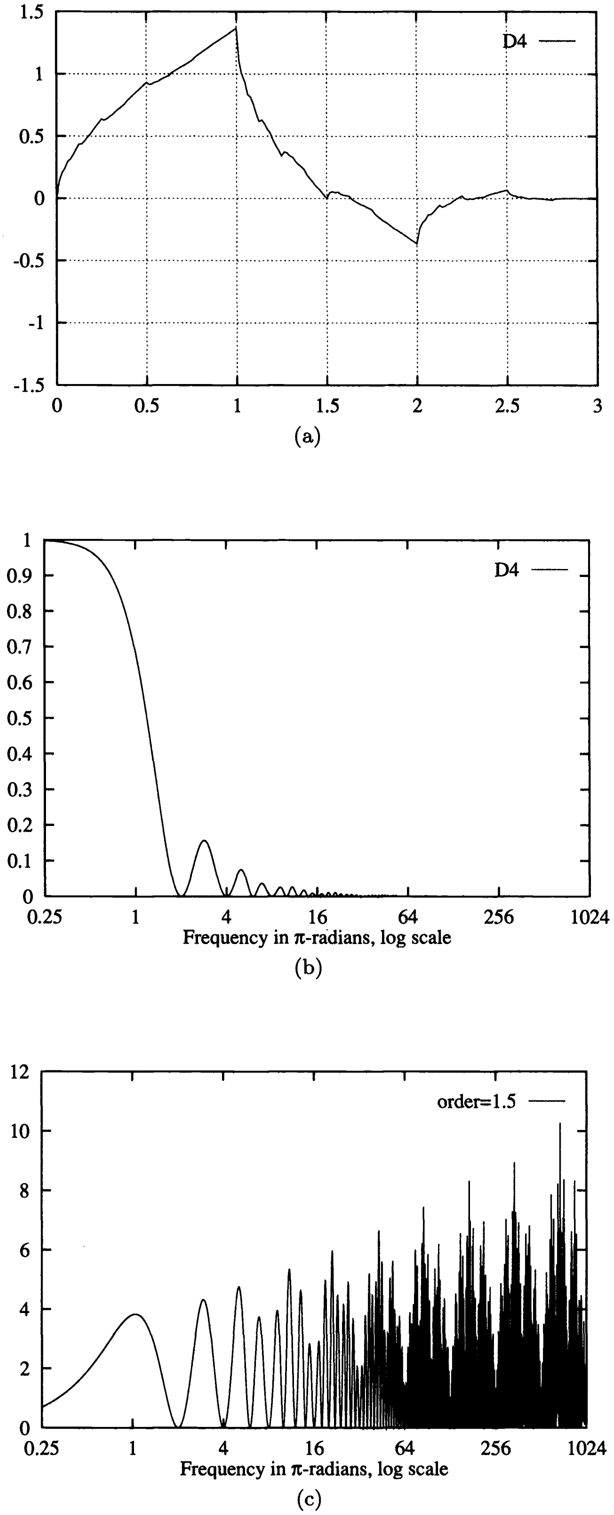

Lip.exp.
2
1.5
1
0.5
0
-1
-0.5
0
c0 axis
0.5
1
1.5
-1
-0.5
0
0.5
1
1.5
c3 axis
2
1.5
1
0.5
0
-0.75
-0.5
-0.25
0
0.25
0.5
0.75
1
1.25
c0 (c3=0.5-c0)
2.5
2
1.5
1
0.5
0
-0.75
-0.5
-0.25
0
0.25
0.5
0.75
1
1.25
c0 (c3=c0)

(a)

(b)

(c)

FIG. 3.

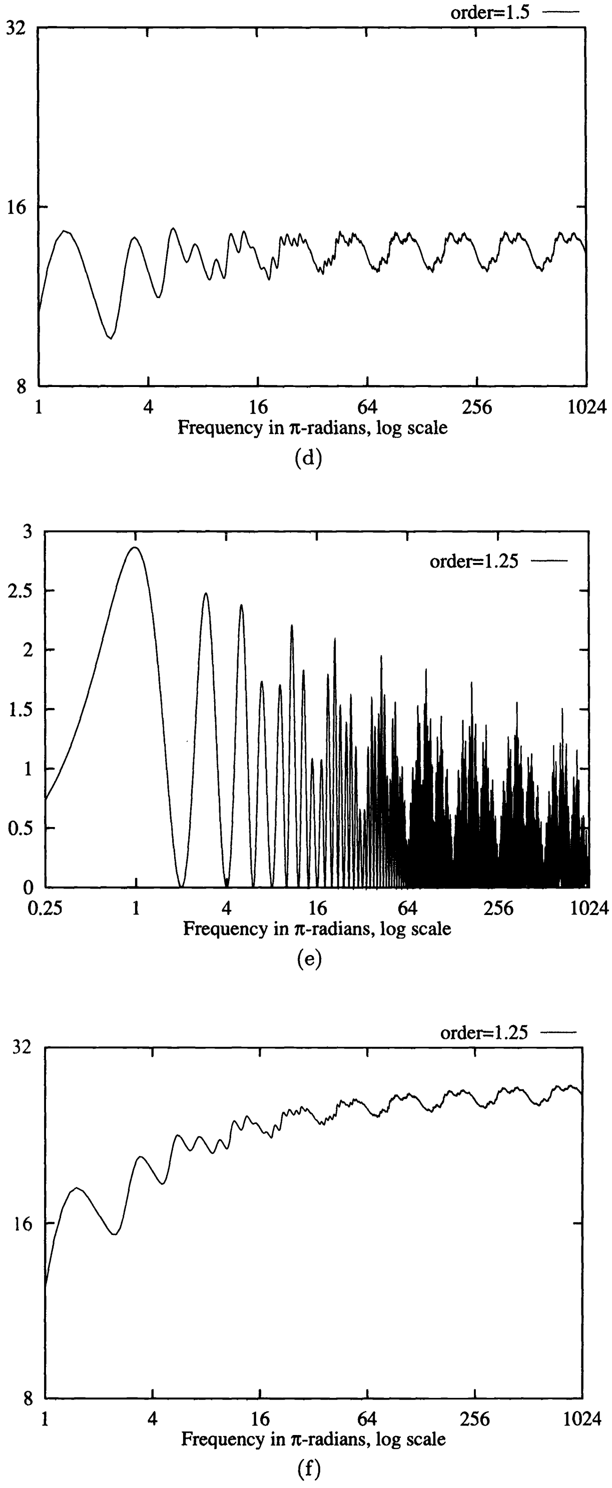

1.5
1
0.5
0
-0.5
-1
-1.5
0
0.5
1
1.5
2
2.5
3
D4
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.25
1
4
16
64
256
1024
Frequency in π-radians, log scale
D4
12
10
8
6
4
2
0
0.25
1
4
16
64
256
1024
Frequency in π-radians, log scale
order=1.5

(a)

(b)

(c)

FIG. 4.

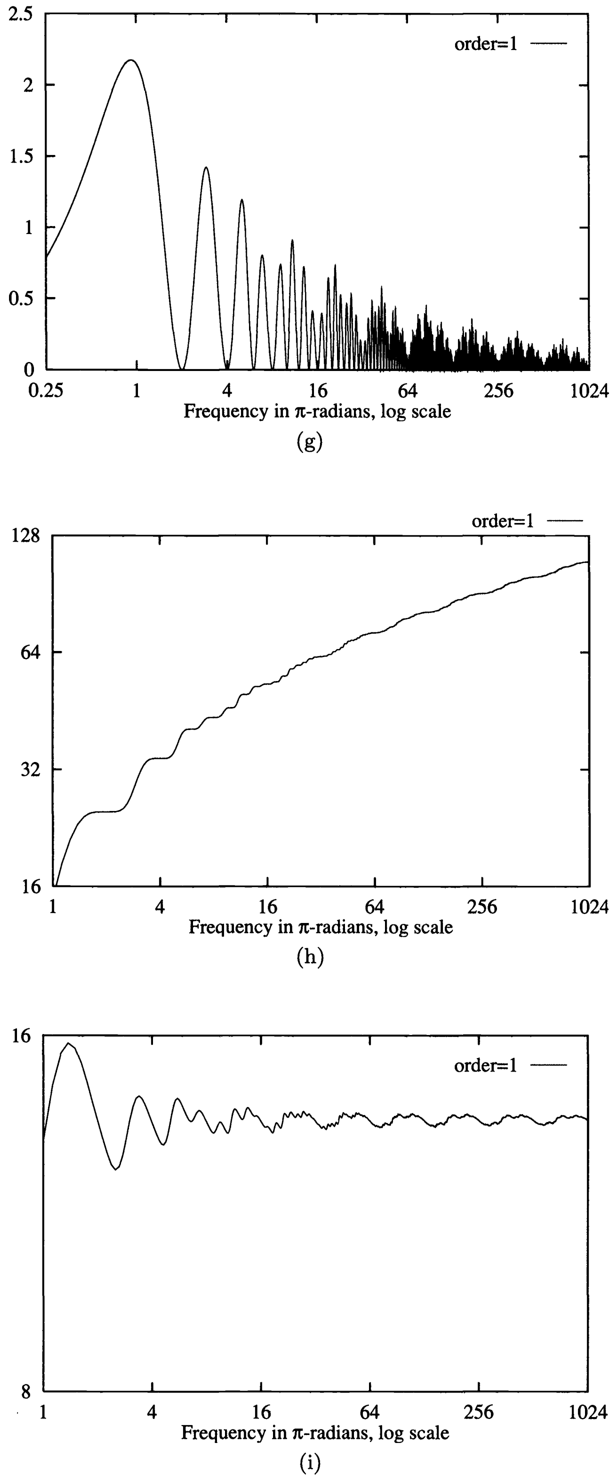


(d)

(e)

(f)

FIG. 4. (*cont.*)



(g)

(h)

(i)

FIG. 4. (*cont.*)

REFERENCES

[C] A. Cohen, *Ondelettes, analysis multiresolutions et filtres mirroirs en quadrature*, Ann. Inst. H. Poincaré Anal Non Linéare, 7 (1990), pp. 439–459.

[CD] A. Cohen and I. Daubechies, *A stability criterion for biorthogonal waveletbases and their related subband coding scheme*, Duke Math. J., 68 (1992), pp. 313–335.

[Ch] C. Chui, *An introduction to wavelets*, Academic Press, New York, 1992.

[CH] D. Colella and C. Heil, *Characterizations of scaling functions* I: *Continuous solutions*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 496–518.

[CR] J. P. Conze and A. Raugi, *Fonctions harmonique pour un operateur de transition et applications*, Bull. Soc. Math. France, 273 (1990), pp. 273–301.

[D] I. Daubechies, *Ten Lectures on Wavelets*, CBMS–NSF Region Series Conference in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[DL1] I. Daubechies and J. Lagarias, *Two-scale difference equation* I: *Global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388–1410.

[DL2] ———, *Two-scale difference equation* II: *Local regularity, infinite products and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[E] T. Eirola, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[H] C. Heil, *Methods of solving dilation equations*, in Probability and Stochastic Methods, Anal. and Prob., ASI Ser. C: Math. Phys. Sci., 372 (1992), pp. 15–45.

[Hu] J. Hutchinson, *Fractals and self-similarity*, Indiana Univ. Math. J., 30 (1981), pp. 713–747.

[K] J. P. Kahane, *Lectures on mean perodic functions*, in Letures on Mathematics and Physics, Tata Institute, 1959.

[L1] K. S. Lau, *Fractal measures and mean p-variations*, J. Funct. Anal., 108 (1992), pp. 427–457.

[L2] ———, *Dimension of a family of singular Bernoulli convolutions*, J. Funct. Anal., 116 (1993), pp. 335–358.

[L3] ———, *A weighted Tauberian theorem*, J. Fourier Anal. Appl., to appear.

[LW1] K. S. Lau and J. Wang, *Characterization of L^p-solution for the two scale dilation equations*, SIAM J. Math. Anal. 26 (1995), pp. 1018–1046.

[LW2] ———, *Mean quadratic variations and Fourier asymptotics of self-similar measures*, Monatsch. Math., 115 (1993), pp. 99–132.

[La] W. Lawton, *Necessary and sufficient conditions for constructing orthonormal wavelet bases*, J. Math. Phys., 32 (1991), pp. 57–64.

[P] J. Peetre, *New Thoughts on Besov Spaces*, Duke University Mathematics Series I, Duke University, Durham, NC, 1976.

[RL] B. Ramachandran and K. S. Lau, *Functional equations in probability theory*, Academic Press, New York, 1991.

[Sch] L. Schwartz, *Fonctions moyenne periodiques*, Ann. of Math., 48 (1947), pp. 857–927.

[S1] R. Strichartz, *Self-similar measures and their Fourier transforms*, Indiana Univ. Math. J., 39 (1990), pp. 797–817.

[S2] ——— R. Strichartz, *Self-similar measures and their Fourier transforms* II, Trans. Amer. Math. Soc., 336 (1993), pp. 335–361.

[V] L. Villemoes, *Energy moments in time and frequency for two-scale difference equation solutions and wavelets*, SIAM J. Math. Anal., 23 (1992), pp. 1519–1543.

[W] Y. Wang, *On two-scale dilation equations*, Random Comput. Dynamics, to appear.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 865–890, May 1996

014

DYADIC AFFINE DECOMPOSITIONS AND FUNCTIONAL WAVELET TRANSFORMS*

CHARLES K. CHUI† AND CHUN LI‡

Abstract. Decomposition of continuous functions can be accomplished by considering the difference of consecutive interpolation operators. When such a difference is expressed as an infinite series of some "wavelets" basis, the coefficient sequence becomes Donoho's "interpolating wavelet transform." Here, in contrast to the usual L^2-setting, no analyzing wavelet is used to describe the wavelet transform. The objective of this paper is to study the structure of such decomposition spaces, including the formulation of bases and their duals, which leads to the notion of functional wavelet transforms (FnWT) using the duals as analyzing wavelets. Such a transform retains some of the most important properties of the integral wavelet transform of Grossmann and Morlet, such as the property of vanishing moments, which has significant applications to engineering problems.

Key words. interpolating wavelets, wavelet decompositions, functional wavelet transform, vanishing moments, space of continuous functions, Dirac delta functions

AMS subject classifications. Primary 41A58; Secondary 42C30

1. Introduction.

In the study of orthogonal wavelets, a multiresolution analysis $\{V_j\}_{j\in\mathbb{Z}}$ of nested closed subspaces of $L^2 := L^2(\mathbb{R})$ along with their orthogonal complementary subspaces $\{W_j\}_{j\in\mathbb{Z}}$ are considered, namely,

$$\begin{cases} \operatorname{clos}_{L^2}\left(\bigcup_{j\in\mathbb{Z}} V_j\right) = L^2, \\ \bigcap_{j\in\mathbb{Z}} V_j = \{0\}, \\ V_{j+1} = V_j \oplus W_j, \quad j \in \mathbb{Z}, \end{cases}$$

so that the orthogonal decomposition

$$L^2 = \bigoplus_{j\in\mathbb{Z}} W_j$$

is achieved, where $V_{j+1} = V_j \oplus W_j$ means that $V_{j+1} = V_j + W_j$ and $V_j \perp W_j$. The spaces W_j can also be defined by

$$W_j = \{f - P_j^\perp f\colon\ f \in V_{j+1}\}, \quad j \in \mathbb{Z},$$

where $P_j^\perp$ is the orthogonal projection from L^2 to V_j. In this paper, we will study the decomposition of continuous functions induced by the cardinal interpolation projection.

To formulate a firm setting, the function space under consideration must be a subspace of $C := C(\mathbb{R})$, the space of continuous functions on $\mathbb{R}$. More precisely, we will consider the subspace

$$C_u = C_u(\mathbb{R}) := \{f \in C\colon\ f \text{ uniformly continuous and bounded on } \mathbb{R}\} \tag{1.1}$$

*Received by the editors May 5, 1993; accepted for publication (in revised form) September 16, 1994. This research was supported by NSF grant DMS-92-06928 and ARO contract DAAH 04-93-G-0047.

†Center for Approximation Theory, Texas A&M University, College Station, TX 77843.

‡Center for Approximation Theory, Texas A&M University, College Station, TX 77843. Current address: Institute of Mathematics, Academia Sinica, Beijing 100080, People's Republic of China.

and will use the sup norm $\|\cdot\|_\infty$. Note that C_u is closed in the topology of this norm. Suppose that $\phi \in C_u$ has compact support and

$$V_j := \left\{ \sum_{k\in\mathbb{Z}} c_k\phi(2^j \cdot -k) \colon \{c_k\} \in \ell^\infty \right\}, \quad j \in \mathbb{Z}. \tag{1.2}$$

Then each V_j is a subspace of C_u. We will call ϕ a scaling function that generates an interpolating multiresolution analysis $\{V_j\}$ of C_u provided that the following conditions are satisfied:

(i) $V_j \subset V_{j+1}$, for all $j \in \mathbb{Z}$;

(ii) $\operatorname{clos}_{C_u}\left(\bigcup_{j\in\mathbb{Z}} V_j\right) = C_u$;

(iii) $\bigcap_{j\in\mathbb{Z}} V_j = \mathbb{C}$, the set of all complex numbers;

(iv) $f \in V_j \Leftrightarrow f(2\cdot) \in V_{j+1}, j \in \mathbb{Z}$;

(v) $f \in V_j \Rightarrow f(\cdot - 2^{-j}k) \in V_j, j \in \mathbb{Z}, k \in \mathbb{Z}$;

and

(vi) $\Phi(z) := \sum_k \phi(k)z^k \neq 0$ for all $|z| = 1$.

Condition (vi) is instrumental to solving the cardinal interpolation problem with "fundamental function"

$$\phi_L \in V_0 \text{ satisfying } \phi_L(k) = \delta_{k,0}, \qquad k \in \mathbb{Z}, \tag{1.3}$$

where $\delta_{k,0}$ denotes, as usual, the Kronecker delta symbol. Under assumption (vi), ϕ_L is unique and is given by

$$\phi_L(x) = \sum_k r_k\phi(x-k), \tag{1.4}$$

where $\{r_k\}$ is the coefficient sequence of the Laurent series expansion:

$$\frac{1}{\Phi(z)} = \sum_{k\in\mathbb{Z}} r_k z^k, \qquad |z| = 1. \tag{1.5}$$

We would like to point out that condition (vi) can be weakened to be

(vi′) $\Phi_c(z) := \sum_k \phi(k+c)z^k \neq 0$ for all $|z| = 1$

and for some $c \in [0,1)$, which may not be 0. In addition, the assumption that ϕ is compactly supported is not necessary for some of our results to be valid. Of course, ϕ still has to have sufficiently fast decay such as $\phi(x) = O\left(\frac{1}{(1+|x|)^{1+\varepsilon}}\right)$ for some $\varepsilon > 0$. In this paper, however, in order to significantly simplify our presentation, we will only consider compactly supported scaling functions $\phi \in C_u$ which satisfy conditions (i) and (vi). We will see later (in §3) that (ii)–(v) are only consequences of (i) and (vi). Hence, in the study of a multiresolution analysis of C_u, only conditions (i) and (vi) are required.

From the interpolating property of ϕ_L in (1.3), it follows that the cardinal interpolation projection operators P_j from C_u to V_j are given by

$$(P_jf)(x) := \sum_{k\in\mathbb{Z}} f\left(\frac{k}{2^j}\right)\phi_L(2^jx-k), \qquad f \in C_u. \tag{1.6}$$

Parallel to the study of orthogonal wavelets, we introduce the complementary subspaces $\{W_j\}$ of $\{V_j\}$ defined by

$$W_j := \{f - P_j f\colon\ f \in V_{j+1}\}, \qquad j \in \mathbb{Z}. \tag{1.7}$$

It is clear that

$$V_{j+1} = V_j + W_j \quad \text{and} \quad V_j \cap W_j = \{0\}, \quad j \in \mathbb{Z}, \tag{1.8}$$

and we will use the direct-sum notation

$$V_{j+1} = V_j \dot{+} W_j, \qquad j \in \mathbb{Z}, \tag{1.9}$$

to describe (1.8).

One of the objectives of this paper is to study the structure of the spaces W_j along with their duals. In particular, we will discuss the properties of the two bases $\{\psi(\cdot - k)\}$ and $\{\psi_L(\cdot - k)\}$ of W_0, where ψ and ψ_L are defined as follows:

$$\psi(x) := \sum_k (-1)^{k-1}\phi(k-1)\phi(2x-k) \tag{1.10}$$

and

$$\psi_L(x) := \phi_L(2x-1). \tag{1.11}$$

It should be mentioned that the function ψ in (1.10) has already appeared in Micchelli [8], while ψ_L has been investigated by Donoho [5] for the special case when ϕ_L is assumed to have compact support.

Since ϕ has compact support, the function ψ in (1.10) is only a finite linear combination of $\phi(2\cdot -k)$ and consequently also has compact support. As to the second function ψ_L, we point out that it does not have compact support in general. In fact, in view of (1.4) and (1.5), ϕ_L, and hence ψ_L, has compact support only if $\Phi(z)$ is a monomial, which is the trivial case. In this paper, we will call both ψ and ψ_L "interpolating wavelets." We also remark that

$$\phi \equiv \phi_L \Rightarrow \psi \equiv \psi_L. \tag{1.12}$$

From (ii), (iii), and the decomposition relation (1.9), it follows that

$$C_u = \mathbb{C} \dot{+} \sum_{j \in \mathbb{Z}} W_j. \tag{1.13}$$

Hence, by using the basis $\{\psi(2^j \cdot -k)\}$ of W_j, we have

$$f(x) = f(0) + \sum_{j,k \in \mathbb{Z}} d_k^j \psi(2^j x - k), \qquad x \in \mathbb{R}, \quad f \in C_u. \tag{1.14}$$

Another objective of this paper is to study the importance of the coefficient sequence $\{d_k^j\}$ in the (unique) representation of any $f \in C_u$. This sequence may be called Donoho's interpolating wavelet transform of f, studied in some detail in [5]. Our investigation leads to the construction of the "dual" $\tilde{\psi}$ of ψ, which naturally introduces the notion of "functional wavelet transform" relative to the "analyzing wavelet functional" $\tilde{\psi}$.

Let C_u^* be the dual space of C_u; that is, C_u^* is the space that consists of all the continuous linear functionals on C_u. By $\langle f, \tilde{f}\rangle$ we mean the value of $\tilde{f} \in C_u^*$ evaluated at $f \in C_u$. Thus, $\widetilde{\psi} \in C_u^*$ is called the dual of ψ if

$$\langle \psi(\cdot - k), \widetilde{\psi}\rangle = \delta_{k,0}, \qquad k \in \mathbb{Z}. \tag{1.15}$$

It will be clear in §3 that the coefficient sequence $\{d_k^j\}$ in (1.14) is given by

$$d_k^j = \langle f(2^{-j} \cdot + 2^{-j}k), \widetilde{\psi}\rangle = 2^j \langle f, \widetilde{\psi}(2^j \cdot - k)\rangle, \qquad j, k \in \mathbb{Z}, \tag{1.16}$$

where the dilation and translation (shift) $\widetilde{\psi}(a \cdot - b)$ of the linear functional $\widetilde{\psi}$ will be defined in §2. Thus, as in the L^2-setting, this formulation naturally leads to the notion of the "functional wavelet transform," defined by

$$(W_{\widetilde{\psi}} f)(b, a) := \frac{1}{a}\left\langle f, \widetilde{\psi}\left(\frac{\cdot - b}{a}\right)\right\rangle. \tag{1.17}$$

Let π_{m-1} denote the collection of all polynomials of degree $\leq m-1$ (or order m). It will be shown that if ϕ (locally) reproduces all polynomials in π_{m-1} for some positive integer m, in the sense that there exist constants $b_{j,k}$ with polynomial growth as $k \to \infty$, such that

$$x^j = \sum_k b_{j,k}\phi(x - k), \qquad x \in \mathbb{R}, \quad j = 0, 1, \ldots, m-1, \tag{1.18}$$

then

$$(W_{\widetilde{\psi}} p)\left(\frac{k}{2^j}, \frac{1}{2^j}\right) = 2^j \langle p, \widetilde{\psi}(2^j \cdot - k)\rangle = 0, \qquad p \in \pi_{m-1}, \quad j, k \in \mathbb{Z}. \tag{1.19}$$

This property of vanishing moment has significant applications to engineering problems that require detection, data compression, etc.

2. Bases and their duals. As in §1, let the scaling function ϕ generate the spaces V_j, $j \in \mathbb{Z}$, in the sense of (1.2). Since $V_0 \subset V_1$, there exists a sequence $\{p_k\} \in \ell^\infty$ such that

$$\phi(x) = \sum_k p_k \phi(2x - k), \qquad x \in \mathbb{R}. \tag{2.1}$$

Moreover, since ϕ is assumed to be a compactly supported continuous function, the sequence $\{p_k\}_{k\in\mathbb{Z}}$ has exponential decay (see [7]), so that the two-scale symbol

$$P(z) = P_\phi(z) := \frac{1}{2}\sum_k p_k z^k \tag{2.2}$$

is an analytic function in a neighborhood of the unit circle $|z| = 1$. Let $\Phi(z)$ be the symbol of $\{\phi(k)\}$ as defined before and assume that it does not vanish on the unit circle as in hypothesis (vi) in §1. We have the following.

LEMMA 2.1. *$P(z)$ and $\Phi(z)$ are governed by the identity*

$$\Phi(z^2) = P(z)\Phi(z) + P(-z)\Phi(-z), \qquad |z| = 1. \tag{2.3}$$

This result can be found in Rioul [9] and its proof is very similar to that of some similar results in our earlier work [2, 3]. For completeness, we include a short derivation here.

Proof.

$$\begin{aligned}\Phi(z^2) &= \sum_n \phi(n)z^{2n} = \sum_n\sum_k p_k\phi(2n-k)z^{2n}\\ &= \sum_k p_k z^k \sum_n \phi(2n-k)z^{2n-k}\\ &= \sum_k p_{2k}z^{2k}\sum_n \phi(2n)z^{2n} + \sum_k p_{2k+1}z^{2k+1}\sum_n \phi(2n-1)z^{2n-1}\\ &= [P(z)+P(-z)][\Phi(z)+\Phi(-z)]/2 + [P(z)-P(-z)][\Phi(z)-\Phi(-z)]/2\\ &= P(z)\Phi(z) + P(-z)\Phi(-z). \qquad \square\end{aligned}$$

Next, consider

$$Q(z) := \frac{1}{2}z\Phi(-z) = \frac{1}{2}\sum_k (-1)^{k-1}\phi(k-1)z^k \tag{2.4}$$

and let the interpolating wavelet ψ be defined as in (1.10). Then its Fourier transform is given by

$$\hat{\psi}(\omega) = Q(z)\hat{\phi}\left(\frac{\omega}{2}\right) = \frac{1}{2}z\Phi(-z)\hat{\phi}\left(\frac{\omega}{2}\right), \qquad z = e^{-i\frac{\omega}{2}}, \tag{2.5}$$

where, as usual, the Fourier transform $\hat{f}$ of a function f on $\mathbb{R}$ is defined by

$$\hat{f}(\omega) := \int_{\mathbb{R}} f(x)e^{-i\omega x}dx, \qquad \omega\in\mathbb{R}, \tag{2.6}$$

whenever it makes sense and/or distributionally otherwise. Of course from the two-scale equation (2.1), we have also

$$\hat{\phi}(\omega) = P(z)\hat{\phi}\left(\frac{\omega}{2}\right), \qquad z = e^{-i\frac{\omega}{2}}. \tag{2.7}$$

LEMMA 2.2. *Let $\Phi(z)$ satisfy* (vi) *in* §1; *that is,*

$$\Phi(z) := \sum_k \phi(k)z^k \neq 0, \quad |z| = 1. \tag{2.8}$$

Then there are positive constants $A \le B$ such that

$$A(\|\mathbf{c}\|_{\ell^\infty} + \|\mathbf{d}\|_{\ell^\infty}) \le \left\|\sum_k c_k\phi(\cdot-k) + \sum_k d_k\psi(\cdot-k)\right\|_\infty \le B(\|\mathbf{c}\|_{\ell^\infty} + \|\mathbf{d}\|_{\ell^\infty}) \tag{2.9}$$

for all $\mathbf{c} = \{c_k\}_{k\in\mathbb{Z}} \in \ell^\infty$ and $\mathbf{d} = \{d_k\}_{k\in\mathbb{Z}} \in \ell^\infty$.

Proof. Since the functions ϕ and ψ are compactly supported and continuous, the second inequality in (2.9) clearly holds for some constant B. To establish the first inequality in (2.9), we recall from a result of Jia and Micchelli [6] that it is sufficient to prove that $\{\hat{\phi}(\omega+2k\pi)\}$ and $\{\hat{\psi}(\omega+2k\pi)\}$ are linearly independent for all $\omega \in \mathbb{R}$. We first observe, by applying the Poisson summation formula

$$\sum_k \hat{\phi}(\omega+2k\pi) = \sum_n \phi(n)e^{-in\omega} \tag{2.10}$$

and condition (2.8), that

$$\sup_{k\in\mathbb{Z}} |\hat{\phi}(\omega+2k\pi)| > 0, \quad \omega \in \mathbb{R}. \tag{2.11}$$

Thus, combining this fact with the two-scale relations (2.5) and (2.7), we can easily conclude that the above statement is equivalent to

$$\Delta(z) := \begin{vmatrix} P(z) & Q(z) \\ P(-z) & Q(-z) \end{vmatrix} \neq 0 \quad \text{for all} \quad |z| = 1. \tag{2.12}$$

By (2.4), (2.3), Lemma 2.1, and condition (2.8), we have, for all $|z| = 1$,

$$\Delta(z) = \begin{vmatrix} P(z) & \frac{1}{2}z\Phi(-z) \\ P(-z) & -\frac{1}{2}z\Phi(z) \end{vmatrix} = -\frac{1}{2}z[P(z)\Phi(z) + P(-z)\Phi(-z)] = -\frac{1}{2}z\Phi(z^2) \neq 0.$$

This completes the proof of Lemma 2.2. □

In view of (2.9), the integer translates $\phi(\cdot - k)$ and $\psi(\cdot - k)$, $k \in \mathbb{Z}$, are said to be ℓ^∞-stable (see [6]). For simplicity, we will also say that ϕ and ψ are ℓ^∞-stable. An immediate consequence of (2.9) is that

$$A\|\mathbf{c}\|_{\ell^\infty} \leq \left\| \sum_k c_k\phi(\cdot - k) \right\|_\infty \leq B\|\mathbf{c}\|_{\ell^\infty}, \tag{2.13}$$

$$A\|\mathbf{c}\|_{\ell^\infty} \leq \left\| \sum_k c_k\psi(\cdot - k) \right\|_\infty \leq B\|\mathbf{c}\|_{\ell^\infty} \tag{2.14}$$

for all $\mathbf{c} = \{c_k\} \in \ell^\infty$, so that ϕ and ψ are ℓ^∞-stable themselves. It follows from (2.13) that the spaces V_j defined in (1.2) are closed subspaces of C_u, and hence, we can write

(2.15)

$$V_j = \text{clos}_{C_u} \text{ span}\{\phi(2^j \cdot -k)\colon\ k \in \mathbb{Z}\} := \text{clos}_{C_u} \left\{ \sum_k c_k\phi(2^j \cdot -k) \in C_u \right\}, \quad j \in \mathbb{Z},$$

where, somewhat differently from the usual definition, the span is not restricted to finite linear combinations. This leads to the following.

DEFINITION 2.3. *Let $X = X(\mathbb{R})$ be a normed linear space of functions over $\mathbb{R}$ and $\{S_j\}_{j\in\mathbb{Z}}$ a family of closed subspaces of X. If a function $f \in X$ exists such that*

$$S_j = \text{clos}_X \text{ span}\{f(2^j \cdot -k)\colon\ k \in \mathbb{Z}\}, \qquad j \in \mathbb{Z}, \tag{2.16}$$

then f is called a generator *that generates the subspaces S_j. Moreover, if there are constants $0 < A \leq B < \infty$ and $1 \leq p \leq \infty$ such that*

$$A\|\mathbf{c}\|_{\ell^p} \leq \left\| \sum_k c_k f(\cdot - k) \right\| \leq B\|\mathbf{c}\|_{\ell^p}, \quad \mathbf{c} = \{c_k\} \in \ell^p,$$

then we say that f is an ℓ^p-stable generator *of $S_j, j \in \mathbb{Z}$.*

Thus, in the sense of Definition 2.3, we see that, under condition (vi), or equivalently (2.8), ϕ is an ℓ^∞-stable generator of V_j. As to the fundamental function ϕ_L given in (1.4), it is clear from the interpolating property (1.3) and the property of exponential decay (because of the exponential decay property of $\{r_k\}$ in (1.5)), that ϕ_L is ℓ^∞-stable in the sense of (2.13) (with ϕ_L instead of ϕ there). Moreover, in view of (1.4) and the fact that

$$\phi(x) = \sum_k \phi(k)\phi_L(x-k), \tag{2.17}$$

we see that ϕ_L is also an ℓ^∞-stable generator that generates the closed subspaces V_j.

THEOREM 2.4. *Let* $\{V_j\}$ *be defined as in* (1.2) *and satisfy the conditions* (i) *and* (vi) *in* §1. *Also, let the subspaces* W_j *of* C_u *be defined as in* (1.7). *Then both* ψ *and* ψ_L, *defined in* (1.10) *and* (1.11), *respectively, are* ℓ^∞*-stable generators of* W_j, $j \in \mathbb{Z}$.

Proof. Since ϕ_L is ℓ^∞-stable, it follows that $\psi_L = \phi_L(2\cdot -1)$ is ℓ^∞-stable also. By (2.14), it is sufficient to prove that

$$W_j = \left\{\sum_k c_k\psi_L(2^j\cdot -k)\colon \{c_k\}\in\ell^\infty\right\} = \left\{\sum_k c_k\psi(2^j\cdot -k)\colon \{c_k\}\in\ell^\infty\right\} \tag{2.18}$$

holds for $j=0$. Let $g\in W_0$. Then there is an $f\in V_1$ such that $g(x) = f(x)-(P_0f)(x)$. Since P_1 is a projection from C_u to V_1, we can write

$$f(x) = (P_1f)(x) = \sum_k f\left(\frac{k}{2}\right)\phi_L(2x-k). \tag{2.19}$$

Also, since the fundamental function ϕ_L has the two-scale relation

$$\begin{aligned}\phi_L(x) &= \sum_k \phi_L\left(\frac{k}{2}\right)\phi_L(2x-k)\\ &= \phi_L(2x) + \sum_k \phi_L\left(k+\frac{1}{2}\right)\phi_L(2x-2k-1)\\ &= \phi_L(2x) + \sum_k \phi_L\left(k+\frac{1}{2}\right)\psi_L(x-k),\end{aligned} \tag{2.20}$$

we see from (2.19) and (2.20) that

$$\begin{aligned} g(x) &= \sum_k f\left(\frac{k}{2}\right)\phi_L(2x-k) - \sum_\ell f(\ell)\phi_L(x-\ell)\\ &= \sum_k f\left(\frac{k}{2}\right)\phi_L(2x-k)\\ &\quad -\sum_\ell f(\ell)\left[\phi_L(2x-2\ell)+\sum_k\phi_L\left(k+\frac{1}{2}\right)\psi_L(x-\ell-k)\right]\\ &= \sum_k f\left(k+\frac{1}{2}\right)\phi_L(2x-2k-1)\\ &\quad -\sum_\ell\sum_k f(\ell)\phi_L\left(k+\frac{1}{2}\right)\psi_L(x-\ell-k)\\ &= \sum_k f\left(k+\frac{1}{2}\right)\psi_L(x-k) - \sum_k\left[\sum_\ell f(\ell)\phi_L\left(k-\ell+\frac{1}{2}\right)\right]\psi_L(x-k)\\ &= \sum_k (f-P_0f)\left(k+\frac{1}{2}\right)\psi_L(x-k).\end{aligned} \tag{2.21}$$

The sequence $\{c_k\}$, defined by

$$c_k := (f - P_0 f)\left(k + \frac{1}{2}\right) = f\left(k + \frac{1}{2}\right) - \sum_{\ell} f(\ell)\phi_L\left(k - \ell + \frac{1}{2}\right), \tag{2.22}$$

is clearly in ℓ^∞. Hence, by (2.21), we have

$$W_0 \subset \left\{\sum_k c_k \psi_L(\cdot - k) \colon\ \{c_k\} \in \ell^\infty\right\}. \tag{2.23}$$

On the other hand, for any $g(x) = \sum_k c_k \psi_L(x - k)$ with $\{c_k\} \in \ell^\infty$, we have $g \in V_1$ and $g(\ell) = 0$ (since $\psi_L(\ell) = \phi_L(2\ell - 1) = 0$) for all $\ell \in \mathbb{Z}$, so that $g = g - P_0 g \in W_0$. This proves that

$$W_0 \supset \left\{\sum_k c_k \psi_L(\cdot - k) \colon\ \{c_k\} \in \ell^\infty\right\}. \tag{2.24}$$

Combining (2.23) and (2.24), we have the first equality in (2.18) for $j = 0$. To establish the second equality in (2.18) for $j = 0$, we only need to represent ψ and ψ_L as (possibly infinite) linear combinations of integer translates of each other with coefficient sequences in ℓ^1. For this purpose, we consider

$$\sum_j v_j z^{2j} := \Phi(z)\Phi(-z). \tag{2.25}$$

Then by (1.5) and (2.25), we have

$$\begin{aligned}\sum_n \left(\sum_j v_j r_{n-2j}\right) z^n &= \sum_j v_j z^{2j} \sum_n r_{n-2j} z^{n-2j} \\ &= \sum_j v_j z^{2j} \sum_k r_k z^k = \Phi(z)\Phi(-z)[\Phi(z)]^{-1} \\ &= \Phi(-z) = \sum_n (-1)^n \phi(n) z^n,\end{aligned}$$

so that

$$\sum_j v_j r_{n-2j-1} = (-1)^{n-1}\phi(n-1), \qquad n \in \mathbb{Z}. \tag{2.26}$$

Hence, it follows from (1.11), (1.4), (2.26), and (1.10) that

$$\begin{aligned}\sum_j v_j \psi_L(x - j) &= \sum_j v_j \phi_L(2x - 2j - 1) \\ &= \sum_j v_j \sum_k r_k \phi(2x - 2j - 1 - k) \\ &= \sum_n \left(\sum_j v_j r_{n-2j-1}\right) \phi(2x - n) \\ &= \sum_n (-1)^{n-1} \phi(n-1) \phi(2x - n) = \psi(x).\end{aligned} \tag{2.27}$$

On the other hand, it also follows from (2.27) that

$$\psi_L(x) = \sum_j w_j \psi(x-j), \tag{2.28}$$

with

$$\sum_j w_j z^{2j} := \left(\sum_j v_j z^{2j}\right)^{-1} = [\Phi(z)\Phi(-z)]^{-1}. \tag{2.29}$$

This proves our assertion and therefore completes the proof of Theorem 2.4. □

We turn to the discussion of duals. By using the classical Dirac delta function (distribution) δ, namely,

$$\langle f, \delta \rangle = f(0), \quad f \in C_u, \tag{2.30}$$

we can express the fundamental interpolating property of ϕ_L in (1.3) as

$$\langle \phi_L(\cdot + k), \delta \rangle = \phi_L(k) = \delta_{k,0}, \qquad k \in \mathbb{Z}. \tag{2.31}$$

In view of (2.31), we will say that ϕ_L and δ are duals to each other. In the following, we give a precise notion of dual functionals and describe what we mean by dilations, translations, and convergence of functionals.

DEFINITION 2.5. *Let $X = X(\mathbb{R})$ be a normed linear space of functions over $\mathbb{R}$, and $X^* = X^*(\mathbb{R})$ be its dual space, consisting of all the continuous linear functionals on X.*

(i) *For an element $f^* \in X^*$, its a-*dilation *and b-*translation (shift) $f^*(a \cdot -b)$, *where $a, b \in \mathbb{R}$, $a \neq 0$, is defined by*

$$\langle f, f^*(a \cdot -b) \rangle := \frac{1}{a}\left\langle f\left(\frac{\cdot + b}{a}\right), f^* \right\rangle, \quad f \in X. \tag{2.32}$$

(ii) *$f \in X$ and $f^* \in X^*$ are said to be* dual to each other *if*

$$\langle f(\cdot + k), f^* \rangle = \langle f, f^*(\cdot - k) \rangle = \delta_{k,0} \quad \text{for all} \quad k \in \mathbb{Z}. \tag{2.33}$$

Thus, f^* will also be called a dual of *f.*

(iii) *Let $\{\tilde{f}_k\}$ be a sequence in X^*. We say that the series $\sum_k \tilde{f}_k$ is* convergent *in X^* if $\sum_k \langle f, \tilde{f}_k \rangle$ is convergent for all $f \in X$ and its limit satisfies*

$$\left| \sum_k \langle f, \tilde{f}_k \rangle \right| \leq C\|f\|, \qquad f \in X,$$

for some positive constant C independent of f. Consequently, the series $\sum_k \tilde{f}_k$ can be considered as an element of X^* in the sense that

$$\left\langle f, \sum_k \tilde{f}_k \right\rangle := \sum_k \langle f, \tilde{f}_k \rangle, \quad f \in X. \tag{2.34}$$

Now, corresponding to the coefficient sequence $\{r_k\}$ of the Laurent expansion of $\Phi^{-1}(z)$ in (1.5), we consider the functional

$$\tilde{\phi} := \sum_k r_k \delta(\cdot + k). \tag{2.35}$$

Since $\{r_k\} \in \ell^1$, we see that $\tilde{\phi}$ is a continuous linear functional on C_u in the sense of Definition 2.5 (iii). It now follows from (1.4) that

$$\tag{2.36} \begin{aligned} \langle \phi, \tilde{\phi}(\cdot - j)\rangle &= \langle \phi(\cdot + j), \tilde{\phi}\rangle \\ &= \sum_k r_k \langle \phi(\cdot + j), \delta(\cdot + k)\rangle \\ &= \sum_k r_k \phi(j-k) = \phi_L(j) = \delta_{j,0}, \qquad j \in \mathbb{Z}, \end{aligned}$$

so that $\tilde{\phi}$ is a dual of ϕ.

Note that although a function $f \in X$ may have more than one dual in different subspaces of X^*, the dual f^* of f in the subspace generated by the integer translates of f^* is unique. Before we go into futher details, let us introduce the "dual subspaces"

$$\tag{2.37} \widetilde{V}_j := \left\{ \sum_j d_k \delta(2^j \cdot - k) \colon \ \{d_k\} \in \ell^1 \right\}, \qquad j \in \mathbb{Z}.$$

We will simply use the notation $\|\cdot\|$ for the functional norm for $\widetilde{V}_j$. We have the following result.

THEOREM 2.6. *The functional $\tilde{\phi}$ in* (2.35) *is in $\widetilde{V}_0$ and is the dual of the scaling function ϕ that generates $\{V_j\}$. Moreover, by setting*

$$\tag{2.38} \phi^0(x) := \sum_k |\phi(x-k)|, \quad \phi_L^0(x) := \sum_k |\phi_L(x-k)|,$$

then

$$\tag{2.39} \begin{cases} \|\phi^0\|_\infty^{-1}\|\mathbf{d}\|_{\ell^1} \le \left\| \sum_j d_j \tilde{\phi}(\cdot - j) \right\| \le \|\tilde{\phi}\| \|\mathbf{d}\|_{\ell^1}, \\ \|\phi_L^0\|_\infty^{-1}\|\mathbf{d}\|_{\ell^1} \le \left\| \sum_j d_j \delta(\cdot - j) \right\| \le \|\mathbf{d}\|_{\ell^1} \end{cases}$$

for all $\mathbf{d} = \{d_j\} \in \ell^1$. Furthermore,

$$\tag{2.40} \delta = \sum_j \phi(j) \tilde{\phi}(\cdot + j).$$

Consequently, both δ and $\tilde{\phi}$ are ℓ^1-stable generators of the dual spaces $\widetilde{V}_j$, $j \in \mathbb{Z}$.

Proof. The second inequalities in both sets of inequalities in (2.39) are obvious, since $\|\tilde{\phi}(\cdot - j)\| = \|\tilde{\phi}\|$ and $\|\delta(\cdot - j)\| = \|\delta\| = 1$. Now observe that

$$\left\langle \sum_k c_k \phi(\cdot - k), \sum_j d_j \tilde{\phi}(\cdot - j) \right\rangle = \sum_{j,k} d_j c_k \delta_{j,k} = \sum_j c_j d_j, \qquad \{c_k\} \in \ell^\infty, \ \{d_j\} \in \ell^1.$$

Thus, for all $\mathbf{c} = \{c_k\} \in \ell^\infty$ with $\|\mathbf{c}\|_{\ell^\infty} = 1$, we have, from (2.13),

$$\begin{aligned} \left| \sum_j d_j c_j \right| &\le \left\| \sum_j d_j \tilde{\phi}(\cdot - j) \right\| \left\| \sum_k c_k \phi(\cdot - k) \right\|_\infty \\ &\le \left\| \sum_j d_j \tilde{\phi}(\cdot - j) \right\| B \|\mathbf{c}\|_{\ell^\infty} = B \left\| \sum_j d_j \tilde{\phi}(\cdot - j) \right\|, \end{aligned}$$

where B is the absolute constant given in (2.13). Of course, it is clear that the constant B can be chosen as $B = \|\phi^0\|_\infty$. This yields

$$\begin{aligned}\left\|\sum_j d_j\tilde{\phi}(\cdot - j)\right\| &\geq B^{-1} \sup_{\|\mathbf{c}\|_{\ell^\infty}=1} \left|\sum_j d_j c_j\right| \\ &= B^{-1}\sum_j |d_j| = \|\phi^0\|_\infty^{-1}\|\mathbf{d}\|_{\ell^1}.\end{aligned}$$

This establishes the first inequality in the first of the two sets of inequalities in (2.39). The proof of the first inequality in the second set in (2.39) is similar. From (2.39) and (2.37), we see that δ is indeed an ℓ^1-stable generator of $\widetilde{V}_j$. Moreover, from (2.35) and (1.5), we also have (2.40), so that $\tilde{\phi}$ generates $\widetilde{V}_j$ as well. □

It is clear that δ satisfies the two-scale relation

$$\delta = 2\delta(2\cdot), \tag{2.41}$$

so that

$$\widetilde{V}_j \subset \widetilde{V}_{j+1}, \quad j \in \mathbb{Z}. \tag{2.42}$$

In addition, it follows from (2.18) and (2.37) that

$$W_j \perp \widetilde{V}_j, \qquad j \in \mathbb{Z}, \tag{2.43}$$

in the sense that

$$\langle g, \tilde{f}\rangle = 0 \quad \text{for all} \quad g \in W_j, \quad \tilde{f} \in V_j, \tag{2.44}$$

since

$$\langle \psi_L(\cdot + k), \delta\rangle = \psi_L(k) = \phi_L(2k-1) = 0, \quad k \in \mathbb{Z}. \tag{2.45}$$

The "orthogonality" property (2.43) is one of the main reasons that the dual spaces $\widetilde{V}_j$ are worth investigating. Next, following Cohen, Daubechies, and Feauveau [4], we will look for a subspace $\widetilde{W}_j$ of $\widetilde{V}_{j+1}$ that satisfies

$$\widetilde{V}_{j+1} = \widetilde{V}_j \dot{+} \widetilde{W}_j, \qquad j \in \mathbb{Z}, \tag{2.46}$$

$$V_j \perp \widetilde{W}_j, \qquad j \in \mathbb{Z}, \tag{2.47}$$

so that the generator of $\widetilde{W}_j$ is a dual of the generator of W_j. For this purpose, set

$$\tilde{\psi}_L := 2\sum_k (-1)^{k-1}\phi_L\left(\frac{1-k}{2}\right)\delta(2\cdot - k). \tag{2.48}$$

Then we have

$$\begin{aligned}\langle \psi_L, \tilde{\psi}_L(\cdot - j)\rangle &= \langle \psi_L(\cdot + j), \tilde{\psi}_L\rangle \\ &= \sum_k (-1)^{k-1}\phi_L\left(\frac{1-k}{2}\right) 2\langle \phi_L(2\cdot + 2j - 1), \delta(2\cdot - k)\rangle \\ &= \sum_k (-1)^{k-1}\phi_L\left(\frac{1-k}{2}\right)\phi_L(k + 2j - 1) \\ &= \sum_k (-1)^{k-1}\phi_L\left(\frac{1-k}{2}\right)\delta_{k+2j-1,0} \\ &= \phi_L(j) = \delta_{j,0}, \qquad j \in \mathbb{Z},\end{aligned} \tag{2.49}$$

so that $\tilde{\psi}_L$ is indeed a dual of ψ_L. Furthermore, observe that

$$\begin{aligned}\langle \phi_L, \tilde{\psi}_L(\cdot - j)\rangle &= \langle \phi_L(\cdot + j), \tilde{\psi}_L\rangle \\ &= \sum_k (-1)^{k-1} \phi_L\left(\frac{1-k}{2}\right) 2\langle \phi_L(\cdot + j), \delta(2\cdot -k)\rangle \\ &= \sum_k (-1)^{k-1} \phi_L\left(\frac{1-k}{2}\right) \phi_L\left(\frac{k}{2} + j\right) \\ &= \sum_\ell (-1)^{\ell} \phi_L\left(\frac{\ell}{2} + j\right) \phi_L\left(\frac{1-\ell}{2}\right) \\ &= -\langle \phi_L, \tilde{\psi}_L(\cdot - j)\rangle, \qquad j \in \mathbb{Z},\end{aligned}$$

so that

$$\langle \phi_L, \tilde{\psi}_L(\cdot - j)\rangle = 0, \qquad j \in \mathbb{Z}. \tag{2.50}$$

Hence, by setting

$$\widetilde{W}_j := \left\{ \sum_k d_k \tilde{\psi}_L(2^j \cdot -k) \colon \ \{d_k\} \in \ell^1 \right\}, \qquad j \in \mathbb{Z}, \tag{2.51}$$

we see that (2.47) holds. Later, in §3, we will see that (2.46) is also valid. Since $\tilde{\psi}_L$ is a dual of ψ_L and ψ_L is ℓ^∞-stable, it is clear from the proof of Theorem 2.6 that $\tilde{\psi}_L$ is an ℓ^1-stable generator of $\widetilde{W}_j$.

We now give another ℓ^1-stable generator of $\widetilde{W}_j$, which is a dual of ψ.

THEOREM 2.7. *Let*

$$S(z) = \sum_k s_k z^k := 2P(z)/\Phi(z^2), \qquad |z| = 1, \tag{2.52}$$

where $P(z)$ and $\Phi(z)$ are given in (2.2) *and* (2.8)*, respectively. Also, let ψ be defined as in* (1.10)*. Then the dual of ψ is given by*

$$\tilde{\psi} := 2\sum_k (-1)^{k-1} s_{1-k} \tilde{\phi}(2\cdot -k), \tag{2.53}$$

where $\tilde{\phi}$ is defined in (2.35)*. Furthermore, both $\tilde{\psi}_L$ and $\tilde{\psi}$ are ℓ^1-stable generators of $\widetilde{W}_j$, $j \in \mathbb{Z}$.*

Proof. Since ϕ and $\tilde{\phi}$ are dual to each other, we have from (1.10) and (2.53) that

$$\begin{aligned}&\langle \psi(\cdot + j), \tilde{\psi}\rangle \\ &= \left\langle \sum_\ell (-1)^{\ell-1} \phi(\ell - 1)\phi(2\cdot + 2j - \ell), 2\sum_k (-1)^{k-1} s_{1-k} \tilde{\phi}(2\cdot -k)\right\rangle \\ &= \sum_\ell \sum_k (-1)^{\ell+k} s_{1-k} \phi(\ell - 1)\langle \phi(\cdot - \ell + 2j), \tilde{\phi}(\cdot - k)\rangle \\ &= \sum_\ell \sum_k (-1)^{\ell+k} s_{1-k} \phi(\ell - 1)\delta_{\ell - 2j, k} \\ &= \sum_k s_{1-k}\phi(2j + k - 1) = \sum_k s_k \phi(2j - k), \qquad j \in \mathbb{Z}.\end{aligned} \tag{2.54}$$

On the other hand, by (2.52) and Lemma 2.1, we also have

$$\begin{aligned}(2.55)\qquad &\sum_j \left(\sum_k s_k \phi(2j-k)\right) z^{2j}\\ &= \sum_k s_k z^k \sum_j \phi(2j-k) z^{2j-k}\\ &= \sum_k s_{2k} z^{2k} \sum_j \phi(2j) z^{2j} + \sum_k s_{2k+1} z^{2k+1} \sum_j \phi(2j-1) z^{2j-1}\\ &= \frac{1}{4}[S(z)+S(-z)][\Phi(z)+\Phi(-z)] + \frac{1}{4}[S(z)-S(-z)][\Phi(z)-\Phi(-z)]\\ &= \frac{1}{2}S(z)\Phi(z) + \frac{1}{2}S(-z)\Phi(-z)\\ &= [P(z)\Phi(z)+P(-z)\Phi(-z)]/\Phi(z^2) = 1, \quad |z|=1.\end{aligned}$$

Hence, it follows from (2.54) and (2.55) that

$$(2.56)\qquad \langle \psi(\cdot + j), \widetilde{\psi}\rangle = \delta_{j,0}, \qquad j \in \mathbb{Z},$$

so that $\widetilde{\psi}$ is a dual of ψ.

We have already shown that $\widetilde{\psi}_L$ generates $\widetilde{W}_j$, $j \in \mathbb{Z}$ (see the discussion following (2.51)). To see that $\widetilde{\psi}$ also generates $\widetilde{W}_j$, it is sufficient to show that

$$(2.57)\qquad \widetilde{\psi}_L = \sum_\ell v_\ell \widetilde{\psi}(\cdot + \ell),$$

where $\{v_\ell\}$ is given by (2.25). Now, by (2.53) and (2.35), we have

$$\begin{aligned}(2.58)\qquad \sum_\ell v_\ell \widetilde{\psi}(\cdot+\ell) &= \sum_\ell v_\ell \left(2\sum_k (-1)^{k-1} s_{1-k} \widetilde{\phi}(2\cdot+2\ell-k)\right)\\ &= 2\sum_\ell \sum_k (-1)^{k-1} s_{1-k} v_\ell \sum_j r_j \delta(2\cdot+2\ell-k+j)\\ &= 2\sum_n (-1)^{n-1}\left(\sum_j \sum_\ell (-1)^j r_j v_\ell s_{1-j-n-2\ell}\right)\delta(2\cdot-n).\end{aligned}$$

Consequently, by (2.25) and (2.52), we obtain

$$\begin{aligned}(2.59)\qquad \sum_k \left(\sum_\ell v_\ell s_{k-2\ell}\right) z^k &= \sum_\ell v_\ell z^{2\ell} \sum_k s_{k-2\ell} z^{k-2\ell}\\ &= \sum_\ell v_\ell z^{2\ell} \sum_k s_k z^k = \Phi(z)\Phi(-z)2P(z)/\Phi(z^2)\\ &=: B(z) = \sum_k b_k z^k,\end{aligned}$$

and moreover, we have, by (1.5),

$$\begin{aligned}(2.60)\qquad \sum_m \left(\sum_j (-1)^j r_j b_{m-j}\right) z^m &= \sum_j (-1)^j r_j z^j \sum_m b_{m-j} z^{m-j}\\ &= [\Phi(-z)]^{-1} B(z) = 2\Phi(z)P(z)/\Phi(z^2).\end{aligned}$$

On the other hand, by (1.4) and (2.1), we also have

$$\phi_L\left(\frac{m}{2}\right) = \sum_k r_k \phi\left(\frac{m}{2} - k\right) = \sum_k r_k \sum_\ell p_\ell \phi(m - 2k - \ell),$$

so that

$$\begin{aligned}\sum_m \phi_L\left(\frac{m}{2}\right) z^m &= \sum_m \sum_k \sum_\ell r_k p_\ell \phi(m-2k-\ell) z^m \\ &= \sum_k r_k z^{2k} \sum_\ell p_\ell z^\ell \sum_m \phi(m-2k-\ell) z^{m-2k-\ell} \\ &= [\Phi(z^2)]^{-1} 2P(z)\Phi(z).\end{aligned} \tag{2.61}$$

Combining (2.61), (2.60), and (2.59), we arrive at

$$\begin{aligned}\phi_L\left(\frac{m}{2}\right) &= \sum_j (-1)^j r_j b_{m-j} \\ &= \sum_j \sum_\ell (-1)^j r_j v_\ell s_{m-j-2\ell}.\end{aligned} \tag{2.62}$$

Hence, substituting (2.62) into (2.58) (with $m = 1 - n$) and applying (2.48), we obtain (2.57). As in the proof of Theorem 2.6, the ℓ^1-stability of $\tilde\psi$ follows from the ℓ^∞-stability of its dual ψ (see Theorem 2.4). This completes the proof of Theorem 2.7. □

We end this section by pointing out that the functional

$$\tilde\eta := 2\sum_k (-1)^{k-1} \phi\left(\frac{1-k}{2}\right) \delta(2\cdot - k) \tag{2.63}$$

is also an ℓ^1-stable generator of $\widetilde{W}_j$. The advantage of $\tilde\eta$ over $\tilde\psi_L$ and $\tilde\psi$ is that it has compact support whenever ϕ is a compactly supported scaling function. The relation between $\tilde\eta$ and $\tilde\psi_L$ is given by

$$\tilde\eta = \sum_\ell \phi(\ell) \tilde\psi_L(\cdot + \ell). \tag{2.64}$$

Similarly, the dual of $\tilde\eta$ in W_0 is given by

$$\eta := \sum_k r_k \psi_L(\cdot - k), \tag{2.65}$$

and its ℓ^1-stability follows from the ℓ^∞-stability of η. The proof of these facts is similar to and even easier than the previous ones. Hence, it is safe to omit the details here.

3. Decompositions, multiresolution analyses, and the functional wavelet transform.

We will now complete our proof of the dual (orthogonal) decomposition (2.46)–(2.47) by establishing the decomposition relation (2.46). This and other related results are summarized in the following theorem.

THEOREM 3.1. *Let V_j, W_j be defined as in* (1.2), (1.7) *and $\widetilde{V}_j$, $\widetilde{W}_j$ as in* (2.37), (2.51). *Suppose that conditions* (i) *and* (vi) *stated in* §1 *hold. Then*

$$V_{j+1} = V_j \dot{+} W_j, \qquad j \in \mathbb{Z} \tag{3.1}$$

$$\widetilde{V}_{j+1} = \widetilde{V}_j \dot{+} \widetilde{W}_j, \qquad j \in \mathbb{Z}; \tag{3.2}$$

$$W_j \perp \widetilde{V}_j, \quad V_j \perp \widetilde{W}_j, \qquad j \in \mathbb{Z}; \quad \textit{and} \tag{3.3}$$

$$W_j \perp \widetilde{W}_k, \quad j \neq k, \quad j, k \in \mathbb{Z}. \tag{3.4}$$

Proof. Assertion (3.1) is the result (1.9) in §1, and (3.3) is the summary of (2.43) and (2.47)). Hence, it remains to establish (3.2) and (3.4). First, we see from (2.48) and (2.51) that $\widetilde{W}_j \subset \widetilde{V}_{j+1}$. So, by applying (2.42), we have

$$\widetilde{V}_{j+1} \supset \widetilde{V}_j + \widetilde{W}_j, \qquad j \in \mathbb{Z}. \tag{3.5}$$

On the other hand, it follows from (2.41) and (2.48) that

$$\begin{aligned}\tilde{\psi}_L &= 2\sum_k \phi_L(-k)\delta(2\cdot -2k-1) - 2\sum_k \phi_L\left(\frac{1}{2}-k\right)\delta(2\cdot -2k)\\ &= 2\delta(2\cdot -1) - \sum_k \phi_L\left(\frac{1}{2}-k\right)\delta(\cdot - k).\end{aligned} \tag{3.6}$$

This yields

$$\delta(2\cdot -1) = \frac{1}{2}\sum_k \phi_L\left(\frac{1}{2}-k\right)\delta(\cdot - k) + \frac{1}{2}\tilde{\psi}_L. \tag{3.7}$$

From (2.41) and (3.7), we see that $\widetilde{V}_{j+1} \subset \widetilde{V}_j + \widetilde{W}_j$. This, together with (3.5), gives

$$\widetilde{V}_{j+1} = \widetilde{V}_j + \widetilde{W}_j, \qquad j \in \mathbb{Z}. \tag{3.8}$$

To establish

$$\widetilde{V}_j \cap \widetilde{W}_j = \{0\}, \qquad j \in \mathbb{Z}, \tag{3.9}$$

we let $\tilde{f} \in \widetilde{V}_j \cap \widetilde{W}_j$ and write both

$$\tilde{f} = \sum_\ell c_\ell \delta(2^j \cdot -\ell), \quad \{c_\ell\} \in \ell^1, \tag{3.10}$$

and

$$\begin{aligned}\tilde{f} &= \sum_\ell d_\ell \tilde{\psi}_L(2^j \cdot -\ell)\\ &= 2\sum_\ell d_\ell \delta(2^{j+1} \cdot -2\ell - 1) - \sum_\ell d_\ell \sum_k \phi_L\left(\frac{1}{2}-k\right)\delta(2^j \cdot -\ell - k)\\ &= 2\sum_\ell d_\ell \delta(2^{j+1} \cdot -2\ell - 1) - \sum_\ell\left(\sum_k d_{\ell-k}\phi_L\left(\frac{1}{2}-k\right)\right)\delta(2^j \cdot -\ell)\end{aligned} \tag{3.11}$$

for some $\{d_\ell\} \in \ell^1$, where the second equality in (3.11) follows from (3.6). By (2.41), (3.10), and (3.11), we obtain

$$\sum_\ell b_\ell \delta(2^{j+1} \cdot -\ell) = 0, \tag{3.12}$$

where $b_{2\ell+1} = d_\ell$ and $b_{2\ell} = -(c_\ell + \sum_k \phi_L(\frac{1}{2}-k)d_{\ell-k})$, $\ell \in \mathbb{Z}$. Clearly $\{b_\ell\} \in \ell^1$, so that the ℓ^1-stability of δ (cf. (2.39)) and (3.12) together imply that $b_\ell = 0$, $\ell \in \mathbb{Z}$. This turns out to be $d_\ell = b_{2\ell+1} = 0$ and $c_\ell = -(b_{2\ell} + \sum_k \phi_L(\frac{1}{2}-k)d_{\ell-k}) = 0$, $\ell \in \mathbb{Z}$. Hence $\tilde{f} = 0$ and (3.9) is established. Combining (3.8) and (3.9) gives (3.2).

Finally, to prove (3.4), we may assume, without loss of generality, that $j < k$. Thus,

$$W_j \subset V_{j+1} \subset \cdots \subset V_k \perp \widetilde{W}_k, \tag{3.13}$$

where the last orthogonality relation comes from (3.3), so that (3.4) holds. This completes the proof of Theorem 3.1. □

In the notion of interpolating multiresolution analysis $\{V_j\}$ of C_u as described in §1, six conditions are imposed on $\{V_j\}$. In the following theorem, we will see that these conditions are not independent.

THEOREM 3.2. *Let the spaces* V_j *be defined as in* (1.2). *Suppose that conditions* (i) *and* (vi) *described in* §1 *hold. Then conditions* (ii)–(v) *there also hold.*

Proof. It is clear that (iv) and (v) are consequences of the definition (1.2). It only remains to prove (ii) and (iii).

Since we have assumed that the generator ϕ of V_j is a compactly supported continuous function, it is obvious that each element of V_j is a uniformly continuous bounded function, so that

$$V_j \subset C_u, \qquad j \in \mathbb{Z}. \tag{3.14}$$

For each $f \in C_u$, let $P_j f$ be given as in (1.6). Then $P_j f \in V_j$, and as proven in [5], we have

$$\begin{aligned} \|f - P_j f\|_\infty &\le \omega(f; 2^{-j}) + \omega(P_j f; 2^{-j}) \\ &\le (1 + C)\omega(f; 2^{-j}), \end{aligned} \tag{3.15}$$

where $C := \|\phi_L^0\|_\infty$ is a constant with ϕ_L^0 already introduced in (2.38). Here, the standard notation

$$\omega(f; t) := \sup_{|h| \le t} \|f(\cdot + h) - f(\cdot)\|_\infty \tag{3.16}$$

of the uniform modulus of continuity is used. Since f is uniformly continuous, it follows from (3.15) that

$$\|f - P_j f\|_\infty \to 0 \quad \text{as} \quad j \to \infty. \tag{3.17}$$

This together with (3.14) yields condition (ii). To establish (iii), let $f \in \bigcap_{j \in \mathbb{Z}} V_j$. Then we can write

$$f = \sum_k c_k^j \phi(2^j \cdot - k), \quad \{c_k^j\} \in \ell^\infty, \qquad j \in \mathbb{Z}.$$

Since ϕ is ℓ^∞-stable, we have $\|f\|_\infty \ge A\|\mathbf{c}^j\|_{\ell^\infty}$, where $A > 0$ is the absolute constant in (2.13). Thus, for any $x, y \in \mathbb{R}$, we obtain

$$\begin{aligned} |f(x) - f(y)| &\le \|\mathbf{c}^j\|_{\ell^\infty} \sum_k |\phi(2^j x - k) - \phi(2^j y - k)| \\ &\le A^{-1}\|f\|_\infty \sum_k |\phi(2^j x - k) - \phi(2^j y - k)| \to 0 \end{aligned}$$

as $j \to -\infty$. That is, $f(x) = f(y)$ for all $x, y \in \mathbb{R}$. This yields $\bigcap_{j \in \mathbb{Z}} V_j \subset \mathbb{C}$. On the other hand, since ϕ satisfies the two-scale relation (2.1) with ℓ^∞-stability, we have

$$\sum_k \phi(x - k) = \hat{\phi}(0) \neq 0, \qquad x \in \mathbb{R} \tag{3.18}$$

(cf. [6]), so that $\bigcap_{j\in\mathbb{Z}} V_j$ indeed consists only of the constant functions. This establishes (iii) and completes the proof of Theorem 3.2. □

It now follows from Theorems 3.1 and 3.2 that

$$C_u = V_j \dot{+} \dot{\sum_{\ell\geq j}} W_\ell = \mathbb{C} + \dot{\sum_{\ell\in\mathbb{Z}}} W_\ell, \tag{3.19}$$

where the $\dot{\sum}$ denotes direct summation. Of course, we must be careful in making sure that the statements in (3.19) actually make sense, since direct summation is usually an algebraic concept. We will return to this after Theorem 3.5 is established.

For the cardinal interpolation projection operators P_j in (1.6), we can write, on one hand,

$$(P_j f)(x) = \sum_k 2^j \langle f, \delta(2^j \cdot - k)\rangle \phi_L(2^j x - k), \tag{3.20}$$

and by applying (1.4), (1.6), and (2.35), we can also write, on the other hand,

$$\begin{aligned}(P_j f)(x) &= \sum_k f(2^{-j}k) \sum_\ell r_\ell \phi(2^j x - k - \ell) \\ &= \sum_k \left(\sum_\ell r_\ell f(2^{-j}(k-\ell))\right) \phi(2^j x - k) \\ &= \sum_k \langle f(2^{-j}(\cdot + k)), \tilde{\phi}\rangle \phi(2^j x - k) \\ &= \sum_k 2^j \langle f, \tilde{\phi}(2^j \cdot - k)\rangle \phi(2^j x - k).\end{aligned} \tag{3.21}$$

We define another family of projection operators Δ_j from C_u to W_j as follows:

$$(\Delta_j f)(x) := \sum_k 2^j \langle f, \tilde{\psi}(2^j \cdot - k)\rangle \psi(2^j x - k), \tag{3.22}$$

where the dual $\tilde{\psi}$ of ψ is given in (2.53). From (2.27) and (2.57), similar to the deduction of (3.21), we also have

$$(\Delta_j f) = \sum_k 2^j \langle f, \tilde{\psi}_L(2^j \cdot - k)\rangle \psi_L(2^j x - k). \tag{3.23}$$

Hence, for each of the projection operators P_j and Δ_j, we have two different representations. These two operators are related as follows.

LEMMA 3.3. *The projection operators P_j and Δ_j satsify the relation*

$$P_{j+1} = P_j + \Delta_j, \qquad j \in \mathbb{Z}. \tag{3.24}$$

Proof. It is sufficient to prove (3.24) for $j = 0$ and we will apply (3.23) and (3.20) (i.e., (1.6)) to establish this assertion. To this end, we need to use the two-scale equation (2.20) for the fundamental function ϕ_L, which has the equivalent form

$$\phi_L(2x) = \phi_L(x) - \sum_\ell \phi_L\left(\ell + \frac{1}{2}\right) \phi_L(2x - 2\ell - 1). \tag{3.25}$$

Thus, for all $f \in C_u(\mathbb{R})$, we have

$$
\begin{aligned}
(3.26)\qquad (P_1 f)(x) &= \sum_k f\left(\frac{k}{2}\right)\phi_L(2x-k)\\
&= \sum_k f(k)\phi_L(2x-2k) + \sum_k f\left(k+\frac{1}{2}\right)\phi_L(2x-2k-1)\\
&= \sum_k f(k)[\phi_L(x-k) - \sum_\ell \phi_L\left(\ell+\frac{1}{2}\right)\phi_L(2x-2k-2\ell-1)]\\
&\quad + \sum_k f\left(k+\frac{1}{2}\right)\phi_L(2x-2k-1)\\
&= \sum_k f(k)\phi_L(x-k) + \sum_k \left[f\left(k+\frac{1}{2}\right)\right.\\
&\quad \left. - \sum_\ell f(k-\ell)\phi_L\left(\ell+\frac{1}{2}\right)\right]\phi_L(2x-2k-1)\\
&= (P_0 f)(x)\\
&\quad + \sum_k \left[\sum_\ell (-1)^{\ell-1} f\left(k+\frac{\ell}{2}\right)\phi_L\left(\frac{1-\ell}{2}\right)\right]\phi_L(2x-2k-1).
\end{aligned}
$$

On the other hand, from (3.23), (2.48), and (1.11), we also have
(3.27)

$$
\begin{aligned}
(\Delta_0 f)(x) &= \sum_k \langle f, \tilde{\psi}_L(\cdot - k)\rangle \psi_L(x-k)\\
&= \sum_k 2\sum_\ell (-1)^{\ell-1}\phi_L\left(\frac{1-\ell}{2}\right)\langle f, \delta(2\cdot -2k-\ell)\rangle \phi_L(2x-2k-1)\\
&= \sum_k \left[\sum_\ell (-1)^{\ell-1}\phi_L\left(\frac{1-\ell}{2}\right) f\left(k+\frac{\ell}{2}\right)\right]\phi_L(2x-2k-1),
\end{aligned}
$$

so that combining this fact with (3.26) yields

$$
(P_1 f)(x) = (P_0 f)(x) + (\Delta_0 f)(x), \qquad x \in \mathbb{R}.
$$

This proves (3.24) for $j = 0$. □

LEMMA 3.4. *The dual pair* ψ, $\tilde{\psi}$ *in* (1.10) *and* (2.53) *satisfies*

$$
(3.28)\qquad \langle \psi(2^j \cdot -k), \tilde{\psi}(2^{j'} \cdot -k')\rangle = 2^{-j}\delta_{j,j'}\delta_{k,k'}, \qquad j, j', k, k' \in \mathbb{Z},
$$

and

$$
(3.29)\qquad \langle 1, \tilde{\psi}(2^j \cdot -k)\rangle = 0, \qquad j, k \in \mathbb{Z},
$$

where 1 *is viewed as a constant function. Similar results also hold for the dual pair* ψ_L, $\tilde{\psi}_L$ *in* (1.11) *and* (2.48).

Proof. By Theorem 2.7, we see that for $j' = j$,

$$
\langle \psi(2^j \cdot -k), \tilde{\psi}(2^j \cdot -k')\rangle = 2^{-j}\langle \psi(\cdot - k), \tilde{\psi}(\cdot - k')\rangle = 2^{-j}\delta_{k,k'}.
$$

For different j and j', since $\psi(2^j \cdot -k) \in W_j$, $\widetilde{\psi}(2^{j'} \cdot -k) \in \widetilde{W}_{j'}$, we see that (3.4) implies

$$\langle \psi(2^j \cdot -k), \widetilde{\psi}(2^{j'} \cdot -k') \rangle = 0.$$

This proves (3.28). To establish (3.29), we have, from (3.18),

$$1 = [\hat{\phi}(0)]^{-1} \sum_k \phi(2^j x - k) \in V_j,$$

so that (3.29) is a consequence of (3.3) and the fact that $\widetilde{\psi}(2^j \cdot -k) \in \widetilde{W}_j$. This establishes Lemma 3.4. □

One of the main results in this section is the following.

THEOREM 3.5. *Every $f \in C_u$ has a unique and pointwise convergent decomposition*

$$f(x) = f(0) + \sum_{j \in \mathbb{Z}} g_j(x), \qquad x \in \mathbb{R}, \tag{3.30}$$

where $g_j = \Delta_j f \in W_j$ with the projection operators Δ_j given as in (3.22) *or* (3.23). *Furthermore, for each $m \in \mathbb{Z}$ and $f \in C_u$, the series*

$$f = P_m f + \sum_{j \geq m} \Delta_j f \tag{3.31}$$

is unformly convergent in C_u (under the sup norm).

Proof. Given $\varepsilon > 0$, by (3.15) there exists an $N \in \mathbb{Z}$ such that

$$\|f - P_n f\|_\infty < \varepsilon \quad \text{for all} \quad n > N.$$

Now, for any $m \in \mathbb{Z}$ and integer $n \geq \max\{m, N\}$, since we have, by Lemma 3.3, $\sum_{j=m}^n \Delta_j = P_{n+1} - P_m$, it follows that

$$\left\| f - P_m f - \sum_{j=m}^{n} \Delta_j f \right\|_\infty = \|f - P_{n+1} f\|_\infty < \varepsilon.$$

This proves that the convergence in (3.31) is uniform. Also, according to (3.31), pointwise convergence in (3.30) is equivalent to pointwise convergence of

$$\lim_{m \to -\infty} (P_m f)(x) = f(0), \qquad x \in \mathbb{R}. \tag{3.32}$$

Hence, since we have, by (1.6),

$$(P_m f)(x) = f(0)\phi_L(2^m x) + \sum_{k \neq 0} f(2^{-m}k)\phi_L(2^m x - k)$$

and therefore

$$\begin{aligned} |(P_m f)(x) - f(0)| &\leq |f(0)||1 - \phi_L(2^m x)| + \sum_{k \neq 0} |f(2^{-m}k)||\phi_L(2^m x - k)| \\ &\leq \|f\|_\infty (|1 - \phi_L(2^m x)| + \sum_{k \neq 0} |\phi_L(2^m x - k)|), \end{aligned}$$

it follows that

$$\overline{\lim_{m \to -\infty}} |(P_m f)(x) - f(0)| \leq \|f\|_\infty (|1 - \phi_L(0)| + \sum_{k \neq 0} |\phi_L(k)|) = 0,$$

which yields (3.32) and hence (3.30). To establish the uniqueness of the decomposition, we see that if $c + \sum_{j\in\mathbb{Z}} g_j$ is another (pointwise convergent) representation for f with $c \in \mathbb{R}$, $g_j \in W_j$, then since every $g_j \in W_j$, $j \in \mathbb{Z}$, satisfies $g_j(0) = 0$ (see (2.45)), we immediately have $c = f(0)$. Moreover, according to Lemma 3.4 and (3.22) or (3.23), one can easily verify that $\Delta_j f = g_j$, $j \in \mathbb{Z}$. This completes the proof of Theorem 3.5. □

Remark. We need to point out that the convergence in (3.30) is not uniform in general. As a simple example, consider $f = \phi_L$. Then it is clear that $(P_j f)(x) = \phi_L(2^j x)$ for $j < 0$, and hence,

$$\|P_j f - f(0)\|_\infty = \|\phi_L(2^j\cdot) - \phi_L(0)\|_\infty = \|\phi_L - 1\|_\infty \geq 1.$$

This shows that the convergence in (3.32), and hence (3.30), is not uniform.

As a corollary of Theorem 3.5, we have from (3.3) that

$$f(x) = f(0) + \sum_{\ell<j}(\Delta_\ell f)(x), \qquad x \in \mathbb{R}, \tag{3.33}$$

for every $f \in V_j$. The reason is that for $\ell \geq j$, we have $f \in V_j \subset V_\ell \perp \widetilde{W}_\ell$, so that $\langle f, \widetilde{\psi}(2^\ell\cdot -k)\rangle = 0$, $k \in \mathbb{Z}$, and $(\Delta_\ell f)(x) = \sum_k 2^\ell\langle f, \widetilde{\psi}(2^\ell\cdot -k)\rangle\psi(2^\ell x - k) = 0$. As a consequence of (3.30) and (3.33), we see that the infinite direct sums in (3.19) as well as the direct sum

$$V_j = \mathbb{C}\dot{+}\dot{\sum_{\ell<j}} W_j \tag{3.34}$$

do make sense (with pointwise convergence).

From (3.22) and (3.30), we can write, for each $f \in C_u$,

$$f(x) = f(0) + \sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} d_k^j\psi(2^j x - k), \qquad x \in \mathbb{R}, \tag{3.35}$$

where

$$d_k^j = 2^j\langle f, \widetilde{\psi}(2^j\cdot -k)\rangle, \qquad j,k \in \mathbb{Z}. \tag{3.36}$$

We are now ready to introduce the notion of "functional wavelet transforms" (FnWT)

$$(W_{\widetilde{\psi}}f)(b,a) := \frac{1}{a}\left\langle f, \widetilde{\psi}\left(\frac{\cdot - b}{a}\right)\right\rangle, \qquad a \neq 0, \quad b \in \mathbb{R}, \tag{3.37}$$

as mentioned in (1.17). Recall that the usual integral wavelet transform (IWT) is an inner product in L^2, which is the dual space of itself. Hence, we may consider the FnWT as a generalization of the IWT. Recall also that the values of the IWT at the dyadic points $\left(\frac{k}{2^j}, \frac{1}{2^j}\right)$ can be computed via certain pyramid decomposition-reconstruction algorithms under the structure of a multiresolution analysis of L^2. The same is true for the situation of the FnWT, though the FnWT can be computed directly without integration (e.g., by the point-value functional and its linear combinations). This will be done in the next section. We end this section by deriving the property of vanishing moments of $\widetilde{\psi}$ as mentioned in §1.

THEOREM 3.6. *Let $\widetilde{\psi}$ be defined as in* (2.53). *If the scaling function ϕ (locally) reproduces all polynomials of degree up to $m-1$, where m is a positive integer, in the sense that*

$$x^j = \sum_k b_k^j \phi(x-k), \qquad x \in \mathbb{R}, \quad j = 0, 1, \ldots, m-1, \tag{3.38}$$

for some constants b_k^j with polynomial growth as $k \to \infty$, then

$$\langle p, \widetilde{\psi} \rangle = 0, \quad \textit{for all} \quad p \in \pi_{m-1}. \tag{3.39}$$

Similar results also hold for $\widetilde{\psi}_L$ in (2.48).

Remark. Since a nontrivial polynomial $p(x)$ does not belong to the space C_u, the meaning of $\langle p, \widetilde{\psi} \rangle$ must be clarified. However, this will be clear from the following proof of the theorem. In addition, since dilation and translation are invariant for polynomials, (1.19) follows from (3.39).

Proof. By (3.29), we have seen that (3.39) is valid for $p(x) \equiv 1$. For the general case, the proof of (3.39) is different from that of (3.29), and we first need to make sure that $\langle p, \widetilde{\psi} \rangle$ is well defined for any polynomial p.

From (2.53) and (2.35), we can write

$$\begin{aligned} \widetilde{\psi} &= 2\sum_k (-1)^{k-1} s_{1-k} \sum_\ell r_\ell \delta(2\cdot -k+\ell) \\ &= 2\sum_n (-1)^{n-1} \left(\sum_\ell (-1)^\ell r_\ell s_{1-n-\ell} \right) \delta(2\cdot -n) \\ &= 2\sum_n t_n \delta(2\cdot -n), \end{aligned} \tag{3.40}$$

where

$$t_n = (-1)^{n-1} \sum_\ell (-1)^\ell r_\ell s_{1-n-\ell}, \qquad n \in \mathbb{Z}. \tag{3.41}$$

Since $\{r_k\}$ and $\{s_k\}$ in (1.5) and (2.52), respectively, have exponential decay, so does $\{t_n\}$ in (3.41). Thus, the series

$$2\sum_n t_n \langle p, \delta(2\cdot -n) \rangle = \sum_n t_n p\left(\frac{n}{2}\right) \tag{3.42}$$

is absolutely convergent for any polynomial p, and according to (3.40) and (3.42), we may define

$$\langle p, \widetilde{\psi} \rangle := \sum_n t_n p\left(\frac{n}{2}\right), \qquad p \in \pi_{m-1}. \tag{3.43}$$

This is similar to but not completely the same as Definition 2.5 (iii).

Now, for $p(x) = x^j$, $0 \le j \le m-1$, from (3.38) and (3.43) we obtain

$$\langle p, \widetilde{\psi} \rangle = \sum_n \sum_k b_k^j t_n \phi\left(\frac{n}{2} - k\right). \tag{3.44}$$

Since b_k^j is of polynomial growth, t_n is of exponential decay, and ϕ is compactly supported, the double series in (3.44) is absolutely convergent, and consequently, we can interchange the order of summations in (3.44) to yield

$$\langle p,\widetilde{\psi}\rangle=\sum_k b_k^j\sum_n t_n\phi\left(\frac{n}{2}-k\right)=\sum_k b_k^j\langle\phi(\cdot-k),\widetilde{\psi}\rangle, \tag{3.45}$$

where the last equality in (3.45) follows from (3.40). From (3.45), it is clear that $\langle p,\widetilde{\psi}\rangle=0$ for all $p\in\pi_{m-1}$, since $\langle\phi(\cdot-k),\widetilde{\psi}\rangle=0$, for all $k\in\mathbb{Z}$ (see (2.47) and Theorem 2.7). This proves (3.39). From (2.57), we also have $\langle p,\widetilde{\psi_L}\rangle=0$ for all $p\in\pi_{m-1}$. This completes the proof of Theorem 3.6. □

4. Algorithms and examples. To derive the decomposition-reconstruction algorithms for the interpolating wavelets ψ_L and ψ, we need an explicit decomposition formula for $\phi_L(2x-\ell)$ and $\phi(2x-\ell)$, $\ell\in\mathbb{Z}$.

Taking $f(x)=\phi_L(2x-\ell)$ as in the proof of Theorem 2.4 and applying (2.21), (1.6), and (1.3), we immediately obtain

$$\begin{aligned}\phi_L(2x-\ell)&=f(x)=(P_0f)(x)+(f-P_0f)(x)\\&=\sum_k\delta_{2k,\ell}\phi_L(x-k)\\&\quad+\sum_k\left[\delta_{2k+1,\ell}-\sum_m\delta_{2m,\ell}\phi\left(k-m+\frac{1}{2}\right)\right]\psi_L(x-k).\end{aligned}\tag{4.1}$$

However, in practical applications, we may wish to use ϕ instead of ϕ_L. For this, we need the following result.

THEOREM 4.1. *Let the sequences $\{g_m\}\in\ell^1$ and $\{h_n\}\in\ell^1$ be determined by the following equations:*

$$\Phi(z)/\Phi(z^2)=\sum_n g_nz^n, \tag{4.2}$$

$$2z^{-1}P(-z)/\Phi(z^2)=\sum_n h_nz^n, \tag{4.3}$$

where $P(z)$ and $\Phi(z)$ are given in (2.2) and (2.8), respectively. Then

$$\phi(2x-\ell)=\sum_k[g_{2k-\ell}\phi(x-k)+h_{2k-\ell}\psi(x-k)],\qquad \ell\in\mathbb{Z}. \tag{4.4}$$

Proof. Since $\phi(2x-\ell)\in V_1=V_0\dot{+}W_0$ (cf. (3.1)), we can write

$$\phi(2x-\ell)=\sum_k g_k^\ell\phi(x-k)+\sum_k h_k^\ell\psi(x-k) \tag{4.5}$$

for some $\{g_k^\ell\}\in\ell^\infty$ and $\{h_k^\ell\}\in\ell^\infty$, $\ell\in\mathbb{Z}$. By applying the duals $\tilde{\phi}$ and $\widetilde{\psi}$ on (4.5) and observing that (3.3) holds, we have

$$\begin{cases}g_k^\ell=\langle\phi(2\cdot-\ell),\tilde{\phi}(\cdot-k)\rangle=\langle\phi(2\cdot+2k-\ell),\tilde{\phi}\rangle,\\ h_k^\ell=\langle\phi(2\cdot-\ell),\widetilde{\psi}(\cdot-k)\rangle=\langle\phi(2\cdot+2k-\ell),\widetilde{\psi}\rangle,\end{cases}\tag{4.6}$$

so that by setting

$$(4.7)\qquad \begin{cases} g_n := \langle \phi(2\cdot +n), \tilde{\phi}\rangle, \\ h_n := \langle \phi(2\cdot +n), \tilde{\psi}\rangle, \quad n \in \mathbb{Z}, \end{cases}$$

we see that (4.4) is a consequence of (4.5)–(4.7). It remains to prove that the sequences $\{g_n\}$ and $\{h_n\}$ defined in (4.7) satisfy equations (4.2) and (4.3).

By (2.35) and (4.7), we have

$$\begin{aligned} g_n &= \left\langle \phi(2\cdot +n), \sum_k r_k \delta(\cdot + k)\right\rangle \\ &= \sum_k r_k \langle \phi(2\cdot +n), \delta(\cdot + k)\rangle \\ &= \sum_k r_k \phi(n-2k), \end{aligned}$$

and consequently,

$$\begin{aligned} \sum_n g_n z^n &= \sum_n \sum_k r_k \phi(n-2k) z^n \\ &= \sum_k r_k z^{2k} \sum_n \phi(n-2k) z^{n-2k} \\ &= \sum_k r_k z^{2k} \sum_n \phi(n) z^n \\ &= \Phi(z)/\Phi(z^2), \end{aligned}$$

where the last equality follows from (2.8) and (1.5). This establishes (4.2). By (2.53), (4.7), and the duality between ϕ and $\tilde{\phi}$, we have

$$\begin{aligned} h_n &= \left\langle \phi(2\cdot +n), 2\sum_k (-1)^{k-1} s_{1-k} \tilde{\phi}(2\cdot -k)\right\rangle \\ &= \sum_k (-1)^{k-1} s_{1-k} \langle \phi(\cdot + n), \tilde{\phi}(\cdot - k)\rangle \\ &= \sum_k (-1)^{k-1} s_{1-k} \delta_{k,-n} = (-1)^{n+1} s_{n+1}, \end{aligned}$$

so that

$$\begin{aligned} \sum_n h_n z^n &= \sum_n (-1)^{n+1} s_{n+1} z^n \\ &= z^{-1} \sum_n s_n (-z)^n = 2z^{-1} P(-z)/\Phi(z^2), \end{aligned}$$

where the last equality follows from (2.52). This proves (2.43). From (4.2) and (4.3) it is clear that $\{g_n\} \in \ell^1$ and $\{h_n\} \in \ell^1$. This completes the proof of Theorem 4.1. □

With the decomposition formula (4.4) in hand, the algorithms for our interpolating wavelet ψ are almost the same as that in the L^2 setting. For any $f \in C_u$, let $f_N = P_N f$ be the projection of f to V_N for a fixed $N \in \mathbb{Z}$. More generally, we don't have to use the cardinal interpolant $P_N f$. We may consider V_N as the "sample space" and f_N the "data" (or measurement) of f on V_N. Since

$$V_N = W_{N-1} \dotplus V_{N-1} = \cdots \\ = W_{N-1} \dotplus \cdots \dotplus W_{N-M} \dotplus V_{N-M} \tag{4.8}$$

for any positive integer M, f_N has the unique decomposition

$$f_N(x) = g_{N-1}(x) + f_{N-1}(x) = \cdots \\ = g_{N-1}(x) + g_{N-2}(x) + \cdots + g_{N-M}(x) + f_{N-M}(x), \tag{4.9}$$

where

$$f_{j+1} = g_j + f_j, \quad f_j \in V_j, \quad g_j \in W_j, \quad j = N-M, \ldots, N-1. \tag{4.10}$$

Let us write

$$f_j(x) = \sum_k c_k^j \phi(2^j x - k) \in V_j, \quad \mathbf{c}^j = \{c_k^j\}, \tag{4.11}$$

and

$$g_j(x) = \sum_k d_k^j \psi(2^j x - k), \quad \mathbf{d}^j = \{d_k^j\}. \tag{4.12}$$

Then the decomposition in (4.9) is uniquely determined by the sequences $\mathbf{c}^j$ and $\mathbf{d}^j$ in (4.11) and (4.12), respectively. As shown in the last section, we have

$$d_k^j = 2^j \langle g_j, \tilde{\psi}(2^j \cdot - k) \rangle \\ = 2^j \langle f_N, \tilde{\psi}(2^j \cdot - k) \rangle = (W_{\tilde{\psi}} f_N) \left(\frac{k}{2^j}, \frac{1}{2^j} \right), \tag{4.13}$$

where $W_{\tilde{\psi}} f$ is the FnWT of f defined in (3.36). By using the dual property between ϕ and $\tilde{\phi}$ and noticing that $\tilde{\phi}(2^j \cdot - k) \in \tilde{V}_j \subset \tilde{V}_\ell \perp W_\ell$, $j \leq \ell \leq N$, we have

$$c_k^j = 2^j \langle f_j, \tilde{\phi}(2^j \cdot - k) \rangle \\ = 2^j \langle f_N, \tilde{\phi}(2^j \cdot - k) \rangle = (W_{\tilde{\phi}} f_N) \left(\frac{k}{2^j}, \frac{1}{2^j} \right). \tag{4.14}$$

4.1. Decomposition algorithm. Let the sequences $\{g_n\}$ and $\{h_n\}$ be as in Theorem 4.1. We have the following formula to produce $\mathbf{c}^j = \{c_k^j\}$ and $\mathbf{d}^j = \{d_k^j\}$, $j = N-M, \ldots, N-1$, from $\mathbf{c}^N = \{c_k^N\}$:

$$\begin{cases} c_k^{j-1} = \sum_\ell g_{2k-\ell} c_\ell^j, \\ d_k^{j-1} = \sum_\ell h_{2k-\ell} c_\ell^j, \quad k \in \mathbb{Z} \quad \text{and} \quad j = N, N-1, \ldots, N-M+1. \end{cases} \tag{4.15}$$

$$\begin{tikzcd} & \mathbf{d}^{N-1} & & \mathbf{d}^{N-2} & & & & \mathbf{d}^{N-M} \\ \mathbf{c}^N \arrow[r] \arrow[ur] & \mathbf{c}^{N-1} \arrow[rr] \arrow[urr] & & \mathbf{c}^{N-2} \arrow[r] \arrow[ur] & \cdots \arrow[rrr] \arrow[urrr] & & & \mathbf{c}^{N-M}. \end{tikzcd}$$

4.2. Reconstruction algorithm. Let $\{p_n\}$ be the two-scale sequence in (2.1). Then we have the following formula to produce $\mathbf{c}^j$, $N-M+1 \leq j \leq N$, from $\mathbf{c}^{N-M}$ and $\mathbf{d}^j$, $N-M \leq j \leq N-1$:

$$c_k^j = \sum_\ell [p_{k-2\ell} c_\ell^{j-1} + (-1)^{k-1}\phi(k-2\ell-1) d_\ell^{j-1}], \tag{4.16}$$

$k \in \mathbb{Z}$, $j = N-M+1, \ldots, N$.

$$\begin{tikzcd}
\mathbf{d}^{N-M} \arrow[dr] & \mathbf{d}^{N-M+1} \arrow[dr] & {} \arrow[dr] & \mathbf{d}^{N-1} \arrow[dr] & \\
\mathbf{c}^{N-M} \arrow[r] & \mathbf{c}^{N-M+1} \arrow[r] & \cdots \arrow[r] & \mathbf{c}^{N-1} \arrow[r] & \mathbf{c}^{N}.
\end{tikzcd}$$

The proof of the above decomposition-reconstruction algorithms is the same as the L^2 setting (see [1, pp. 158–159]).

Examples. (i) Let $\phi_1 \in L^2$ be a compactly supported scaling function which generates a (dyadic) multiresolution analysis of L^2. Then it is easy to see that the autocorrelation function

$$\phi(x) := (\phi_1(-\cdot) * \bar{\phi}_1)(x) = \int_{\mathbb{R}} \overline{\phi_1(x+y)}\phi_1(y)dy \tag{4.17}$$

of ϕ_1 generates the multiresolution analysis of C_u as introduced in this paper.

(ii) Consider the mth-order cardinal B-spline function N_m, defined by

$$\begin{cases} N_1(x) := \chi_{[0,1]}(x), \\ N_m(x) := (N_{m-1} * N_1)(x) = \displaystyle\int_0^1 N_{m-1}(x-t)dt, \quad m \geq 2. \end{cases} \tag{4.18}$$

Then N_m satisfies the two-scale relation

$$N_m(x) = \sum_{k=0}^{m} 2^{-m+1}\binom{m}{k} N_m(2x-k) \tag{4.19}$$

and has compact support $[0, m]$. It is well known that

$$\sum_k N_m\left(k + \frac{1+(-1)^{m-1}}{4}\right) z^k \neq 0, \quad \text{for all } |z| = 1 \tag{4.20}$$

(see [10]). In particular, we have

$$\sum_k N_{2m}(k) z^k \neq 0, \qquad |z| = 1, \quad m \geq 1. \tag{4.21}$$

Thus, for the scaling function $\phi = N_{2m}$, our interpolating wavelet ψ is given by

$$\psi_{2m}(x) := \sum_{k=2}^{2m} (-1)^{k-1} N_{2m}(k-1) N_{2m}(2x-k). \tag{4.22}$$

Note that the support of ψ is $[1, 2m]$, which is even smaller than the support of N_{2m}. The graphs of ψ_4 and ψ_6 are shown in Figures 1 and 2.

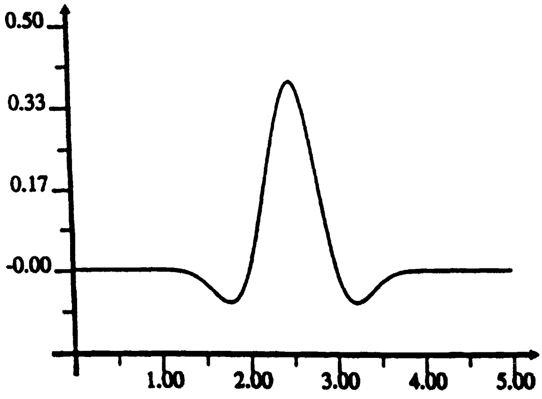


FIG. 1.

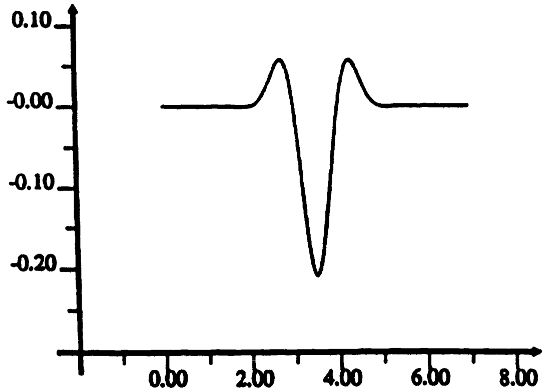


FIG. 2.

Acknowledgments. We are grateful to the most conscientious referee for pointing out an error in the statement of one of our results and other typos in the original manuscript.

REFERENCES

[1] C. K. CHUI, *An Introduction to Wavelets*, Academic Press, Boston, 1992.

[2] C. K. CHUI AND C. LI, *Non-orthogonal wavelet packets*, SIAM J. Math. Anal., 24 (1993), pp. 712–738.

[3] ———, *A general framework of multivariate wavelets with duals*, Applied and Computational Harmonic Analysis (ACHA), 1 (1994), pp. 368–390.

[4] A. COHEN, I. DAUBECHIES, AND J. C. FEAUVEAU, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.

[5] D. L. DONOHO, *Interpolating wavelet transform*, preprint, 1992.

[6] R. Q. JIA AND C. A. MICCHELLI, *Using the refinement equation for the construction of pre-wavelets* II: *Power of two*, in Curves and Surfaces, P. J. Laurent, A. Le Méhauté, and L.L. Schumaker, eds., Academic Press, Boston, 1991, pp. 209–246.

[7] ———, *Using the refinement equation for the construction of pre-wavelets V: extensibility of trigonometric polynomials*, Computing, 48 (1992), pp. 61–72.

[8] C. A. MICCHELLI, *Using the refinement equation for the construction of pre-wavelets*, Numer. Algorithms, 1 (1991), pp. 75–116.

[9] O. RIOUL, *Simple regularity criteria for subdivision schemes*, SIAM J. Math. Anal., 23 (1992), pp. 1544–1576.

[10] I. J. SCHOENBERG, *Cardinal Spline Interpolation*, CBMS–NSF Series in Applied Mathematics, Vol. 12, Society for Industrial and Applied Mathematics, Philadelphia, 1973.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 891–904, May 1996

015

STABILITY OF REFINABLE FUNCTIONS, MULTIRESOLUTION ANALYSIS, AND HAAR BASES*

DING-XUAN ZHOU†

Abstract. The stability of the integer translates of a univariate refinable function is characterized in terms of the mask sequence in the corresponding k-scale ($k \geq 2$) refinement equation. We show that the stability and refinement of some kinds of basis functions lead to a multiresolution analysis in $L^p(\mathbb{R}^s)(1 \leq p \leq \infty, s \in \mathbb{N})$ based on general lattices. As an application we determine explicitly all those multiresolution analyses in $L^2(\mathbb{R})$ associated with $(\mathbb{Z}, k)$ whose scaling functions are characteristic functions.

Key words. stability, wavelets, refinement equations, multiresolution analysis, Haar bases

AMS subject classifications. 41A30, 41A58, 42C05

1. Introduction.

Wavelet decompositions are based on basis functions that satisfy refinement equations. The stability of the integer translates of the basis function plays an essential role in the study of wavelets. Since the basis function can be determined by the mask of the corresponding refinement equation, it is natural to characterize the stability in terms of the mask sequence.

Let $s \in \mathbb{N}$, $1 \leq p \leq \infty$, $\phi \in L^p := L^p(\mathbb{R}^s)$; we say that the integer translates of ϕ are l^p-stable if there exist positive constants A_p and B_p such that for any $a \in l^p(\mathbb{Z}^s)$

$$A_p\|a\|_p \leq \left\| \sum_{\alpha \in \mathbb{Z}^s} a_\alpha \phi(\cdot - \alpha) \right\|_p \leq B_p\|a\|_p. \tag{1.1}$$

A locally integrable function ϕ is said to be refinable if it satisfies the refinement equation

$$\phi = \sum_{\alpha \in \mathbb{Z}^s} b_\alpha \phi(M \cdot -\alpha), \tag{1.2}$$

where $b \in l^\infty(\mathbb{Z}^s)$ is a complex-valued sequence called the mask of (1.2) and M is an $s \times s$ matrix called a scaling matrix such that $M\mathbb{Z}^s := \{M\alpha : \alpha \in \mathbb{Z}^s\} \subset \mathbb{Z}^s$ and $k := |\det M| \in \mathbb{N}$ with all the eigenvalues satisfying $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_s| > 1$.

Recently, Jia and Wang [7] gave a criterion for l^2-stability of the integer translates of a univariate compactly supported refinable distribution in terms of the compactly supported mask of the corresponding refinement equation. In this paper, in the cases of $s = 1$ and power of k with $k \geq 2$, we consider the stability of the integer translates of noncompactly supported refinable functions. For a function that decays exponentially fast in $L^2(\mathbb{R})$ we give a criterion for this l^2-stability in terms of the mask sequence of the refinement equation that it satisfies.

Once we have obtained the stability of the integer translates of a refinable function, we can construct a multiresolution analysis and then wavelet decompositions for different kinds of function spaces as well as applications. This procedure has been discussed by many authors; see [1, 3, 5–12]. Here we only mention a recent work of Jia and Micchelli [5].

*Received by the editors August 26, 1992; accepted for publication (in revised form) September 16, 1994. This research was supported by the Alexander von Humboldt Foundation of Germany.

†Institute of Mathematics, Academia Sinica, Beijing 100080, People's Republic of China. Current address: Department of Mathematical Sciences, University of Alberta, Edmonton, AB T6G 2G1, Canada.

For a measurable function ϕ on $\mathbb{R}^s$, set

$$\phi^0(x) = \sum_{\alpha\in\mathbb{Z}^s} |\phi(x-\alpha)|.$$

Then ϕ^0 is a 1-periodic function. Define

$$|\phi|_p = \|\phi^0\|_{L^p([0,1)^s)}.$$

For $1 \le p \le \infty$, let $\mathcal{L}^p(\mathbb{R}^s)$ be the Banach space of all measurable functions ϕ for which $\|\phi\|_{\mathcal{L}^p} = |\phi|_p < \infty$. These spaces have many interesting properties that can be found in [5]. By means of these properties, Jia and Micchelli proved that the stability and refinement of an $\mathcal{L}^p$ basis function are sufficient to generate a multiresolution analysis of power of two in $L^p(\mathbb{R}^s)$ for $1 \le p < \infty$.

In the second part of this paper we give similar results for $1 \le p \le \infty$ and general lattices.

Finally, we use the above results to determine explicitly all the multiresolution analyses in $L^2(\mathbb{R})$ associated with $(\mathbb{Z}, k)$ whose scaling functions are characteristic functions, which is the main purpose of this paper. Such a scaling function can be expressed as the characteristic function χ_Q of some set Q defined by

$$Q = \left\{ \sum_{j=1}^{\infty} k^{-j}\epsilon_j : \epsilon_j \in \{n_0, \dots, n_{k-1}\} \right\}, \tag{1.3}$$

where $\{n_0, \dots, n_{k-1}\}$ is a collection of representatives of distinct cosets of $\mathbb{Z}/(k\mathbb{Z})$, say, $n_j \equiv j \pmod{k}$ for $0 \le j \le k-1$. Gröchenig and Madych [4] gave some conditions on Q that are sufficient and necessary for χ_Q to generate a multiresolution analysis in $L^2(\mathbb{R})$. However, these conditions cannot be verified easily. Using our above results and Euler's theorem from number theory, we show in the last section that χ_Q can generate a multiresolution analysis in $L^2(\mathbb{R})$ if and only if the numbers $\{n_1 - n_0, n_2 - n_0, \dots, n_{k-1} - n_0\}$ are relatively prime. Moreover, for any collection of representatives $\{n_0, n_1, \dots, n_{k-1}\}$ the set Q defined by (1.3) has measure $(n_1 - n_0, n_2 - n_0, \dots, n_{k-1} - n_0) \in \mathbb{N}$. Here for a set $\{m_1, \dots, m_J\}$ of integer numbers we use $(m_1, \dots, m_J)$ to denote the greatest common divisor (g.c.d.).

For further convenience, we need some notation.

For $x = (x_1, \dots, x_s) \in \mathbb{R}^s$, we let $|x| = \sum_{j=1}^s |x_j|$ be its norm in $\mathbb{R}^s$. We denote $|E|$ as the measure of the set $E \subset \mathbb{R}^s$.

For a sequence $b \in \mathbb{Z}^s$, we denote $\tilde{b}$ as its symbol

$$\tilde{b}(z) = \sum_{\alpha\in\mathbb{Z}^s} b_\alpha z^\alpha, \qquad z \in \mathbb{C}^s. \tag{1.4}$$

Given a function $\phi \in \mathcal{L}^p(\mathbb{R}^s)$ and a sequence $a \in l^\infty(\mathbb{Z}^s)$, the semidiscrete convolution product $\phi *' a$ is, by definition, the sum $\sum_{\alpha\in\mathbb{Z}^s} \phi(\cdot - \alpha)a(\alpha)$.

The Fourier transform of a function $\phi \in L^1(\mathbb{R}^s)$ is given by

$$\hat{\phi}(\omega) = \int_{\mathbb{R}^s} \phi(x) e^{-ix\cdot\omega}\, dx,$$

where for $x = (x_1, \dots, x_s)$ and $\omega = (\omega_1, \dots, \omega_s) \in \mathbb{R}^s, x \cdot \omega = \sum_{j=1}^s x_j\omega_j$. This operation can be uniquely extended to $L^2(\mathbb{R}^s)$. For $\phi \in L^2(\mathbb{R}^s)$, set for $\omega \in \mathbb{R}^s$

$$\Pi_\phi(\omega) = \sum_{\alpha\in\mathbb{Z}^s} |\hat{\phi}(\omega + 2\pi\alpha)|^2. \tag{1.5}$$

With all the above notation we can now state our main results.

2. A criterion for stability. In the univariate case the refinement equation (1.2) becomes

$$\phi = \sum_{n\in\mathbb{Z}} b_n \phi(k\cdot -n) \tag{2.1}$$

with $2 \leq k \in \mathbb{N}$, $(b_n) \in l^\infty(\mathbb{Z})$.

We want to characterize the stability of the integer translates of an exponentially decaying solution to (2.1) in terms of the zero distribution of the symbol of the mask sequence.

For generality we denote E^2 as the space of all the functions f in $\mathcal{L}^2(\mathbb{R})$ that satisfy

$$\|f(\cdot+n)\|_{L^2[0,1)} \leq Cq^{|n|} \tag{2.2}$$

for some constants $C > 0$ and $0 < q < 1$ that depend only on f.

Then we can state our first main result as follows.

THEOREM 1. *Let $\phi \in E^2$ be a solution to (2.1) with $\hat{\phi}(0) \neq 0$. Then the integer translates of ϕ are l^2-stable if and only if the following conditions are satisfied.*

(1) $b \in l^1(\mathbb{Z})$.

(2) $\sum_{l=0}^{k-1} |\tilde{b}(e^{-i2l\pi/k}z)| > 0$ *for any z on the unit circle $T := \{z \in \mathbb{C} : |z| = 1\}$.*

(3) *For any $m \in \mathbb{N}$ and $z \in T$ satisfying $z^{k^m} = z \neq 1$, there exists an integer $d \geq 0$ such that*

$$\sum_{l=1}^{k-1} |\tilde{b}(e^{-i2l\pi/k} z^{k^d})| > 0. \tag{2.3}$$

Proof of Theorem 1. Necessity. Suppose that the integer translates of $\phi \in E^2 \subset \mathcal{L}^2(\mathbb{R})$ are l^2-stable. Then by [5, Thm. 3.3] there exists $g = \phi *' a$, $a \in l^1$, such that for $\alpha, \beta \in \mathbb{Z}$

$$\langle \phi(\cdot - \alpha), g(\cdot - \beta)\rangle = \delta_{\alpha,\beta}.$$

Note that $b \in l^\infty$ and $\phi, g \in \mathcal{L}^2(\mathbb{R})$. Using the dominated convergence theorem, we obtain from (2.1) that

$$b_n = \left\langle \phi\left(\frac{1}{k}\cdot\right), g(\cdot - n)\right\rangle.$$

Hence, condition (1) is necessary:

$$\|b\|_1 \leq \left|\phi\left(\frac{1}{k}\cdot\right)\right|_2 |g|_2 \leq k^2|\phi|_2|g|_2 < \infty.$$

Therefore, $\tilde{b}(z)$ is continuous for $z \in T$.

To see the necessity of condition (2), we take the Fourier transforms of both sides of (2.1) and obtain

$$\hat{\phi}(\omega) = \frac{1}{k}\tilde{b}(e^{-i\omega/k})\hat{\phi}\left(\frac{\omega}{k}\right). \tag{2.4}$$

Note that for $\phi \in \mathcal{L}^2(\mathbb{R})$, $\Pi_\phi(\omega)$ is continuous for $\omega \in \mathbb{R}$. We have

$$\Pi_\phi(\omega) = \frac{1}{k^2}\sum_{l=0}^{k-1}\left|\tilde{b}(e^{-i2l\pi/k}e^{-i\omega/k})\right|^2 \Pi_\phi\left(\frac{\omega+2l\pi}{k}\right).$$

It was proved in [5, Thm. 3.3] that for $\phi \in \mathcal{L}^2(\mathbb{R})$, the integer translates of ϕ are l^2-stable if and only if $\Pi_\phi(\omega) > 0$ for any $\omega \in \mathbb{R}$. Therefore, we must have

$$\sum_{l=0}^{k-1}\left|\tilde{b}(e^{-i2l\pi/k}e^{-i\omega/k})\right| > 0.$$

Hence condition (2) is also necessary.

Finally, we prove the necessity of condition (3). Suppose to the contrary that there exist $m \in \mathbb{N}$ and $z_0 \in T$ such that $z_0^{k^m} = z_0 \neq 1$ and for all integers $d \geq 0$,

$$\sum_{l=1}^{k-1}\left|\tilde{b}(e^{-i2l\pi/k}z_0^{k^d})\right| = 0. \tag{2.5}$$

Let $z_0 = e^{-i2n\pi/(k^m-1)} \neq 1$. Then $n/(k^m-1) \notin \mathbb{Z}$. We claim that for all $\alpha \in \mathbb{Z}$,

$$\hat{\phi}(2n\pi/(k^m-1)+2\alpha\pi) = 0, \tag{2.6}$$

which implies that the integer translates of ϕ are l^2-unstable.

To prove (2.6), we set $n+(k^m-1)\alpha = k^p q$, where $p \geq 0$ and q are integers and $k \not| |q|$ or $q = 0$. Since $n/(k^m-1) \notin \mathbb{Z}$, we must have $q \neq 0$.

By (2.4) we have

$$\begin{aligned}\hat{\phi}(2n\pi/(k^m-1)+2\alpha\pi) &= \hat{\phi}(2k^p q\pi/(k^m-1)) \\ &= \prod_{j=1}^{p+1}\left[\frac{1}{k}\tilde{b}(e^{-i2k^{p-j}q\pi/(k^m-1)})\right]\hat{\phi}(2q\pi/(k(k^m-1))).\end{aligned}$$

Under assumption (2.5) we now prove that

$$\tilde{b}(e^{-i2q\pi/(k(k^m-1))}) = 0, \tag{2.7}$$

which implies (2.6).

To this end, we choose $r = (k^{m(p+1)}-1)/(k^m-1) \in \mathbb{N}$ and set $-qr = uk+v$ with $u, v \in \mathbb{Z}$ and $0 \leq v < k$. Then we must have $v \neq 0$, since $(k,r) = 1$ and $k \not| |q|$.

Therefore, we have

$$\begin{aligned}\tilde{b}(e^{-i2q\pi/(k(k^m-1))}) &= \tilde{b}(e^{-i2q\pi[k^{m(p+1)}-r(k^m-1)]/[k(k^m-1)]}) \\ &= \tilde{b}(e^{i2qr\pi/k}e^{-i2\pi k^p q k^{(m-1)(p+1)}/(k^m-1)}) \\ &= \tilde{b}(e^{-i2v\pi/k}(e^{-i2\pi n/(k^m-1)})^{k^{(m-1)(p+1)}}) \\ &= \tilde{b}(e^{-i2v\pi/k}z_0^{k^{(m-1)(p+1)}}) \\ &= 0.\end{aligned}$$

Thus, the proof of the necessity is complete.

Sufficiency. We note that $\phi \in \mathcal{L}^2(\mathbb{R})$ satisfies

$$\sum_{n\in\mathbb{Z}} |\hat{\phi}(\omega + 2n\pi)|^2 = \sum_{n\in\mathbb{Z}} \langle \phi(\cdot + n), \phi \rangle e^{-in\omega}.$$

For this, see [5, Thm. 3.2].

Let

$$K(\phi) = \left\{ z = e^{-i\omega} \in T : \omega \in \mathbb{R}, \sum_{n\in\mathbb{Z}} |\hat{\phi}(\omega + 2n\pi)|^2 = 0 \right\}. \tag{2.8}$$

Since $\phi \in E^2$, the sequence $\{\langle \phi(\cdot + n), \phi\rangle\}_{n\in\mathbb{Z}}$ decays exponentially fast. Hence the function

$$\sum_{n\in\mathbb{Z}} \langle \phi(\cdot + n), \phi \rangle z^n$$

is analytic in a domain $\{z \subset \mathbb{C} : r_1 < |z| < r_2\}$ with $0 < r_1 < 1 < r_2$ and, therefore, has only finitely many zero points on T. That is, $K(\phi)$ is a finite set.

Suppose that the three conditions of Theorem 1 are satisfied. We show that $K(\phi)$ is empty, which implies that ϕ has l^2-stable integer translates.

We first prove the following statement.

If $z \in K(\phi)$, then there exists a positive integer m such that

$$z^{k^m} = z \neq 1. \tag{2.9}$$

Let $z_0 = e^{-i\zeta} \in K(\phi)$. Then $z_0 \neq 1$ since $\hat{\phi}(0) \neq 0$. From condition (1), we know that $\tilde{b}(z)$ is continuous on T. Thus, we can use (2.4) and obtain

$$\Pi_\phi(\zeta) = \frac{1}{k^2} \sum_{l=0}^{k-1} \left| \tilde{b}(e^{-i(\zeta+2l\pi)/k}) \right|^2 \Pi_\phi\left(\frac{\zeta}{k} + \frac{2l\pi}{k} \right) = 0.$$

From condition (2) we know that there exists $l \in \{0, 1, \ldots, k-1\}$ such that $\tilde{b}(e^{-i2l\pi/k} e^{-i\zeta/k}) \neq 0$. Hence,

$$\Pi_\phi\left(\frac{\zeta}{k} + \frac{2l\pi}{k} \right) = 0$$

and $e^{-i(\zeta+2l\pi)/k} \in K(\phi)$. This shows that there exists $z_1 \in K(\phi)$ such that $z_1^k = z_0$. Repeating this process, we can find a sequence $\{z_n\}_{n=0,1,\ldots}$ in $K(\phi)$ such that $z_n^k = z_{n-1}$ for $n \in \mathbb{N}$. But $K(\phi)$ is a finite set. We must have integers $0 \leq p < q$ such that $z_p = z_q$. Then we obtain

$$z_p = z_q^{k^{q-p}} = z_q$$

and

$$z_0 = z_q^{k^q} = z_q^{k^{2q-p}} = z_0^{k^{q-p}}.$$

Here $q - p \in \mathbb{N}$. Therefore, (2.9) holds and the statement is true.

We use this statement to prove the sufficiency. Suppose to the contrary that the integer translates of ϕ are l^2-unstable. Equivalently, $K(\phi)$ is not empty, say, $z \in K(\phi)$.

Then, by (2.9) there exists $m \in \mathbb{N}$ such that $z^{k^m} = z \neq 1$. Set $z = e^{-i2\pi k^p q/(k^m-1)}$, where $q, p \geq 0$ and $k \not| q$.

We show that for $z_0 = e^{-i2\pi q/(k^m-1)} \neq 1$ and any integers $d \geq 0, l = 1, 2, \ldots, k-1$,

$$\tilde{b}(e^{-i2l\pi/k} z_0^{k^d}) = 0, \tag{2.10}$$

which is a contradiction to condition (3). Hence the integer translates must be l^2-stable.

To prove (2.10) for d and l, we let $\omega_0 = 2\pi k^{p+m(d+1)} q/(k^m - 1)$. Then $e^{-i\omega_0} = z \in K(\phi)$, i.e., $\Pi_\phi(\omega_0) = 0$.

By (2.4), we have for $l = 0, 1, \ldots, k-1$,

$$\begin{aligned}\sum_{n\in\mathbb{Z}} &|\hat{\phi}(\omega_0 + 2\pi(kn + l))|^2 \\ &= \frac{1}{k^2}|\tilde{b}(e^{-i(\omega_0+2l\pi)/k})|^2 \Pi_\phi\left(\frac{\omega_0 + 2l\pi}{k}\right) = 0.\end{aligned}$$

Thus, either $\tilde{b}(e^{-i(\omega_0+2l\pi)/k})$ or $\Pi_\phi\left(\frac{\omega_0+2l\pi}{k}\right)$ must be zero.

On the other hand, we claim that for $l = 1, 2, \ldots, k-1$,

$$\Pi_\phi\left(\frac{\omega_0 + 2l\pi}{k}\right) \neq 0.$$

Since otherwise, $e^{-i(\omega_0+2l\pi)/k} \in K(\phi)$, by (2.9), there exists a positive integer α such that $(e^{-i(\omega_0+2l\pi)/k})^{k^\alpha - 1} = 1$. Hence,

$$\frac{1}{2\pi}(\omega_0 + 2l\pi)(k^\alpha - 1)/k = (k^{p+m(d+1)}q/(k^m - 1) + l)(k^\alpha - 1)/k \in \mathbb{Z}.$$

We observe that for any $n \in \mathbb{N}$, $(k^n - 1, k) = 1$. Therefore,

$$k^{p+m(d+1)-1}q(k^\alpha - 1) + l(k^\alpha - 1)(k^m - 1)/k \in \mathbb{Z}$$

and

$$l/k \in \mathbb{Z},$$

which is a contradiction.

We have proved our claim and obtain for $l = 1, 2, \ldots, k-1$,

$$\tilde{b}(e^{-i(\omega_0+2l\pi)/k}) = 0. \tag{2.11}$$

By condition (2) we also have from (2.11)

$$\tilde{b}(e^{-i\omega_0/k}) \neq 0.$$

Hence, $\Pi_\phi(\omega_0/k) = 0$.

Let $\omega_j = k^{-j}\omega_0 = 2\pi k^{p+m(d+1)-j}q/(k^m - 1)$. Our process for ω_0 can be repeated for ω_j with $j = 1, 2, \ldots, p + m(d+1) - 1$. In fact, for $1 \leq j \leq p + m(d+1) - 1$, we have

$$\tilde{b}(e^{-i(\omega_j+2\pi l)/k}) = 0, \qquad l = 1, 2, \ldots, k-1,$$

and

$$\Pi_\phi\left(\frac{\omega_j}{k}\right) = 0.$$

In particular, for $l = 1, 2, \ldots, k-1$,

$$\tilde{b}(e^{-i2\pi l/k} z_0^{k^d}) = \tilde{b}(e^{-i(\omega_{p+m(d+1)-d}+2l\pi)/k}) = 0.$$

Thus, we have proved (2.10) for any integer $d \geq 0$ and $l = 1, 2, \ldots, k-1$.

The proof of Theorem 1 is complete. □

Remark. If we replace the condition $\phi \in E^2$ by the assumption that $\phi \in \mathcal{L}^2(\mathbb{R})$ and $\Pi_\phi(\cdot)$ has only finitely many zeros on $[0, 2\pi)$, then Theorem 1 still holds. In fact, this observation is also true for multivariate cases, which we shall discuss elsewhere.

Remark. Even if a refinable function that has l^2-stable integer translates is of compact support, it may happen that the mask of the corresponding refinement equation does not have compact support. Let $k = 2$ and $\phi \in \mathcal{L}^2(\mathbb{R})$ be a compactly supported refinable function whose integer translates are l^2-stable. Suppose that ϕ satisfies the refinement equation

$$\phi\left(\frac{1}{2}\cdot\right) = \phi *' b$$

with the mask b of compact support. Then for a finitely supported sequence a, the function $\psi = \phi *' a$ has l^2-stable integer translates if and only if $\tilde{a}(z)$ has no zeros on T. In this case, ψ is also refinable. The mask of the corresponding refinement equation is of compact support if and only if $\tilde{a}(z^2)\tilde{b}(z)/\tilde{a}(z)$ is a Laurent polynomial. For example, we choose $\phi \in C(\mathbb{R})$ to be of compact support, to be refinable, and to have orthogonal integer translates. Then $\phi(\cdot) + c\phi(\cdot + 2n)$ with $n \in \mathbb{N}$ and $0 \neq c \notin T$ is a refinable compactly supported function and has l^2-stable integer translates. But the mask of the corresponding refinement equation is of no compact support.

By the method in the first part of the proof of Theorem 1 and some estimates in [5] we can easily obtain the following result.

THEOREM 2. *Suppose that $s \in \mathbb{N}$, $\{\phi_1, \ldots, \phi_n\} \subset \mathcal{L}^p(\mathbb{R}^s)$ have l^p-stable integer translates with $2 \leq p \leq \infty$ (for a definition, see* [5]*). Denote $S_p(\phi_1, \ldots, \phi_n) = \{\sum_{j=1}^n \phi_j *' a_j : a_1, \ldots, a_n \in l^p(\mathbb{Z}^s)\}$. Then the mapping*

$$L_{\phi_1,\ldots,\phi_n} : (l^p)^n \longrightarrow L^p(\mathbb{R}^s)$$

given by

$$L_{\phi_1,\ldots,\phi_n}(a_1, \ldots, a_n) = \sum_{j=1}^n \phi_j *' a_j$$

defines an isomorphism from $(l^1)^n$ onto the space $\mathcal{L}^p(\mathbb{R}^s) \cap S_p(\phi_1, \ldots, \phi_n)$ with the norm $|\ |_p$.

Combining Theorem 1 with [2], we can give a similar criterion for the orthogonality.

THEOREM 3. *Let $\phi \in E^2$ be a solution to* (2.1) *with $\hat{\phi}(0) = 1$. Then the integer translates of ϕ are orthogonal if and only if the mask b satisfies the conditions* (1) *and* (3) *in Theorem* 1 *and the following condition:*

$$\sum_{l=0}^{k-1} \left|\tilde{b}(e^{-i2l\pi/k} z)\right|^2 = k^2 \quad \textit{for all } z \in T.$$

By our method here and some ideas from [7] we can also give some criteria for the stability and linear independence of the integer translates of a refinable (associated with $(\mathbb{Z}, k)$) compactly supported distribution (for the definitions, see [7]).

Let $\{b_n\}_{n\in\mathbb{Z}}$ be a finitely supported sequence with $\sum_{n\in\mathbb{Z}} b_n = k, k \geq 2$; and let ϕ be the compactly supported distribution solution to the refinement equation (2.1) with $\hat{\phi}(0) = 1$. Then we have the following result.

THEOREM 4. *The integer translates of ϕ are stable if and only if conditions* (2) *and* (3) *in Theorem* 1 *are satisfied.*

THEOREM 5. *The integer translates of ϕ are linearly independent if and only if the symbol of the mask $\tilde{b}(z)$ satisfies condition* (3) *in Theorem* 1 *and the following condition:*

$$\sum_{l=0}^{k-1} \left|\tilde{b}(e^{-i2l\pi/k}z)\right| > 0 \quad \textit{for all } z \in \mathbb{C} \setminus \{0\}.$$

3. Multiresolution analysis in $L^p(\mathbb{R}^s)$. Most constructions of wavelet decompositions are based on multiresolution analyses.

We say that a sequence $\{V_j\}_{j\in\mathbb{Z}}$ of closed subspaces of $L^p(\mathbb{R}^s)$ $(1 \leq p \leq \infty, s \in \mathbb{N})$ forms a multiresolution analysis in $L^p(\mathbb{R}^s)$ with respect to a scaling matrix M as defined in §1 if it satisfies the following conditions.

(R1) $V_j \subset V_{j+1}$ for all $j \in \mathbb{Z}$.

(R2) $f \in V_0$ if and only if $f(\cdot - \alpha) \in V_0$ for all $\alpha \in \mathbb{Z}^s$.

(R3) $f \in V_j$ if and only if $f(M\cdot) \in V_{j+1}$ for all $j \in \mathbb{Z}$.

(R4) There exists an isomorphism from $l^p(\mathbb{Z}^s)$ onto V_0 that commutes with shift operators.

(R5) $\cap_{j\in\mathbb{Z}} V_j = \{0\}$ for $1 \leq p < \infty$;

$$\dim\left(\bigcap_{j\in\mathbb{Z}} V_j\right) < \infty \quad \text{for } p = \infty.$$

(R6) $\cup_{j\in\mathbb{Z}} V_j$ is dense in $(L^p(\mathbb{R}^s), \|\cdot\|_p)$ for $1 \leq p < \infty$ and in $(L^\infty(\mathbb{R}^s), \sigma(L^\infty, L^1))$ for $p = \infty$.

The concept of multiresolution analysis was first introduced by Mallat [10] and Meyer [11] for the case $p = 2$, which is of most interest. In this case, there is a function ϕ in V_0 whose integer translates form an unconditional basis of V_0. Such a function is called scaling function.

Jia and Micchelli [5] discussed the cases of $1 \leq p < \infty$ and power two in more detail. They proved that the refinement and stability of integer translates of certain basis function are sufficient to lead to a multiresolution analysis of $L^p(\mathbb{R}^s)$ for $1 \leq p < \infty$. For general scaling matrices and $p = 2$, see also [6, 9]. In this section we develop this theory to the case $p = \infty$. Our definition for this case is somewhat different from that of Jia and Micchelli.

In the next section we shall use the results for a general scaling matrix in the univariate case. Hence, we present our statements for the general lattices and for $1 \leq p \leq \infty$. We shall only give the detailed proof for the case $p = \infty$ and omit the proof for $1 \leq p < \infty$ because it can be obtained by the same methods as in [5, 6].

We first state our main result in this section and prove it later.

THEOREM 6. *Let $\phi \in \mathcal{L}^p(\mathbb{R}^s)$ for $1 \leq p < \infty$; $\phi, \phi^0 \in C(\mathbb{R}^s)$ for $p = \infty$; $s \in \mathbb{N}$; and M is a scaling matrix. Define $V_0 = S_p(\phi)$ and $V_j = \{f(M^j\cdot) : f \in V_0\}$. If ϕ is* *refinable with an l^1-mask and has l^p-stable integer translates, then $\{V_j\}_{j\in\mathbb{Z}}$ forms a multiresolution analysis in $L^p(\mathbb{R}^s)$.*

The proof of Theorem 6 is divided into a few parts. We first prove the property (R5).

THEOREM 7. *Let $\phi \in \mathcal{L}^p(\mathbb{R}^s)$ for $1 \le p < \infty$; $\phi, \phi^0 \in C(\mathbb{R}^s)$ for $p = \infty$; $V_0 = S_p(\phi)$; and $V_j = \{f(M^j\cdot) : f \in V_0\}$. If ϕ has l^p-stable integer translates, then $\cap_{j\in\mathbb{Z}} V_j = \{0\}$ for $1 \le p < \infty$ while $\cap_{j\in\mathbb{Z}} V_j$ is the set of constant functions in $\mathbb{R}^s$ for $p = \infty$.*

Proof of Theorem 7 for $p = \infty$. Let $f \in \cap_{j\in\mathbb{Z}} V_j$. Then $f(M^j\cdot) \in V_0$ for any $j \in \mathbb{Z}$. Hence there exists a sequence $a^{(j)} \in l^\infty$ such that

$$f(M^j x) = \phi *' a^{(j)}(x) = \sum_{\alpha\in\mathbb{Z}^s} a_\alpha^{(j)} \phi(x - \alpha)$$

and

$$\|a^{(j)}\|_\infty \le A_\infty^{-1} \|f(M^j\cdot)\|_\infty = A_\infty^{-1} \|f\|_\infty.$$

For any fixed $x \neq y \in \mathbb{R}^s$, we have

$$\begin{aligned} |f(x) - f(y)| &\le \|a^{(j)}\|_\infty \sum_{\alpha\in\mathbb{Z}^s} |\phi(M^{-j}x - \alpha) - \phi(M^{-j}y - \alpha)| \\ &\le A_\infty^{-1} \|f\|_\infty \sum_{\alpha\in\mathbb{Z}^s} |\phi(M^{-j}x - \alpha) - \phi(M^{-j}y - \alpha)|. \end{aligned}$$

We observe that $\phi, \phi^0 \in C(\mathbb{R}^s)$; hence the series

$$\sum_{\alpha\in\mathbb{Z}^s} |\phi(\omega - \alpha)|$$

is uniformly convergent for $\omega \in [-1, 1)^s$. We also note that $M^{-j}x, M^{-j}y \to 0$ as $j \to \infty$. Therefore,

$$|f(x) - f(y)| \to 0 \quad \text{as } j \to \infty.$$

Thus, we know that f is a constant function. The proof of Theorem 7 for $p = \infty$ is complete.

Remark. In the case $p = \infty$, the assumption $\phi, \phi^0 \in C(\mathbb{R}^s)$ in Theorem 7 cannot be replaced by $\phi \in \mathcal{L}^\infty(\mathbb{R}^s)$, which can be seen from the Haar basis.

Let ϕ be the characteristic function of the set $[0, 1)^s$. Then $\phi \in \mathcal{L}^\infty(\mathbb{R}^s)$ and has l^∞-stable integer translates while

$$\dim\left(\bigcap_{j\in\mathbb{Z}} V_j\right) = 2^s.$$

THEOREM 8. *If $\phi \in L^1(\mathbb{R}^s)$ satisfies the refinement equation* (1.2) *with the mask $b \in l^1(\mathbb{Z}^s)$, then $\hat{\phi}(2\beta\pi) = 0$ for $\beta \in \mathbb{Z}^s\backslash\{0\}$ and*

$$\sum_{\alpha\in\mathbb{Z}^s} \phi(\cdot - \alpha) = \hat{\phi}(0). \tag{3.1}$$

Now if for some $1 \leq p \leq \infty$, $\phi \in \mathcal{L}^p(\mathbb{R}^s) \subset L^1(\mathbb{R}^s)$ satisfies the refinement equation (1.2) with the mask $b \in l^1$ and has l^p-stable integer translates, we know from [5, Thm. 3.5] and Theorem 8 that $\hat{\phi}(0) \neq 0$. Therefore, after normalization we can assume that $\hat{\phi}(0) = 1$. In this case we can state property (R6) as follows.

THEOREM 9. *If $\phi \in \mathcal{L}^p(\mathbb{R}^s)$, $1 \leq p \leq \infty$, $\sum_{\alpha\in\mathbb{Z}^s} \phi(\cdot - \alpha) = 1$, then $\cup_{j\in\mathbb{Z}} V_j$ is dense in $(L^p(\mathbb{R}^s), \|\cdot\|_p)$ for $1 \leq p < \infty$ and in $(L^\infty(\mathbb{R}^s), \sigma(L^\infty, L^1))$ for $p = \infty$.*

Proof of Theorem 9 *for $p = \infty$.* Assume first that $f \in C_0(\mathbb{R}^s)$. We state that for any $g \in L^1(\mathbb{R}^s)$,

$$\epsilon_j := \left| \int_{\mathbb{R}^s} \left[f(x) - \sum_{\alpha\in\mathbb{Z}^s} \phi(M^j x - \alpha) f(M^{-j}\alpha) \right] g(x)\, dx \right| \to 0 \tag{3.2}$$

as $j \to \infty$.

Note that $C_0(\mathbb{R}^s)$ is dense in $L^1(\mathbb{R}^s)$ and

$$\left| f(x) - \sum_{\alpha\in\mathbb{Z}^s} \phi(M^j x - \alpha) f(M^{-j}\alpha) \right| \leq \|f\|_\infty + \|f\|_\infty \|\phi^0\|_\infty < \infty.$$

It is sufficient to prove (3.2) under the assumption that $g \in C_0(\mathbb{R}^s)$, which can, in turn, be derived from Theorem 8 by some similar arguments and methods from [6]. We omit the details here.

Once we have the statement (3.2), the proof is easier.

To see that $\cup_{j\in\mathbb{Z}} V_j$ is dense in $(L^\infty(\mathbb{R}^s), \sigma(L^\infty, L^1))$, we let $0 \neq f \in L^\infty(\mathbb{R}^s)$, $\{g_1, \ldots, g_n\} \subset L^1(\mathbb{R}^s)$, and $\epsilon > 0$. We need to find some $h \in \cup_{j\in\mathbb{Z}} V_j$ such that for $j = 1, 2, \ldots, n$,

$$\left| \int_{\mathbb{R}^s} (f(x) - h(x)) g_j(x)\, dx \right| < \epsilon. \tag{3.3}$$

By Lusin's Theorem, for any $\delta > 0$ there exist $f_\delta \in C(\mathbb{R}^s)$ and a measurable set $E_\delta \subset \mathbb{R}^s$ such that

$$|E_\delta| < \delta,$$
$$f_\delta(x) = f(x) \quad \text{if } x \in \mathbb{R}^s \backslash E_\delta,$$

and

$$\|f_\delta\|_\infty \leq \|f\|_\infty.$$

Furthermore, for any $r > 0$ we choose $f_{\delta,r} \in C_0(\mathbb{R}^s)$ such that supp $f_{\delta,r} \subset B_{r+1} := \{x \in \mathbb{R}^s : |x| \leq r + 1\}$, $f_{\delta,r}|_{B_r} = f_\delta|_{B_r}$, and

$$\|f_{\delta,r}\|_\infty \leq \|f_\delta\|_\infty \leq \|f\|_\infty.$$

Since $\{g_j\}_{j=1}^n \subset L^1(\mathbb{R}^s)$, for sufficiently large r and sufficiently small δ we have for $j = 1, 2, \ldots, n$,

$$\int_{E_\delta} |g_j(x)|\, dx < \frac{\epsilon}{8\|f\|_\infty}$$

and

$$\int_{\mathbb{R}^s - B_r} |g_j(x)|\, dx < \frac{\epsilon}{8\|f\|_\infty}.$$

Therefore,

$$\left|\int_{\mathbb{R}^s}(f(x)-f_{\delta,r}(x))g_j(x)\,dx\right| \leq \int_{E_\delta}|f(x)-f_\delta(x)||g_j(x)|\,dx$$
$$+\int_{\mathbb{R}^s-B_r}|f_{\delta,r}(x)-f_\delta(x)||g_j(x)|\,dx<\frac{\epsilon}{2}.$$

Now applying (3.2) to $f_{\delta,r}$ and $\{g_j\}_{j=1}^n$ we can find $m \in \mathbb{N}$ such that for $j = 1,2,\ldots,n$,

$$\left|\int_{\mathbb{R}^s}\left[f_{\delta,r}(x)-\sum_{\alpha\in\mathbb{Z}^s}\phi(M^m x-\alpha)f_{\delta,r}(M^{-m}\alpha)\right]g_j(x)\,dx\right|<\frac{\epsilon}{2}.$$

Hence, for $h(x)=\sum_{\alpha\in\mathbb{Z}^s}\phi(M^m x-\alpha)f_{\delta,r}(M^{-m}\alpha)\in\cup_{j\in\mathbb{Z}}V_j$, (3.3) is valid.

The proof of Theorem 9 is complete.

With all the above results, the proof of Theorem 6 is now easy.

Proof of Theorem 6. By the definition of $\{V_j\}_{j\in\mathbb{Z}}$, (R1), (R2), and (R3) follow immediately since the mask b of the refinement equation for ϕ is in $l^1(\mathbb{Z}^s)$. Note that the integer translates of ϕ are l^p-stable. We know that the mapping $L_\phi : l^p(\mathbb{Z}^s)\to V_0$ defined by

$$L_\phi(a)=\phi *' a$$

is an isomorphism from $l^p(\mathbb{Z}^s)$ onto V_0 that commutes with shift operators. Property (R5) follows from Theorem 7 while (R6) is obtained from Theorem 9 since the assumption $\sum_{\alpha\in\mathbb{Z}^s}\phi(\cdot-\alpha)=\hat{\phi}(0)\neq 0$ is satisfied by Theorem 8 and the stability.

The proof of Theorem 6 is complete.

4. Determination of scaling functions of Haar type. Combining the results in §§2 and 3, we can give some criteria for a refinable function to generate a multiresolution analysis in $L^2(\mathbb{R})$ associated with $(\mathbb{Z},k)$. This has an interesting application in determining all the scaling functions of Haar type.

Let $k\geq 2$ be a positive integer. We want to determine all those multiresolution analyses of $L^2(\mathbb{R})$ associated with $(\mathbb{Z},k)$ whose scaling functions are characteristic functions. It was shown by Gröchenig and Madych [4] that such a scaling function must be the characteristic function χ_Q of some set Q defined by

$$Q=\left\{x\in\mathbb{R}:x=\sum_{j=1}^{\infty}k^{-j}\epsilon_j,\epsilon_j\in\{n_0,n_1,\ldots,n_{k-1}\}\right\}, \tag{4.1}$$

where $\{n_0,n_1,\ldots,n_{k-1}\}$ is a collection of representatives of distinct cosets of $\mathbb{Z}/(k\mathbb{Z})$, say, $n_j\equiv j \pmod{k}$ for $j=0,1,\ldots,k-1$. However, χ_Q may fail to be a scaling function. In fact, for $k=2$, it is well known that $\chi_{[0,1)}$ is the unique scaling function up to an integer translate. For this, see Daubechies [3] and Gröchenig and Madych [4].

The main purpose of this section is to characterize for general $k\in\mathbb{N}$ such scaling functions in terms of the number theory property of the representatives of the cosets. To this end, we need Euler's theorem from number theory. Euler's theorem says that if p and q are relatively prime integers $q>1$, then

$$p^{\varphi(q)}\equiv 1 \pmod{q}, \tag{4.2}$$

where $\varphi(q) \in \mathbb{N}$ is the Euler number of q; i.e., $\varphi(q)$ is the number of positive integers $\leq q$ that are relatively prime to q.

Now we can state our main result in this section.

THEOREM 10. *Suppose that $\{n_0, n_1, \ldots, n_{k-1}\}$ is a collection of representatives of distinct cosets of $\mathbb{Z}/(k\mathbb{Z})$ and the set Q is defined by* (4.1). *If $k > 2$ and for $j = 0, 1, \ldots, k-1$, $n_j \equiv j \pmod k$. Then $\phi = \chi_Q$ is a scaling function of a multiresolution analysis in $L^2(\mathbb{R})$ associated with $(\mathbb{Z}, k)$ if and only if*

$$(n_1 - n_0, n_2 - n_0, \ldots, n_{k-1} - n_0) = 1. \tag{4.3}$$

Proof of Theorem 10. Since $\phi \in \mathcal{L}^2(\mathbb{R})$ and satisfies the refinement equation

$$\phi(x) = \sum_{j=0}^{k-1} \phi(kx - n_j), \tag{4.4}$$

by Theorem 1 in §2 and Theorem 6 in §3 it is sufficient to prove that (4.3) holds if and only if the symbol of the mask

$$\tilde{b}(z) = \sum_{j=0}^{k-1} z^{n_j} \tag{4.5}$$

satisfies conditions (2) and (3) in Theorem 1.

We first claim that condition (2) is always valid for any collection of representatives. To show this, we denote $\omega_0 = e^{-i2\pi/k}$. Then $\omega_0^k = 1$ and $e^{-i2l\pi/k} = \omega_0^l$.

For any $z \in T$, we must have

$$\begin{pmatrix} \tilde{b}(e^{-i2\pi 0/k}z) \\ \tilde{b}(e^{-i2\pi/k}z) \\ \vdots \\ \tilde{b}(e^{-i2\pi(k-1)/k}z) \end{pmatrix} = \begin{pmatrix} \tilde{b}(\omega_0^0 z) \\ \tilde{b}(\omega_0 z) \\ \vdots \\ \tilde{b}(\omega_0^{k-1}z) \end{pmatrix}$$
$$= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & \omega_0 & \cdots & \omega_0^{k-1} \\ \vdots & & & \\ 1 & \omega_0^{k-1} & \cdots & (\omega_0^{k-1})^{k-1} \end{pmatrix} \begin{pmatrix} z^{n_0} \\ z^{n_1} \\ \vdots \\ z^{n_{k-1}} \end{pmatrix} \neq 0$$

since the determinant of the coefficient matrix is a Vandermonde determinant and is therefore not equal to zero, while the vector $(z^{n_0}, \ldots, z^{n_{k-1}}) \neq 0$. Thus we have proved our claim.

Now we need only to prove that $\tilde{b}(z)$ does not satisfy condition (3) in Theorem 1 if and only if (4.3) is not true. We observe that the first statement is equivalent to the fact that there exist $m \in \mathbb{N}$ and $n \in \mathbb{Z}$ such that $n/(k^m - 1) \notin \mathbb{Z}$ and for $z = e^{-i2n\pi/(k^m-1)}$ and any $d \geq 0$, $l = 1, 2, \ldots, k-1$,

$$\tilde{b}(\omega_0^l z^{k^d}) = \sum_{j=0}^{k-1} \omega_0^{jl} e^{-i2\pi n n_j k^d/(k^m-1)} = 0,$$

which is equivalent to

$$A \begin{pmatrix} 1 \\ e^{-i2\pi n(n_1-n_0)k^d/(k^m-1)} \\ \vdots \\ e^{-i2\pi n(n_{k-1}-n_0)k^d/(k^m-1)} \end{pmatrix} = 0. \tag{4.6}$$

Here A is the $(k-1)\times k$ matrix defined by

$$(A)_{l,j} = \omega_0^{l(j-1)}, \qquad l = 1,\ldots,k-1,\ j = 1,\ldots,k.$$

The rows of A are the rows of a $k \times k$ matrix whose determinant is a Vandermonde determinant; hence, A has rank of $k-1$. Therefore, (4.6) has the unique solution $(1,1,\ldots,1)$.

Thus, $\tilde{b}(z)$ does not satisfy condition (3) in Theorem 1 if and only if there exist $m \in \mathbb{N}$ and $n \in \mathbb{Z}$ such that $n/(k^m-1) \notin \mathbb{Z}$ and

$$n(n_j - n_0)k^d/(k^m-1) \in \mathbb{Z} \quad \text{for } j = 1,2,\ldots,k-1 \text{ and } d \geq 0. \tag{4.7}$$

We can now prove the necessity and sufficiency as follows.

Suppose that (4.3) holds. Then (4.7) is not true, which implies that $\tilde{b}(z)$ satisfies condition (3) in Theorem 1. Since otherwise there exist $m \in \mathbb{N}$ and $n \in \mathbb{Z}$ such that $n/(k^m-1) \notin \mathbb{Z}$ and for $j = 1,2,\ldots,k-1, d \geq 0$, it follows that

$$n(n_j - n_0)k^d/(k^m-1) \in \mathbb{Z}.$$

But $(n_1 - n_0, n_2 - n_0, \ldots, n_{k-1} - n_0) = 1$ and $(k^d, k^m - 1) = 1$, so we must have $n/(k^m-1) \in \mathbb{Z}$, which is a contradiction. Thus we have proved the sufficiency.

Conversely, suppose that (4.3) does not hold. We need to show that (4.7) is true. To this end, we use Euler's theorem.

Let $q = (n_1 - n_0, n_2 - n_0, \ldots, n_{k-1} - n_0) \in \mathbb{N}$. By the assumption, $q \geq 2$.

On the other hand, $n_1 - n_0 \equiv 1 \pmod{k}$. Hence $(n_1 - n_0, k) = 1$ and $(q,k) = 1$. By Euler's theorem we have

$$k^{\varphi(q)} \equiv 1 \pmod{q}.$$

Let $m = \varphi(q) \in \mathbb{N}$ and $n = (k^{\varphi(q)} - 1)/q \in \mathbb{Z}$. Then,

$$\frac{n}{k^m - 1} = \frac{1}{q} \notin \mathbb{Z},$$

while for $j = 1,2,\ldots,k-1$,

$$n(n_j - n_0)k^d/(k^m-1) = k^d(n_j - n_0)/q \in \mathbb{Z}.$$

Hence, (4.7) is true, which implies that $\tilde{b}(z)$ does not satisfy condition (3) in Theorem 1. Therefore, the necessity holds.

The proof of Theorem 10 is complete.

More generally, if $(n_1 - n_0, n_2 - n_0, \ldots, n_{k-1} - n_0) \neq 1$, then we know that $\varphi = \chi_Q$ fails to be a scaling function. Equivalently, $|Q| > 1$.

Using Theorem 10, we can give a more exact statement for these cases.

COROLLARY. *Let $2 < k \in \mathbb{N}$ and for $j = 0, 1, \ldots, k-1$, $n_j \in \mathbb{Z}$ satisfies $n_j \equiv j$ (mod k). Then we have*

$$\left| \left\{ x \in \mathbb{R} : x = \sum_{j=1}^{\infty} k^{-j} \epsilon_j,\ \epsilon_j \in \{n_0, n_1, \ldots, n_{k-1}\} \right\} \right|$$
$$= (n_1 - n_0, n_2 - n_0, \ldots, n_{k-1} - n_0).$$

Thus we have determined all the Haar-type scaling functions of multiresolution analyses in $L^2(\mathbb{R})$ associated with $(\mathbb{Z}, k)$ for general k.

The cases of multidimensions and of smooth scaling functions will be discussed elsewhere.

Note added in proof. After my paper was submitted and reviewed, I learned from I. Daubechies that results similar to those in §4 had been obtained independently by K. Gröchenig and A. Haas in [Self-similar lattice tilings, *J. Fourier Anal. Appl.*].

REFERENCES

[1] C. K. CHUI AND J. Z. WANG, *A general framework of compactly supported splines and wavelets*, J. Approx. Theory, 71 (1992), pp. 263–304.

[2] A. COHEN, *Ondelettes, analysis multirésolutions et filtres mirroirs en quadrature*, Ann. Inst. H. Poincaré, 7 (1990), pp. 439–459.

[3] I. DAUBECHIES, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[4] K. GRÖCHENIG AND W. R. MADYCH, *Multiresolution analysis, Haar bases, and self-similar tilings of* $\mathbb{R}^n$, IEEE Trans. Inform. Theory, 38 (1992), pp. 556–568.

[5] R. Q. JIA AND C. A. MICCHELLI, *Using the refinement equation for the construction of pre-wavelets* II: *Powers of two*, in Curves and Surfaces, P. J. Laurent, A. Le Méhauté, and L. L. Schumaker, eds., Academic Press, New York, 1991, pp. 209–246.

[6] ———, *Using the refinement equation for the construction of pre-wavelets* V: *Extensibility of trigonometric polynomial*, Computing, 48 (1992), pp. 61–72.

[7] R. Q. JIA AND J. Z. WANG, *Stability and linear independence associated with wavelet decompositions*, Proc. Amer. Math. Soc., 117 (1993), pp. 1115–1124.

[8] P. G. LEMARIÉ, *Fonctions a support compact dans les analyses multi-résolutions*, Rev. Mat. Iberoamericana, 7 (1991), pp. 157–182.

[9] W. R. MADYCH, *Some elementary properties of multiresolution analyses of* $L^2(\mathbb{R}^n)$, in Wavelets: A Tutorial in Theory and Applications, C. K. Chui, ed., Academic Press, New York, 1992, pp. 259–294.

[10] S. G. MALLAT, *Multiresolution approximation and wavelet orthonormal bases of* $L^2(\mathbb{R})$, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.

[11] Y. MEYER, *Ondelettes et Opérateurs* I: *Ondelettes*, Hermann, Paris, 1990.

[12] S. D. RIEMENSCHNEIDER AND Z. W. SHEN, *Wavelets and prewavelets in low dimensions*, J. Approx. Theory, 71 (1992), pp. 18–38.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 905–913, July 1996

001

GLOBAL SOLVABILITY OF THE ANHARMONIC OSCILLATOR MODEL FROM NONLINEAR OPTICS*

J. L. JOLY[†], G. METIVIER[‡], AND J. RAUCH[§]

Abstract. The field equations describing the propagation of electromagnetic waves in a nonlinear dielectric medium whose polarization responds locally to the electric field as an anharmonic oscillator with potential $V(P)$ have smooth solutions global in space and time for arbitrary smooth initial data as soon as V has bounded derivatives of order less than or equal to three. This is true in spite of the fact that solutions of the nonlinear Shrödinger equation which approximate the fields in the slowly varying envelope approximation may blow up in finite time.

Key words. nonlinear optics, nonlinear Maxwell equations, nonlinear Schrödinger equation, saturated susceptibility equation

AMS subject classifications. 35L60, 35Q60, 78A60

1. Introduction. A standard model, due to Lorentz [L] (see also [F, Chaps. I-31 and II-33]), of the linear dispersive behavior of electromagnetic waves is given by the system of partial differential equations

$$\partial_t B + \operatorname{curl} E = 0, \tag{1.1}$$

$$\partial_t E - \operatorname{curl} B = -\partial_t P, \tag{1.2}$$

$$\partial_t^2 P + \partial_t P/T_1 + aP = bE \tag{1.3}$$

with positive constants a , b, and T_1. The physical origin of equation (1.3) is a model of the electron as bound to the nucleus by a Hooke's law spring force. Here E and B are the electric and magnetic fields and the vector field P is the polarization of the medium. A simple and natural model (see [Bl], [O]) to explain nonlinear dispersive phenomena is to replace the linear restoring force with a nonlinear law:

$$\partial_t^2 P + \partial_t P/T_1 + \nabla V(P) = bE. \tag{1.4}$$

If the Taylor expansion of V at the origin is

$$V(P) = \frac{a}{2}\,|P|^2 - \beta |P|^4 + \text{higher-order terms},$$

then asymptotic analysis of small-amplitude solutions reveals a focusing cubic term; that is, the nonlinear susceptibility, n_2, is strictly positive (see [B], [DR1], [NM]). In

*Received by the editors August 31, 1994; accepted for publication October 21, 1994. This research was supported by North American Treaty Organization grant CRG-890904, Office of Naval Research grant N-014-92-J-1245, and National Science Foundation grant DMS-90-03256.

[†] CEREMAB, Unitè de Recherche Associèe 226 de Centre National de la Recherche Scientifique, Universitè de Bordeaux I, 33405 Talence, France.

[‡] IRMAR, Unitè de Recherche Associèe 305 de Centre National de la Recherche Scientifique, Universitè de Rennes I, 35042 Rennes, France.

[§] Department of Mathematics, University of Michigan, Ann Arbor, MI 48109.

addition, the slowly varying envelope approximation (see [NM], [DR2]) leads to amplitudes which are solutions to nonlinear Schrödinger equations. In the monochromatic case, the equation has a focusing cubic nonlinearity when β is positive. In this case, there are solutions of the Schrödinger equation which explode in finite time. Our main theorem shows that the solutions of the underlying field equations (1.1), (1.2), and (1.4) do not break down. These equations have global smooth solution for arbitrary smooth data under appropriate hypotheses on the potential energy function V. An analogous result for the Maxwell–Bloch equations which come from modelling the matter as a gas of finite-state quantum systems is proved in [DR1]. In both cases, the fundamental nonlinear field equations are globally solvable even when the reduced Schrödinger equation is not. These contradictory predictions are resolved by the observation that near the focal point, amplitudes grow and the assumptions underlying the slowly varying envelope approximation are no longer satisfied. Once it is known that the underlying equations have smooth solutions, it is natural to ask what the behavior is near a focal point. This appears to us to be a very difficult problem. A first step in considering large-amplitude solutions is to go beyond the regime of Taylor expansions about $P = 0$. It might seem reasonable to simply take the potential

$$V(P) = \frac{a}{2}\,|P|^2 - \beta|P|^4.$$

However, for large displacements this is strongly repulsive. In fact, it is so repulsive that solutions of the classical spring equation

$$\partial_t^2 P + \partial_t P/T_1 + \nabla V(P) = 0 \tag{1.5}$$

with large initial energy diverge to infinity in finite time.

MAIN HYPOTHESIS. *The potential-energy function $V : \mathbf{R}^3 \to \mathbf{R}$ satisfies $V(0) = 0$ and is infinitely differentiable with second- and third-order partial derivatives uniformly bounded.*

This implies that $|\nabla V(P)|$ [resp. $V(P)$] grows at most linearly [resp. quadratically] as $P \to +\infty$. That is, there is a C such that for all $\beta \in \mathbf{N}^3$ with $|\beta| \leq 3$ and $P \in \mathbf{R}^3$,

$$|\partial^\beta V(P)| \leq C\ (1\ +\ |P|)^{\max(2-|\beta|,0)}. \tag{1.6}$$

An example is the potential

$$V(P) := \frac{\frac{a}{2}\,|P|^2 - \beta'|P|^4}{1+\gamma|P|^2}, \quad \gamma > 0\,.$$

On one hand, this hypothesis is very strong since the nonlinear term $\nabla V(P)$ is then a globally Lipshitzean function of P. In particular, the ordinary differential equation (1.5) is globally solvable. On the other hand, the hypothesis is reasonable since, as observed above, what is needed to produce a Kerr nonlinearity is that $\partial^2 V/\partial P_i \partial P_j$ be smaller than Hooke's law when $P \approx 0$. The hypothesis roughly asserts that this comparison is valid for all P. A second plausibility argument is that models of nonlinear susceptibility often include saturation effects. For such models, one would have $|\partial^\beta V(P)| \to 0$ as $P \to \infty$ for $|\beta| \geq 2$ and so the hypothesis would be satisfied.

Assuming the Main Hypothesis, it is routine to prove the global solvability of a semilinear equation of the form $(\partial_{tt} - \triangle)P + \nabla V(P) = 0$. However, there is no

Laplacian in the anharmonic spring equation (1.4), and the existence proof is delicate relying on the detailed structure of the system.

Taking the divergence of (1.1) and (1.2) implies that

$$\partial_t \operatorname{div}(E+P) = \partial_t \operatorname{div}(B) = 0. \tag{1.7}$$

The physically relevant solutions are those which satisfy

$$\operatorname{div}(E+P) = \operatorname{div}(B) = 0. \tag{1.8}$$

Thanks to 1.7 this holds as soon as it holds at $t = 0$, so (1.8) is only a constraint on the initial data.

MAIN THEOREM. *If $s \geq 2$ and the initial data $B(0), E(0), P(0)$, and $\partial_t P(0)$ belong to $H^s(\mathbf{R}^3)$ and satisty $\operatorname{div}\big(E(0)+P(0)\big) = \operatorname{div}(B(0)) = 0$, then there is one and only one solution of equations (1.1), (1.2), (1.4), and (1.8) which achieves these initial data and is such that B, E, P, and $\partial_t P$ are continuous functions of $t \in [0, \infty[$ with values in $H^s(\mathbf{R}^3)$. The values of the solution at $\underline{t}, \underline{x}$ depend only on the values of the initial data on the ball of radius $\underline{t}$ with center $\underline{x}$.*

It follows that if the initial data belong to $C_0^\infty(\mathbf{R}^3)$, then the solution has compact support whose diameter grows linearly with t. Using the equation to express time derivatives in terms of spatial derivatives shows that the solution is an infinitely differentiable function of t, x [Ra]. Let

$$U(t,x) := \big(E(t,x)\,, B(t,x)\,, P(t,x)\,, \partial_t P(t,x)\big). \tag{1.9}$$

Equation (1.4) is written as a system for the pair (P, Q) with $Q := \partial_t P$:

$$\partial_t P = Q, \qquad \partial_t Q = bE - \frac{Q}{T_1} - \nabla V(P).$$

The equations for U then take the form of a semilinear symmetric hyperbolic system:

$$\begin{gathered} LU := \partial_t U - \sum_{1 \leq j \leq 3} A_j \partial_j U = F(U), \\ F(U) := \left(-Q\,, 0\,, Q\,, bE - \frac{Q}{T_1} - \nabla V(P)\right). \end{gathered} \tag{1.10}$$

The symmetry means that the matrices A_j are symmetric and real. In this case, the A_j are 12×12 real symmetric matrices whose last six rows vanish. The next result is classical, dating back to Schauder [S]. A short modern proof uses the first theorem in [Re] in the Hilbert space of $U \in H^s(\mathbf{R}^3)$ such that $\operatorname{div}(E+P) = \operatorname{div}(B) = 0$. The operator$\sum_{1 \leq j \leq 3} A_j \partial_j$ is anti-self-adjoint and for $s > 3/2$, the nonlinear term $F(U)$ is locally Lipshitzean.

LOCAL EXISTENCE THEOREM. *If $s > 3/2$ and $U(0,.) \in H^s(\mathbf{R}^3)$, then there are a $T_* \in]\,0, \infty]$ and a unique $U \in C([0, T_{/,*}[: H^s(\mathbf{R}^3))$ which satisfy (1.1), (1.2), and (1.4) and attain these initial values. The solution depends continuously on the initial data in the sense that if $\varphi \in H^s(\mathbf{R}^3)$ and $T < T_*(\varphi)$, then there is an $H^s(\mathbf{R}^3)$ neighborhood $\mathcal{O}$ of φ such that $T_*(\psi) > T$ for all $\psi \in \mathcal{O}$ and the map from initial data to solution is continuous from $\mathcal{O}$ to $C([0,T] : H^s(\mathbf{R}^3))$. There is a lower bound $T_* \geq c(s, {}_{,}\|U(0)\|_{H^s(\mathbf{R}^3)}) > 0$, where $c(s, \lambda)$ is a nonincreasing function of λ. The* *values of U at $(\underline{t}, \underline{x})$ depend only on the values of the initial data on the ball of radius $\underline{t}$ with center $\underline{x}$.*

Our main result asserts that the solutions do not blow up in finite time. The Main Theorem then takes the following form.

MAIN THEOREM 1. *If $s \geq 2$ and the initial data $U(0,.) \in H^s(\mathbf{R}^3)$ satisfy $\operatorname{div}(E(0)+P(0)) = \operatorname{div}(B(0)) = 0$, then $T_* = \infty$.*

This result is a special case of a general result which isolates the essential features of (1.10) which guarantee global existence. Write the nonlinear term as a sum of linear and nonlinear parts by Taylor expansion at $U = 0$:

$$\begin{aligned} &F(U) := -BU + f(U), \\ (1.11) \qquad &BU := \left(-Q\,,\ 0\,,\ Q\,,\ bE - \frac{Q}{T_1} - \left(\nabla^2 V(0)\right) P\right), \\ &f(U) := \left(0\,,0\,,0\,,\ -\nabla V(P) + \left(\nabla^2 V(0)\right) P\right). \end{aligned}$$

The basic system then has the form

$$(1.12) \qquad \partial_t U - \sum_{1 \leq j \leq 3} A_j \partial_j U - BU = f(U).$$

We next introduce a class of $N \times N$ systems of the form of (1.12), including (1.10) as a special case.

ASSUMPTION 1. *The $N \times N$ system (1.12) has hermitian symmetric constant matrix coefficients A_j and constant B. For $\xi \neq 0$, the kernel of $A(\xi) := \sum \xi_j A_j$ has dimension independent of ξ.*

For the anharmonic oscillator model, this dimension is equal to 8, the kernel consisting of vectors such that E and B are parallel to ξ.

ASSUMPTION 2. *The nonlinear function $f : \mathbf{C}^N \to \mathbf{C}^N$ has range contained in $\bigcap \{\operatorname{Ker} A(\xi) : \xi \neq 0\}$. In addition, $f(0) = 0$, $\nabla f(0) = 0$, and the first derivatives $\partial f / \partial U_j$ are uniformly bounded in $\mathbf{C}^N$.*

For the anharmonic oscillator model, the Main Hypothesis yields the derivative bound in Assumption 2. Moreover, the nonlinear term takes values in the vectors whose first nine components vanish, and the kernel of the A_j includes the vectors whose first six components vanish, so Assumption 2 is satisfied . Thus Main Theorem 1 is a special case of the following result.

MAIN THEOREM 2. *For semilinear symmetric hyperbolic systems satisfying Assumptions 1 and 2, $s \geq 2$, and $U(0) \in H^2(\mathbf{R}^3)$, the Cauchy problem is globally solvable, that is, $T_* = \infty$.*

To prove Main Theorem 2, it is sufficient to prove the following a priori estimate. For any $T \in]0, \infty[$ and any $M > 0$, there is a constant $C(T, M)$ such that if $\underline{t} \in]0, T]$ and U is a smooth compactly supported solution of (1.12) on $[0, \underline{t}] \times \mathbf{R}^3$ such that

$$(1.13) \qquad \|U(0)\|_{H^2(\mathbf{R}^3)} \leq M,$$

then

$$(1.14) \qquad \|U(t,.)\|_{H^2(\mathbf{R}^3)} \leq C(T, M) \quad \text{for } 0 \leq t \leq \underline{t}.$$

To prove that (1.14) is sufficient, one must show that if $U(0) \in H^s(\mathbf{R}^3)$, $s \geq 2$, and $T > 0$, then $T_* \geq T$. Gagliardo–Nirenberg estimates imply that for a smooth solution U and $s \geq 2$, one has

$$(1.15) \qquad \|U(t,.)\|_{H^s(\mathbf{R}^3)} \leq C\,\|U(0)\|_{H^s(\mathbf{R}^3)} \quad \text{for } 0 \leq t \leq \underline{t},$$

where C only depends on the L^∞ norm of U on $[0,\underline{t}] \times \mathbf{R}^3$. These estimates and the Local Existence Theorem imply the following corollary (see, e.g., [M], [GR]). If $T_* < +\infty$, then

$$\limsup_{t\to T_*} \|U(t,.)\|_{L^\infty(\mathbf{R}^3)} = +\infty. \tag{1.16}$$

Moreover, by continuity, the a priori estimates (1.14) extend to solutions $U \in C^0([0,\underline{t}] : H^s(\mathbf{R}^3))$. The Sobolev imbedding $H^2(\mathbf{R}^3) \subset L^\infty(\mathbf{R}^3)$ and (1.14) show that the L^∞ norm of an H^s solution U cannot blow up before time T and therefore $T_* \geq T$.

The proof of Main Theorem 2 proceeds by a sequence of estimates leading to (1.14). These estimates are presented in §§2–6.

Remark 1. The divergence conditions $\operatorname{div}\big(E(0)+P(0)\big) = \operatorname{div}\big(B(0)\big) = 0$ are not necessary for global existence. We included them in the statement of Main Theorem 1 since they are required for the physically relevant solutions.

Remark 2. Main Theorem 2 applies to the sytem (1.1)–(1.2) with (1.4) replaced by the more general law

$$\partial_t^2 P = F(E, B, P, \partial_t P), \tag{1.17}$$

where $F \in C^3(\mathbf{R}^{12} : \mathbf{R}^3)$, $F(0) = 0$, and F has bounded first and second derivatives. Another generalization is the case of a finite number of anharmonic oscillators. This corresponds more closely to what one would find from nonlinear terms in quantum perturabtion theory. The dynamics of the polarization is then given by

$$P = \sum_{j=1}^{n} P^j\,, \qquad \partial_t^2 P^j + \partial_t P^j / T_1^j + \nabla V^j(P^j) = b^j E\,.$$

Here each of the potentials V^j satisfies the Main Hypothesis.

2. $L^2(\mathbf{R}^3)$ estimates for U.

Let

$$K := \|(B+B^*)/2\| + \sup_{u\in\mathbf{C}^N} \{\|\nabla f(U)\|\}, \tag{2.1}$$

which is finite thanks to Assumption 2. Then since $f(0) = 0$, one has the Lipshitz bound

$$\|f(U)\|_{\mathbf{C}^N} \leq K\|U\|_{\mathbf{C}^N}. \tag{2.2}$$

For a smooth compactly supported solution U satisfying (1.13), we perform the standard energy estimate. Namely, take the $\mathbf{C}^N$ scalar product of the the partial differential equation (1.12) with U and integrate over $\mathbf{R}^3$. Then take the real part to find that

$$\partial_t \|U(t)\|^2_{L^2(\mathbf{R}^3)} \leq 2K \|U(t)\|^2_{L^2(\mathbf{R}^3)}.$$

It follows that

$$\|U(t)\|_{L^2(\mathbf{R}^3)} \leq e^{Kt} \|U(0)\|_{L^2(\mathbf{R}^3)}. \tag{2.3}$$

3. An $H^1(\mathbf{R}^3)$ estimate for U.

Let ∂ denote one of $\partial/\partial x_j$ for $1 \leq j \leq 3$. Then

$$(L-B)(\partial U) = -\partial\big(f(U)\big) = -(\nabla f(U))\,\partial U. \tag{3.1}$$

The main hypothesis implies that ∇f is bounded and thus

$$|\,\partial(f(U))\,(t,x)\,| \le K\,|\partial U(t,x)|\,. \tag{3.2}$$

The standard energy method, namely, taking the scalar product of (3.1) with ∂U and integration dx over $\mathbf{R}^3$, yields

$$\partial_t\,||\partial U(t)||^2_{L^2(\mathbf{R}^3)} \le 2K\,||\partial U(t)||^2_{L^2(\mathbf{R}^3)}. \tag{3.3}$$

Summing the resulting expressions over the three values of ∂ and applying Gronwall's inequality shows that for $0 \le t \le \underline{t} \le T$,

$$||\partial U(t)||_{L^2(\mathbf{R}^3)} \le C(T,M). \tag{3.4}$$

In particular, we control the $H^1(\mathbf{R}^3)$ norm of $U(t)$.

4. An $H^2(\mathbf{R}^3)$ estimate for the propagating part of U. Denote by $\Pi_0(\xi)$ the orthogonal projector on $\operatorname{Ker}A(\xi)$, defining a smooth matrix-valued function homogeneous of degree zero on $\xi \neq 0$. Let $\Pi_1(\xi) := I - \Pi_0(\xi)$. The next estimates depend on the decomposition

$$U = u_0 + u_1\,, \qquad u_\nu := \Pi_\nu(D)\,U := \mathcal{F}^{-1}\,\Pi_\nu(\xi)\,\mathcal{F}\,U. \tag{4.1}$$

The key intuition is that u_0 corresponds to zero speeds and u_1 to nonzero speeds. This will show that u_0 can only weakly influence u_1 since the corresponding waves cross transversely. Furthermore, the nonlinear term does not influence the u_1 part thanks to Assumption 2. These two facts suffice to pass from an H^1 estimate for U to an H^2 estimate for u_1. Let

$$B_{\mu\nu}(D) := \Pi_\mu(D)\,B\,\Pi_\nu(D). \tag{4.2}$$

Multiply (1.12) by $\Pi_1(D)$ using the facts that $\Pi_1 f = 0$ and $\Pi_1(\xi)A(\xi) = A(\xi)\Pi_1(\xi) = A(\xi)$ to find

$$\Big(\partial_t - \sum A_j\partial_j\Big)u_1 - B_{11}(D)u_1 - B_{10}(D)u_0 = 0. \tag{4.3}$$

Similarly, multiplying by $\Pi_0(D)$ yields

$$\partial_t u_0 - B_{00}u - B_{01}(D)\,u_1 = \Pi_0(D)\,f(U).$$

Differentiate with respect to x to find

$$\partial_t Du_0 - B_{00}u_0 - B_{01}(D)\,Du_1 = \Pi_0(D)\,\big(f'(U)DU\big). \tag{4.4}$$

Since f' is bounded and the $\Pi_\nu(D)$ are bounded in L^2, this yields the estimate

$$||\partial_t u_0(t)||_{H^1(\mathbf{R}^3)} \le C\,||U(t)||_{H^1(\mathbf{R}^3)}. \tag{4.5}$$

Let $\langle\xi\rangle := (1 + |\xi|^2)^{1/2}$ and notice that u_1 then satisfies the modified equation

$$\Big(\partial_t - \sum A_j\partial_j - i\langle D\rangle\Pi_0(D)\Big)u_1 - B_{11}u_1 - B_{10}u_0 = 0 \tag{4.6}$$

involving the anti-self-adjoint elliptic operator

$$G(D) := \sum A_j\partial_j + i\langle D\rangle \Pi_0(D), \quad G(\xi) := -i\langle\xi\rangle\Pi_0(\xi) + \sum i\, A_j\xi_j. \tag{4.7}$$

Then $u_1 = z + w$, where z and w are the solutions of the initial-value problems

$$\big(\partial_t - G(D)\big)z - B_{11}z = 0, \qquad z(0,\,.\,) = u_1(0,\,.\,), \tag{4.8}$$

$$\big(\partial_t - G(D)\,\big)w - B_{11}w - B_{10}u_0 = 0, \quad w(0,\,.\,) = 0. \tag{4.9}$$

For z, we have the elementary estimate

$$||z(t)||_{H^2(\mathbf{R}^3)} \le e^{Kt}\,||z(0)||_{H^2(\mathbf{R}^3)} \le e^{Kt}\,M. \tag{4.10}$$

Duhamel's formula yields the following formula for the Fourier transform of w:

$$\hat{w}(t,\xi)) = \int_0^t e^{(G(\xi)+B_{11}(\xi))(t-s)}\, B_{10}(\xi)\,\hat{u}_0(s,\xi)\; ds. \tag{4.11}$$

Introduce

$$\Phi(t,s,\xi) := -\int_s^t e^{(G(\xi)+B_{11}(\xi))\,(t-\sigma)}\; d\sigma$$

so that

$$\partial_s\,\Phi(t,s,\xi) = e^{(G(\xi)+B_{11}(\xi))\,(t-s)} \quad \text{and} \quad \Phi(t,t,\xi) = 0.$$

Integration by parts in (4.11) yields

$$\hat{w}(t,\xi)) = \Phi(t,0,\xi)B_{10}(\xi)\,\hat{u}_0(0,\xi) - \int_0^t \Phi(t,s,\xi)B_{10}(\xi)\,\partial_t\hat{u}_0(s,\xi)\; ds. \tag{4.12}$$

The symmetric hyperbolicity implies that

$$||\Phi(t,s,\xi)|| \le \int_0^{|t-s|} e^{K\sigma}\, d\sigma. \tag{4.13}$$

The ellipticity of G implies that for large $|\xi|$, $G+B$ is invertible and has norm $O(|\xi|^{-1})$. For such ξ,

$$\Phi(t,s,\xi) = -(G(\xi) + B_{11}(\xi))^{-1}\big(e^{(G(\xi)+B_{11}(\xi))(t-s)} - I\big).$$

Together with (4.13), this implies

$$|\Phi(t,s,\xi)| \le C(T)\langle\xi\rangle^{-1} \quad \text{for } s,t \in [0,T] \text{ and } \xi \in \mathbf{R}^3. \tag{4.14}$$

Then (4.12) yields

$$\begin{aligned} ||w(t)||_{H^2(\mathbf{R}^3)} &\le C(T)\bigg(||u_0(0)||_{H^1(\mathbf{R}^3)} + \int_0^t ||\partial_t u_0(s)||_{H^1(\mathbf{R}^3)}\; ds\bigg) \\ &\le C(T,M), \end{aligned} \tag{4.15}$$

where (3.4) and (4.5) are used in the last step. Estimates (4.10) and (4.15) show that

$$\|u_1(t)\|_{H^2(\mathbf{R}^3)} \leq C(T,M). \tag{4.16}$$

5. An $L^6(\mathbf{R}^3)$ estimate for Du_0. Consider the right-hand side of (4.4). Since f' is bounded, it follows that

$$\|f'(U(t))DU\|_{L^6(\mathbf{R}^3)} \leq C\|DU(t))\|_{L^6(\mathbf{R}^3)}. \tag{5.1}$$

Since the singular integral operator $\Pi_0(D)$ is bounded from $L^6(\mathbf{R}^3)$ to itself, (4.4) and (5.1) yield

$$\begin{aligned} \|\partial_t Du_0\|_{L^6(\mathbf{R}^3)} &\leq C(\|Du_1\|_{L^6(\mathbf{R}^3)} + \|DU(t))\|_{L^6(\mathbf{R}^3)}) \\ &\leq 2C\,\|Du_1\|_{L^6(\mathbf{R}^3)} + C\,\|\,Du_0\|_{L^6(\mathbf{R}^3)}. \end{aligned} \tag{5.2}$$

The Sobolev imbedding $H^1(\mathbf{R}^3) \subset L^6(\mathbf{R}^3)$ implies that

$$\|Du_1\|_{L^6(\mathbf{R}^3)} \leq C'\|Du_1(t)\|_{H^1(\mathbf{R}^3)} \leq C(T,M) \tag{5.3}$$

thanks to (4.16). Thus integrating inequality (5.2) yields

$$\|Du_0\|_{L^6(\mathbf{R}^3)} \leq C(T,M). \tag{5.4}$$

6. Endgame. The second derivatives $v := D_iD_jU$ satisfy

$$Lv = f'(U)\,v + \sum_{1\leq\alpha,\beta\leq 3} \frac{\partial^2 f(U)}{\partial U_\alpha \partial U_\beta}\,(D_iU_\alpha)\,(D_jU_\beta). \tag{6.1}$$

Thanks to the boundedness of f' and f'', the standard energy method applied to (6.1) yields

$$\begin{aligned} \partial_t\|D^2U(t)\|^2_{L^2(\mathbf{R}^3)} &\leq C\|D^2U(t)\|^2_{L^2(\mathbf{R}^3)} + C\int_0^t \|D_iU(s)\,D_jU(s)\|^2_{L^2(\mathbf{R}^3)}\,ds \\ &\leq C\|D^2U(t)\|^2_{L^2(\mathbf{R}^3)} + C\int_0^t \|DU(s)\|^2_{L^4(\mathbf{R}^3)}\,ds. \end{aligned} \tag{6.2}$$

Interpolating between the L^2 estimate (3.4) and the L^6 estimates (5.3)–(5.4) shows that the integrand is bounded by $C(T,M)$. Then Gronwall's method shows that

$$\|D^2U(t)\|_{L^2(\mathbf{R}^3)} \leq C(T,M) \quad \text{for} \quad 0 \leq t \leq T. \tag{6.3}$$

This together with (2.3) and (3.4) proves the desired estimate (1.14), and the proof of the Main Theorems is complete.

Acknowledgments. The authors gratefully acknowledge the support of the North Atlantic Treaty Organization, the U. S. Office of Naval Research, and the National Science Foundation. They also thank the Centre de Mathématiques and the Centre de Mathématiques Appliquées of the Ecole Polytechnique in Palaiseau for the welcome offered to J. Rauch during the academic year 1992–1993. All three thank Simone, Florence, Ludovic, Geraldine, and the citizens of St. Marsal for their hospitality in July 1993, when much of this research was performed.

REFERENCES

[Bl] N. BLOEMBERGEN, *Nonlinear Optics*, W. A. Benjamin, Inc., New York, 1965.

[B] R. BOYD, *Nonlinear Optics*, Academic Press, New York, 1992.

[DR1] P. DONNAT AND J. RAUCH, *Global solvability of the Maxwell–Bloch equations from nonlinear optics*, Arch. Rational Mech. Anal., to appear.

[DR2] ———, *Diffractive dispersive nonlinear optics*, to appear.

[GR] P. GÉRARD AND J. RAUCH, *Propagation de la régularité locale de solutions d'équations hyperboliques non linéaires*, Ann. Inst. Fourier (Grenoble), 37 (1987), pp. 65–84.

[F] R. P. FEYNMANN, *Lectures on Physics*, vols. I and II, Addison–Wesley, Reading, MA, 1963 and 1964.

[L] H. A. LORENTZ, *The Theory of Electrons*, 2nd ed., Teubner, Leipzig, Stuttgart, 1908; reprinted by Dover, New York, 1952.

[M] A. MAJDA, *Compressible Fluid Flows and Systems of Conservation Laws in Several Variables*, Appl. Math. Sci. 33, Springer-Verlag, Berlin, New York, Heidelberg, 1984.

[NM] A. NEWELL AND J. MALONEY, *Nonlinear Optics*, Addison–Wesley, Reading, MA, 1992.

[O] A. OWYOUNG, *The origin of the nonlinear refractive indices of liquids and glasses*, Ph.D. dissertation, California Institute of Technology, Pasadena, CA, 1971.

[Ra] J. RAUCH, *Partial Differential Equations*, Springer-Verlag, New York, 1991.

[Re] M. REED, *Abstract Non-Linear Wave Equations*, Springer-Verlag, New York, 1975.

[S] J. SCHAUDER, *Das Anfangswertproblem einer quasilinearen hyperbolischen Differntialgleichung zweiter Ordnung*, Fund. Math., 24 (1935), pp. 213–246.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 914–931, July 1996

002

INITIAL-VALUE PROBLEMS WITH INFLOW BOUNDARIES FOR MAXWELL FLUIDS*

MICHAEL RENARDY†

Abstract. We consider the two-dimensional flow of an upper convected Maxwell fluid transverse to a domain bounded by parallel planes. We characterize a set of inflow boundary conditions, which leads to a well-posed initial-boundary value problem.

Key words. polymer rheology, inflow boundaries, local existence

AMS subject classifications. 35L20, 35Q99, 76A10

1. Introduction.

Many problems in computational fluid dynamics involve "open" boundaries (i.e., boundaries through which the fluid enters or leaves the domain). Such open boundaries typically arise from the need to truncate the domain. Boundary conditions at open boundaries are therefore not dictated by physics but are a mathematical artifact. For Newtonian fluids, boundary conditions at open boundaries can be chosen of the same type as those at physical boundaries. For example, the prescription of either velocities or tractions leads to a well-posed problem. (Of course, it is a considerable problem to decide which boundary conditions are suitable for a specific application.) Non-Newtonian fluids, on the other hand, present a more fundamental problem. Such fluids have memory, and hence their equations of motion require information which depends on the flow of the fluid before it enters the domain. This manifests itself in the need for additional boundary conditions at inflow boundaries.

A simple model problem for studying this issue is the perturbation of uniform flow in a domain bounded by two parallel planes. In [2], I studied this problem for a fluid with a differential constitutive equation of Maxwell type. Velocities were prescribed on both planes and, in addition, partial information about the extra stresses was needed at the inflow boundary. An alternative way to prescribe such partial information was given in [4]. The solution was constructed by an iteration which alternately solves an elliptic problem of the same kind as the Stokes equation and then determines stresses by integration along stream lines. However, the procedure in [2] is limited to steady flows and does not extend in any straightforward way to time-dependent problems. An existence result for time-dependent problems without open boundaries was given in [3].

It appears natural to approach time-dependent problems with open boundaries by combining the approaches of [2] and [3]. In essence, this is what we shall do in this paper. However, there is a difficulty. Problems which are elliptic in the steady case [2] are replaced by hyperbolic problems in the time-dependent case. Associated with this is a loss of regularity of the solution, which becomes extremely difficult to control in an iteration. In order to overcome this difficulty, we shall have to rely on a special feature which is particular to one constitutive model (the upper convected

*Received by the editors April 11, 1994; accepted for publication (in revised form) November 14, 1994. This research was supported by National Science Foundation grant DMS-9306635 and Office of Naval Research grant N00014-92-J-1664.

†Department of Mathematics and Interdisciplinary Center for Applied Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0123 (renardym@math.vt.edu).

Maxwell fluid), and we shall have to modify the boundary conditions. This is not a very satisfactory state of affairs, and further research on this problem is needed.

2. Governing equations.

The equations of motion for an incompressible fluid are

$$\rho\Big(\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v}\Big) = \operatorname{div}\mathbf{T} - \nabla p + \mathbf{f}, \tag{1}$$

$$\operatorname{div}\mathbf{v} = 0. \tag{2}$$

Here $\mathbf{v}$ is the velocity, $\mathbf{T}$ the extra stress, p the pressure, ρ the density, and $\mathbf{f}$ a given body force. We assume the constitutive law of an upper convected Maxwell fluid:

$$\frac{\partial \mathbf{T}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{T} - (\nabla\mathbf{v})\mathbf{T} - \mathbf{T}(\nabla\mathbf{v})^T + \lambda\mathbf{T} = \mu(\nabla\mathbf{v} + (\nabla\mathbf{v})^T), \tag{3}$$

where λ and μ are positive constants, and the gradient of a vector is defined with the convention that the column index refers to the direction of differentiation and the row index to the component of the vector.

Our task will be to solve the equations for $t \geq 0$ and $\mathbf{x} = (x,y) \in \Omega := (0,L)\times\mathbb{R}$, subject to initial conditions

$$\mathbf{v}(\mathbf{x},0) = \mathbf{v}_0(\mathbf{x}), \qquad \mathbf{T}(\mathbf{x},0) = \mathbf{T}_0(\mathbf{x}), \qquad \mathbf{x}\in\Omega, \tag{4}$$

and boundary conditions. Throughout, we shall assume periodicity in y with a given period M. We shall make assumptions on the velocity which imply that the boundary $x = 0$ is an inflow boundary and the boundary $x = L$ is an outflow boundary. The boundary conditions for the velocity will be as follows:

$$\mathbf{v}(0,y,t) = \mathbf{v}_{\text{in}}(y,t), \ \Big(\frac{\partial^2\mathbf{v}}{\partial t^2} + (\mathbf{v}\cdot\nabla)\frac{\partial\mathbf{v}}{\partial t} - \Big(\frac{\partial\mathbf{v}}{\partial t}\cdot\nabla\Big)\mathbf{v}\Big)(L,y,t) = \mathbf{w}_{\text{out}}(y,t),$$

$$\frac{\partial\mathbf{v}}{\partial t}\cdot\mathbf{e}_1(L,y,0) = u_{\text{out}}(y). \tag{5}$$

Here $\mathbf{e}_1$ is a unit vector in the x-direction. Equation (5) means that we prescribe the velocity at the inflow boundary, while the quantity prescribed at the outflow boundary is the time derivative of the acceleration modified by a correction term which makes it into a divergence-free vector field. The condition on the normal component of $\partial\mathbf{v}/\partial t$ will be needed to determine an initial value of $\partial\mathbf{v}/\partial t$; note that (1) determines $\partial\mathbf{v}/\partial t$ up to a gradient, and the boundary values of the normal component are precisely the additional information that is required. The choice of boundary conditions might appear peculiar at this point; we shall see later in the proof why this particular choice makes the energy estimates "work." In addition, we need inflow boundary conditions for the stress. They are as follows:

$$p(0,y,t) = \pi(y,t), \qquad \frac{1}{2}(T_{11}(0,y,t) - T_{22}(0,y,t)) = \tau(y,t). \tag{6}$$

The inflow boundary conditions chosen here are slightly different from those in [2]. Moreover, the proof given in this paper works only for the upper convective Maxwell model and not for similar models of differential type. The reasons for this will be given below.

3. Construction of solutions. The equations (1)–(3) form a system of combined type, and there appear to be no methods available to deal with such systems directly. Hence all schemes to construct solutions have been based on some preprocessing of the equations, which leads to a decoupling into simpler problems at leading order. We shall also use this idea here. We begin by applying the operation $\partial/\partial t + (\mathbf{v}\cdot\nabla) + \lambda + (\nabla\mathbf{v})^T$ to equation (1). After using (3), this yields

$$\begin{aligned}&\rho\Big(\frac{\partial^2\mathbf{v}}{\partial t^2} + 2(\mathbf{v}\cdot\nabla)\frac{\partial\mathbf{v}}{\partial t} + \Big(\frac{\partial\mathbf{v}}{\partial t}\cdot\nabla\Big)\mathbf{v} + (\mathbf{v}\cdot\nabla)^2\mathbf{v} + ((\nabla\mathbf{v})^T + \lambda)\Big[\frac{\partial\mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v}\Big]\Big)\\ &= \mu\Delta\mathbf{v} + (\mathbf{T}:\partial^2)\mathbf{v} + (\nabla\mathbf{v} + (\nabla\mathbf{v})^T)\text{div }\mathbf{T} - \nabla q + \frac{\partial\mathbf{f}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{f} + (\nabla\mathbf{v})^T\mathbf{f} + \lambda\mathbf{f}.\end{aligned}\tag{7}$$

Here the notation $\mathbf{T}:\partial^2$ stands for $\sum T_{ij}\partial^2/\partial x_i\partial x_j$ and

$$q = \frac{\partial p}{\partial t} + (\mathbf{v}\cdot\nabla)p + \lambda p.\tag{8}$$

In [3], an iteration scheme was used to solve the equations. The scheme alternately uses (7) to update the velocity field and then (3) to update the stress. In order to do this, one has to prescribe $\mathbf{T}$ at the inflow boundary. However, the outcome would then be a solution to (3) and (7), and in order to get back to (1), one has to restrict the inflow boundary data so that (1) is satisfied at the inflow boundary. In [2], this led to a system of ordinary differential equations (ODEs) for the inflow boundary data, which was used to determine some stress components in terms of others. This system has to be solved at each step of the iteration to generate the inflow conditions for (3). Roughly speaking, one solves (3) and (8) for x-derivatives of $\mathbf{T}$ and p and inserts the result in (1). The resulting system is then solved for the values of T_{12} and p on the inflow boundary. In the time-dependent case, following the same procedure leads to a hyperbolic partial differential equation (PDE) system in place of the ODEs. Associated with this is a loss of regularity, which I do not know how to cope with in general. For the upper convected Maxwell model, however, the special features of the equation allow some "slack" so that the loss of regularity does not matter. To take advantage of this, we need to modify the inflow boundary conditions that were used in [2]. Instead of prescribing T_{11} and T_{22} as we did there, we now prescribe p and $(T_{11} - T_{22})/2$.

We need to reformulate (7) a little further. We solve (1) for div $\mathbf{T}$ and substitute the result on the right-hand side of (7). As a result, we find

$$\begin{aligned}&\rho\Big(\frac{\partial^2\mathbf{v}}{\partial t^2} + 2(\mathbf{v}\cdot\nabla)\frac{\partial\mathbf{v}}{\partial t} + \Big(\frac{\partial\mathbf{v}}{\partial t}\cdot\nabla\Big)\mathbf{v} + (\mathbf{v}\cdot\nabla)^2\mathbf{v} + (\lambda - (\nabla\mathbf{v}))\Big[\frac{\partial\mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v}\Big]\Big)\\ &= \mu\Delta\mathbf{v} + (\mathbf{T}:\partial^2)\mathbf{v} + (\nabla\mathbf{v} + (\nabla\mathbf{v})^T)\nabla p - \nabla q + \frac{\partial\mathbf{f}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{f} - (\nabla\mathbf{v})\mathbf{f} + \lambda\mathbf{f}.\end{aligned}\tag{9}$$

In contrast to (7), the right-hand side of (9) does not contain any derivatives of $\mathbf{T}$ but only of p. This trick works only for the upper convected Maxwell model, and it is this feature which allows us to cope with a loss of regularity as long as it occurs only in $\mathbf{T}$ and not in p.

The iteration will now proceed as follows. For given $\mathbf{v}^n$, q^n, p^n, and $\mathbf{T}^n$, we determine p^{n+1} by solving the equation

$$\frac{\partial p^{n+1}}{\partial t} + (\mathbf{v}^n\cdot\nabla)p^{n+1} + \lambda p^{n+1} = q^n\tag{10}$$

with initial condition $p^{n+1}(x,y,0) = p_0(x,y)$ and inflow condition $p^{n+1}(0,y,t) = \pi(y,t)$. Here p_0 is the initial pressure, which can be determined from (1) and the prescribed initial and boundary data. To avoid the usual ambiguity of undetermined constants in the pressure, we shall fix the spatial average of q and of the initial value of p to be zero.

Next, we need to provide complete inflow conditions for the stresses. Let us use the notations $T_+ = (T_{11} + T_{22})/2$ and $T_- = (T_{11} - T_{22})/2$. We rewrite (1) as follows:

$$\begin{aligned} &\frac{\partial T_+}{\partial x} + \frac{\partial T_{12}}{\partial y} = \frac{\partial p}{\partial x} - \frac{\partial T_-}{\partial x} + \rho\left(\frac{\partial v_1}{\partial t} + (\mathbf{v}\cdot\nabla)v_1\right) - f_1, \\ &\frac{\partial T_{12}}{\partial x} + \frac{\partial T_+}{\partial y} = \frac{\partial p}{\partial y} + \frac{\partial T_-}{\partial y} + \rho\left(\frac{\partial v_2}{\partial t} + (\mathbf{v}\cdot\nabla)v_2\right) - f_2. \end{aligned} \tag{11}$$

We now solve (3) and (8) for the x-derivatives of $\mathbf{T}$ and p and insert the result in (11). The resulting system is of the form

$$\begin{aligned} &-\frac{1}{v_1}\frac{\partial T_+}{\partial t} - \frac{v_2}{v_1}\frac{\partial T_+}{\partial y} + \frac{\partial T_{12}}{\partial y} = \phi_1\left(\mathbf{v}, \frac{\partial \mathbf{v}}{\partial t}, \nabla\mathbf{v}, q, \mathbf{T}, \frac{\partial T_-}{\partial t}, \frac{\partial T_-}{\partial y}, p, \frac{\partial p}{\partial t}, \frac{\partial p}{\partial y}, \mathbf{f}\right), \\ &-\frac{1}{v_1}\frac{\partial T_{12}}{\partial t} - \frac{v_2}{v_1}\frac{\partial T_{12}}{\partial y} + \frac{\partial T_+}{\partial y} = \phi_2\left(\mathbf{v}, \frac{\partial \mathbf{v}}{\partial t}, \nabla\mathbf{v}, q, \mathbf{T}, \frac{\partial T_-}{\partial t}, \frac{\partial T_-}{\partial y}, p, \frac{\partial p}{\partial t}, \frac{\partial p}{\partial y}, \mathbf{f}\right). \end{aligned} \tag{12}$$

Since (12) contains to x-derivatives of the stresses, we can impose it as a condition to be satisfied at the inflow boundary $x = 0$. We incorporate it in the iteration as follows:

$$\begin{aligned} &-\frac{1}{v_1^n}\frac{\partial T_+^{n+1}}{\partial t} - \frac{v_2^n}{v_1^n}\frac{\partial T_+^{n+1}}{\partial y} + \frac{\partial T_{12}^{n+1}}{\partial y} \\ &\qquad = \phi_1\left(\mathbf{v}^n, \frac{\partial \mathbf{v}^n}{\partial t}, \nabla\mathbf{v}^n, q^n, \mathbf{T}^{n+1}, \frac{\partial \tau}{\partial t}, \frac{\partial \tau}{\partial y}, \pi, \frac{\partial \pi}{\partial t}, \frac{\partial \pi}{\partial y}, \mathbf{f}\right), \\ &-\frac{1}{v_1^n}\frac{\partial T_{12}^{n+1}}{\partial t} - \frac{v_2^n}{v_1^n}\frac{\partial T_{12}^{n+1}}{\partial y} + \frac{\partial T_+^{n+1}}{\partial y} \\ &\qquad = \phi_2\left(\mathbf{v}^n, \frac{\partial \mathbf{v}^n}{\partial t}, \nabla\mathbf{v}^n, q^n, \mathbf{T}^{n+1}, \frac{\partial \tau}{\partial t}, \frac{\partial \tau}{\partial y}, \pi, \frac{\partial \pi}{\partial t}, \frac{\partial \pi}{\partial y}, \mathbf{f}\right). \end{aligned} \tag{13}$$

Here

$$\mathbf{T}^{n+1} = \begin{pmatrix} T_+^{n+1} + \tau & T_{12}^{n+1} \\ T_{12}^{n+1} & T_+^{n+1} - \tau \end{pmatrix}. \tag{14}$$

Having solved (13), we now have complete inflow data for the stresses, which we use to solve (3):

$$\frac{\partial \mathbf{T}^{n+1}}{\partial t} + (\mathbf{v}^n\cdot\nabla)\mathbf{T}^{n+1} = (\nabla\mathbf{v}^n)\mathbf{T}^n + \mathbf{T}^n(\nabla\mathbf{v}^n)^T - \lambda\mathbf{T}^n + \mu(\nabla\mathbf{v}^n + (\nabla\mathbf{v}^n)^T). \tag{15}$$

Finally, we obtain new values for $\mathbf{v}$ and q by solving the equation

$$\begin{aligned} &\rho\Big(\frac{\partial^2\mathbf{v}^{n+1}}{\partial t^2} + 2(\mathbf{v}^n\cdot\nabla)\frac{\partial \mathbf{v}^{n+1}}{\partial t} + \Big(\frac{\partial \mathbf{v}^n}{\partial t}\cdot\nabla\Big)\mathbf{v}^n + (\mathbf{v}^n\cdot\nabla)^2\mathbf{v}^{n+1} \\ &\quad + (\lambda - (\nabla\mathbf{v}^n))\Big[\frac{\partial \mathbf{v}^n}{\partial t} + (\mathbf{v}^n\cdot\nabla)\mathbf{v}^n\Big]\Big) = \mu\Delta\mathbf{v}^{n+1} + (\mathbf{T}^{n+1} : \partial^2)\mathbf{v}^{n+1} \\ &\quad + (\nabla\mathbf{v}^n + (\nabla\mathbf{v}^n)^T)\nabla p^{n+1} - \nabla q^{n+1} + \frac{\partial \mathbf{f}}{\partial t} + (\mathbf{v}^n\cdot\nabla)\mathbf{f} - (\nabla\mathbf{v}^n)\mathbf{f} + \lambda\mathbf{f} \end{aligned} \tag{16}$$

with the incompressibility condition (2) and initial and boundary conditions for $\mathbf{v}^{n+1}$. Here the outflow condition is implemented in the form

$$\left(\frac{\partial^2 \mathbf{v}^{n+1}}{\partial t^2} + (\mathbf{v}^n \cdot \nabla)\frac{\partial \mathbf{v}^{n+1}}{\partial t} - \left(\frac{\partial \mathbf{v}^{n+1}}{\partial t} \cdot \nabla\right)\mathbf{v}^n\right)(L, y, t) = \mathbf{w}_{\text{out}}(y, t). \tag{17}$$

4. Statement of result. In the following, we use the notation H_p^k for spaces of functions which have k locally square integrable derivatives and are periodic with period M in the unbounded direction. Moreover, we shall, in abuse of proper notation, say that a vector or tensor valued quantity is in a function space X if all components are in X. We shall make the following assumptions:

(S1) $\mathbf{v}_0 \in H_p^5((0,L) \times \mathbb{R})$, $\mathbf{T}_0 \in H_p^5((0,L) \times \mathbb{R})$.

(S2) $\mathbf{v}_{\text{in}} \in H_p^{11/2}((0,T) \times \mathbb{R})$, $\mathbf{w}_{\text{out}} \in H_p^{7/2}((0,T) \times \mathbb{R})$, $u_{\text{out}} \in H_p^{7/2}(\mathbb{R})$. In addition, the first component of $\mathbf{v}_{\text{in}}$ is in $H^6((0,T); H_p^{-1/2}(\mathbb{R}))$ and the first component of $\mathbf{w}_{\text{out}}$ is in $H^{9/2}((0,T); H_p^{-1}(\mathbb{R}))$.

(S3) $\pi, \tau \in H_p^4((0,T) \times \mathbb{R}))$.

(S4) $\mathbf{f} \in \bigcap_{k=0}^5 W^{k,1}((0,T); H_p^{5-k}((0,L) \times \mathbb{R}))$.

(E) $\mathbf{T}_0 + \mu - \mathbf{v}_0\mathbf{v}_0^T$ is uniformly positive definite.

(C1) The initial values of $\mathbf{v}$ and its derivatives up to fourth order, p and its derivatives up to third order, and $\mathbf{T}$ and its derivatives up to second order are compatible with the prescribed boundary conditions.

(C2) div $\mathbf{v}_0 = 0$.

(C3)
$$\frac{d^2}{dt^2}\int_0^M \mathbf{v}_{\text{in}}(y,t) \cdot \mathbf{e}_1 \, dy = \int_0^M \mathbf{w}_{\text{out}} \cdot \mathbf{e}_1 \, dy,$$
$$\int_0^M \frac{\partial \mathbf{v}_{\text{in}}}{\partial t}(y,0) \cdot \mathbf{e}_1 \, dy = \int_0^M u_{\text{out}}(y) \, dy.$$

(I) $\mathbf{v}_{\text{in}} \cdot \mathbf{e}_1 > 0$, $\mathbf{v}_0 \cdot \mathbf{e}_1 > 0$.

In assumption (C1), the initial values of derivatives are to be computed from the equations. For example, (1), together with initial values of $\mathbf{v}$ and $\mathbf{T}$, yields an initial value of $\rho\partial\mathbf{v}/\partial t + \nabla p$. Using the divergence condition and the boundary data for the normal component of $\partial\mathbf{v}/\partial t$, we can calculate initial data of $\partial\mathbf{v}/\partial t$ and ∇p from this. At the inflow boundary, the tangential component of ∇p has to be compatible with π, and from the knowledge of π and ∇p, we can evaluate p. This procedure can be repeated for higher time derivatives.

Our result will be as follows.

THEOREM. *Under the assumptions above, there exists $T' > 0$ such that the initial-boundary value problem* (1)–(6) *has a solution with the regularity*

$$\begin{aligned} &\mathbf{v} \in H_p^5((0,T') \times (0,L) \times \mathbb{R}), \\ &p \in \bigcap_{i=0}^{3} C^i([0,T']; H_p^{4-i}((0,L) \times \mathbb{R})), \\ &\mathbf{T} \in \bigcap_{i=0}^{3} C^i([0,T']; H^{3-i}((0,L) \times \mathbb{R})). \end{aligned} \tag{18}$$

The assumption that $\mathbf{T}_0 \in H^5_p((0,L) \times \mathbb{R}$ may appear excessive in view of the regularity obtained for the solution. However, as in [3], this assumption is needed to ensure regularity of initial values for time derivatives of $\mathbf{v}$.

The proof is based on a contraction argument. We define $Z(M,T')$ to be the set of all functions $(\mathbf{v},q,p,\mathbf{T})$, defined on $(0,T')\times(0,L)\times\mathbb{R}$ with the following properties:

$$
\begin{aligned}
&\mathbf{v} \in H^5_p((0,T') \times (0,L) \times \mathbb{R}),\\
&q \in \bigcap_{i=0}^{3} H^i((0,T'); H^{4-i}_p((0,L) \times \mathbb{R})),\\
&q(0,\cdot,\cdot) \in W^{3,\infty}(0,T'); L^2_p(\mathbb{R})),\\
&p \in \bigcap_{i=0}^{3} W^{i,\infty}((0,T'); H^{4-i}_p((0,L) \times \mathbb{R})) \cap H^4((0,T'); L^2_p((0,L) \times \mathbb{R})),\\
&\mathbf{T} \in \bigcap_{i=0}^{3} W^{i,\infty}((0,T'); H^{3-i}_p((0,L) \times \mathbb{R})),\\
&\sum_{i=0}^{5} \|\mathbf{v}\|_{i,5-i,2} + \sum_{i=0}^{3} \|q\|_{i,4-i,2} + \sum_{i=0}^{3} \|p\|_{i,4-i,\infty} + \sum_{i=0}^{3} \|\mathbf{T}\|_{i,3-i,\infty} \le M,\\
&\operatorname{div} \mathbf{v} = 0,\ \mathbf{v}(0,y,t) = \mathbf{v}_{\text{in}}(y,t),\\
&\int_0^1 \int_0^M q(x,y,t)\,dy\,dx = 0,\\
&\text{The initial values of time derivatives up to the indicated orders}\\
&\quad\text{agree with those determined from the equations:}\\
&\mathbf{v}:\ 4,\ q:\ 2,\ p:\ 3,\ \mathbf{T}:\ 2.
\end{aligned}
\tag{19}
$$

Here $\|\cdot\|_{i,k,q}$ denotes the norm in $W^{i,q}((0,T'); H^k_p((0,L) \times \mathbb{R}))$. We shall show that the mapping defined by the iteration in §3 maps $Z(M,T')$ into itself if M is sufficiently large and T' is sufficiently small relative to M. Moreover, it is a contraction if $Z(M,T')$ is equipped with the norm

$$
\begin{aligned}
d((\mathbf{v},q,p,\mathbf{T}),(\hat{\mathbf{v}},\hat{q},\hat{p},\hat{\mathbf{T}})) &= \sum_{i=0}^{4} \|\mathbf{v} - \hat{\mathbf{v}}\|_{i,4-i,2} + \sum_{i=0}^{2} \|q - \hat{q}\|_{i,3-i,2}\\
&+ \|q(0,\cdot,\cdot) - \hat{q}(0,\cdot,\cdot)\|_{2,0,\infty} + \sum_{i=0}^{2} \|p - \hat{p}\|_{i,3-i,\infty} + \sum_{i=0}^{2} \|\mathbf{T} - \hat{\mathbf{T}}\|_{i,2-i,\infty}.
\end{aligned}
\tag{20}
$$

It is easy to see that $Z(M,T')$ is complete; we need to check that it is nonempty if M is large enough. The existence of functions which, along with their derivatives, satisfy the given initial conditions, follows from the inverse trace theorem ([1, p. 21]), and as a result, we can find functions q, p, and $\mathbf{T}$ which satisfy all the requirements above.

It remains to construct a velocity field which satisfies the inflow conditions, the divergence condition, and the initial conditions. It is easy to find a divergence-free vector field which has the required regularity and satisfies the inflow conditions. Therefore, without loss of generality, we can henceforth require that $\mathbf{v} = \mathbf{0}$ on the boundary. This is the case studied in [3], and the only difference is that we require an additional order of regularity.

Our task is now to find a divergence-free $\mathbf{v}$ such that $\mathbf{v} = \mathbf{0}$ on the boundary and

$$\frac{\partial^i \mathbf{v}}{\partial t^i}(x, y, 0) = \mathbf{v}_i(x, y), \quad i = 0, 1, 2, 3, 4, \tag{21}$$

where $\mathbf{v}_i \in H_p^{5-i}((0, L) \times \mathrm{I\!R})$ is divergence-free and vanishes on the boundary. Following the ideas of [3], we take a_1 and a_2 to be functions such that

$$a_i = \frac{\partial a_i}{\partial x} = \frac{\partial^2 a_i}{\partial x^2} = 0, \quad -\frac{\partial^3 a_i}{\partial x^3} = (\mathcal{S}\mathbf{v}_i) \cdot \mathbf{e}_2, \quad \frac{\partial^4 a_1}{\partial x^4} = 0 \tag{22}$$

on the boundary, where $\mathcal{S}$ is the Stokes operator. The trace theorem shows that such functions a_i exist and $a_i \in H_p^{6-i}((0, L) \times \mathrm{I\!R})$. Next, we decompose $\mathbf{v}$ in the form $\mathbf{v} = \mathbf{v}_0 + \hat{\mathbf{v}} + \tilde{\mathbf{v}}$, where $\mathbf{v}_0$ is the initial value from (21) and $\hat{\mathbf{v}}$ is required to satisfy the initial conditions

$$\hat{\mathbf{v}}(x, y, 0) = \mathbf{0}, \quad \frac{\partial \hat{\mathbf{v}}}{\partial t}(x, y, 0) = \mathbf{v}_1(x, y) - \operatorname{curl} a_1, \quad \frac{\partial^2 \hat{\mathbf{v}}}{\partial t^2}(x, y, 0) = \mathbf{v}_2(x, y) - \operatorname{curl} a_2,$$

$$\frac{\partial^3 \hat{\mathbf{v}}}{\partial t^3}(x, y, 0) = \mathbf{v}_3(x, y), \quad \frac{\partial^4 \hat{\mathbf{v}}}{\partial t^4}(x, y, 0) = \mathbf{v}_4(x, y). \tag{23}$$

Here the curl operator is defined by

$$\operatorname{curl} a = \frac{\partial a}{\partial y}\mathbf{e}_1 - \frac{\partial a}{\partial x}\mathbf{e}_2. \tag{24}$$

It is easily verified that $\mathbf{v}_4 \in D((-\mathcal{S})^{1/2})$, $\mathbf{v}_3 \in D(\mathcal{S})$, $\mathbf{v}_2 - \operatorname{curl} a_2 \in D((-\mathcal{S})^{3/2})$, and $\mathbf{v}_1 \in D(\mathcal{S}^2)$. The inverse trace theorem ([1, p. 21]) shows that a divergence-free vector field $\hat{\mathbf{v}}$ with the required initial conditions and the regularity

$$\hat{\mathbf{v}} \in \bigcap_{i=0}^{5} C^i([0, T']; D((-\mathcal{S})^{(5-i)/2}) \tag{25}$$

exists. The final contribution $\tilde{\mathbf{v}}$ is constructed as $\operatorname{curl} \tilde{a}$, were $\tilde{a}$ has to satisfy

$$\tilde{a}(x, y, 0) = 0, \quad \frac{\partial \tilde{a}}{\partial t}(x, y, 0) = a_1(x, y), \quad \frac{\partial^2 \tilde{a}}{\partial t^2}(x, y, 0) = a_2(x, y),$$

$$\frac{\partial^3 \tilde{a}}{\partial t^3}(x, y, 0) = 0, \quad \frac{\partial^4 \tilde{a}}{\partial t^4} = 0. \tag{26}$$

From the conditions (22), it follows that $a_i \in D(\mathcal{B}^{(6-i)/4})$, where $\mathcal{B}$ is the biharmonic operator with Dirichlet boundary conditions. Invoking the inverse trace theorem again, we find that there exists $\tilde{a}$ satisfying (26) with the regularity

$$\tilde{a} \in \bigcap_{i=0}^{5} C^i([0, T']; D(\mathcal{B}^{(6-i)/4})). \tag{27}$$

This completes the argument.

5. Stress integration. We now need to provide the estimates which will establish that the iteration defined in §3 gives a contraction on $Z(M,T')$. We note that if T' is chosen small enough relative to M, then (I) implies that $\mathbf{v}\cdot\mathbf{e}_1 > 0$ for any element of $Z(M,T')$. We shall assume throughout that this is the case.

Let us assume that $\mathbf{v}^n$, q^n, p^n, and $\mathbf{T}^n$ are given. In this section, we shall obtain estimates for p^{n+1} and $\mathbf{T}^{n+1}$. The equations involved in determining these are easily solved, e.g., by the method of characteristics, and we shall obtain energy estimates for the solutions. For convencience, we make the following definition.

DEFINITION. *A function $\phi(M,T')$ is called* controllable *if there exists a positively valued function $T'(M)$ and a constant C such that $\phi(M,T'(M)) \le C$ for sufficiently large M.*

In general, of course, $T'(M)$ will tend to zero as $M \to \infty$.

We start with (10). We multiply both side by p^{n+1} and integrate over $(0,t)\times\Omega$. This yields

$$\begin{aligned}&\frac{1}{2}\int_\Omega (p^{n+1}(\mathbf{x},t))^2\,dx\,dy + \lambda\int_0^t\int_\Omega (p^{n+1}(\mathbf{x},\tau))^2\,dx\,dy\,d\tau\\ &= \int_0^t\int_\Omega p^{n+1}(\mathbf{x},\tau)q^n(\mathbf{x},\tau)\,dx\,dy\,d\tau + \frac{1}{2}\int_\Omega p_0(\mathbf{x})^2\,dx\,dy\\ &+\frac{1}{2}\int_0^t\int_0^M \mathbf{v}_{\text{in}}(y,\tau)\cdot\mathbf{e}_1\pi(y,\tau)^2\,dy\,d\tau\\ &-\frac{1}{2}\int_0^t\int_0^M \mathbf{v}(L,y,\tau)\cdot\mathbf{e}_1(p^{n+1}(L,y,\tau))^2\,dy\,d\tau.\end{aligned}\tag{28}$$

This yields the estimate

$$\|p^{n+1}\|_{0,0,\infty} \le C(\sqrt{T'}\|q^n\|_{0,0,2} + \|p_0\|_0 + \|\pi\|_0).\tag{29}$$

Here $\|\cdot\|_k$ with a single index k shall denote the norm in $H^k(\Omega)$ or $H^k_p((0,T')\times\mathbb{R})$, respectively.

We can now differentiate (10) and obtain estimates for derivatives of p^{n+1} in the same fashion; note that the value of the x-derivative of p^{n+1} at the inflow boundary is determined by the equation itself in terms of given data and the value of q^n at the inflow boundary. By proceeding in this fashion, we find an estimate of the form

$$\sum_{i=0}^{3}\|p^{n+1}\|_{i,4-i,\infty} \le \phi(M,T'),\tag{30}$$

where the function $\phi(M,T')$ is controllable. (Note that the right-hand side of (29), for instance, has a bound of the form $C_1M\sqrt{T'}+C_2$, which is obviously controllable by setting $T'=1/M^2$.)

In a similar fashion, the solution of (13) yields controllable bounds for $\|\mathbf{T}\|_3$ on the inflow boundary, and this can be used in (15) to get a controllable bound for $\sum_{i=0}^3\|\mathbf{T}\|_{i,3-i,\infty}$.

We have derived all the estimates for stress integration which are needed to show that the iteration maps $Z(M,T')$ into itself. The contraction estimates are derived in the same fashion, and we outline only one step. Consider two iteration sequences for equation (10) and take the difference. We get

$$\left(\frac{\partial}{\partial t} + (\mathbf{v}^n\cdot\nabla)\right)(p^{n+1}-\hat{p}^{n+1}) + \lambda(p^{n+1}-\hat{p}^{n+1}) = q^n - \hat{q}^n - ((\mathbf{v}^n-\hat{\mathbf{v}}^n)\cdot\nabla)\hat{p}^{n+1},\tag{31}$$

with initial and inflow values of $p^{n+1} - \hat{p}^{n+1}$ equal to zero. Proceeding as above, we obtain from (31) the estimate

$$\begin{aligned}(32)\quad &\sum_{i=0}^{2} \|p^{n+1} - \hat{p}^{n+1}\|_{i,3-i,\infty} + \|p - \hat{p}\|_{3,0,2} \le C\Big(\sum_{i=0}^{2}(\|q^n - \hat{q}^n\|_{i,3-i,1} \\ &\qquad + \|q^n(0,\cdot,\cdot) - \hat{q}^n(0,\cdot,\cdot)\|_{2,0,2}) + M\sum_{i=0}^{4} \|\mathbf{v}^n - \hat{\mathbf{v}}^n\|_{i,4-i,1}\Big) \\ &\le C\sqrt{T'}\Big(\sum_{i=0}^{3}(\|q^n - \hat{q}^n\|_{i,3-i,2} + \|q^n - \hat{q}^n\|_{2,0,\infty}) \\ &\qquad + M\sum_{i=0}^{4} \|\mathbf{v}^n - \hat{\mathbf{v}}^n\|_{i,4-i,2}\Big).\end{aligned}$$

Similar considerations apply to the integration of $\mathbf{T}$.

6. The velocity equation. We now consider the solution of the equation (16). To simplify notation, we shall write $\mathbf{v}$ and q for $\mathbf{v}^{n+1}$ and q^{n+1}, $\mathbf{w}$ for $\mathbf{v}^n$, and $\mathbf{T}$ for $\mathbf{T}^{n+1}$. The problem now assumes the form

$$(33)\qquad \rho\Big(\frac{\partial^2 \mathbf{v}}{\partial t^2} + 2(\mathbf{w}\cdot\nabla)\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{w}\cdot\nabla)^2\mathbf{v}\Big) = \mu\Delta\mathbf{v} + \mathbf{T} : \nabla^2\mathbf{v} - \nabla q + \tilde{\mathbf{f}},$$

where $\tilde{\mathbf{f}}$ is a known forcing function. In addition, we have to satisfy the divergence condition

$$(34)\qquad \operatorname{div} \mathbf{v} = 0,$$

the boundary conditions

$$(35)\quad \mathbf{v}(0,y,t) = \mathbf{v}_{\text{in}}(y,t), \qquad \Big(\frac{\partial^2 \mathbf{v}}{\partial t^2} + (\mathbf{w}\cdot\nabla)\frac{\partial \mathbf{v}}{\partial t} - \Big(\frac{\partial \mathbf{v}}{\partial t}\cdot\nabla\Big)\mathbf{w}\Big)(L,y,t) = \mathbf{w}_{\text{out}}(y,t),$$

and the initial conditions

$$(36)\qquad \mathbf{v}(x,y,0) = \mathbf{v}_0(x,y), \qquad \frac{\partial \mathbf{v}}{\partial t}(x,y,0) = \mathbf{v}_1(x,y).$$

Our goal for this section is the following result.

LEMMA. *The problem* (33)–(36) *has a unique solution, which satisfies an estimate of the form*

$$\begin{aligned}(37)\quad &\sum_{i=0}^{5} \|\mathbf{v}\|_{i,5-i,2} + \sum_{i=0}^{3} \|q\|_{i,4-i,2} \le C\Big(\sum_{i=0}^{3} \|\tilde{\mathbf{f}}\|_{i,3-i,\infty} + \sum_{i=0}^{2}\Big\|\frac{\partial^2 \tilde{\mathbf{f}}}{\partial t^2} + (\mathbf{w}\cdot\nabla)\frac{\partial \tilde{\mathbf{f}}}{\partial t}\Big\|_{i,2-i,1} \\ &\qquad + \|\mathbf{v}_0\|_5 + \|\mathbf{v}_1\|_4 + \|\mathbf{v}_{\text{in}}\| + \|\mathbf{w}_{\text{out}}\|\Big).\end{aligned}$$

Here the norms of $\mathbf{v}_{\text{in}}$ *and* $\mathbf{w}_{\text{out}}$ *are in the function spaces specified in assumption* (S2), *and the constant* C *depends only the initial data and on*

$$(38)\qquad \sum_{i=0}^{5} \|\mathbf{w}\|_{i,5-i,1} + \sum_{i=0}^{3} \Big(\|\mathbf{T}\|_{i,3-i,\infty} + \Big\|\frac{\partial \mathbf{T}}{\partial t} + (\mathbf{w}\cdot\nabla)\mathbf{T}\Big\|_{i,3-i,1}\Big).$$

Moreover, in any subdomain away from the outflow boundary, we obtain a bound with the same right-hand side as (37) *for*

$$\sum_{i=0}^{5} \|\mathbf{v}\|_{i,5-i,\infty} + \sum_{i=0}^{3} \|q\|_{i,4-i,\infty}.$$

Since the right-hand side in (37) is easily shown to be controllable, the lemma allows us to complete the proof that the iteration maps $Z(M, T')$ into itself as long as T' is small relative to M. Moreover, to show the contraction estimate, we can, similarly as in the previous section, take the difference between two problems of the form (33)–(36) and use the analogue of the preceding lemma at one less order of differentiability to estimate the difference between the solutions. We omit the details of this rather routine argument and devote the rest of the paper to proving the lemma.

We first construct a divergence-free $\mathbf{v}$ which satisfies the boundary conditions. To satisfy the divergence condition, we express $\mathbf{v}$ in terms of a streamfunction: $\mathbf{v} = (-\psi_y, \psi_x)$. Moreover, let ϕ denote the corresponding streamfunction for $\mathbf{w}$. The function ψ is the sum of a part which is linear in y and independent of x and a part which is periodic in y. The coefficient of the linear part is simply given by the average the first component of $\mathbf{v}_{\text{in}}$ and is hence of class $H^6(0,T)$. We can therefore focus on the periodic part. At the inflow boundary, we have prescribed values for ψ_y and ψ_x, and we can integrate with respect to y to obtain prescribed values of $\psi \in H^6((0,T); H_p^{1/2}(\mathbb{R})) \cap L^2((0,T); H_p^{13/2}(\mathbb{R}))$ and $\psi_x \in H_p^{11/2}((0,T) \times \mathbb{R})$. Using the inverse trace theorem, we can construct a ψ satisfying these boundary conditions which lies in $H^6((0,T); H_p^1((0,L) \times \mathbb{R})) \cap L^2((0,T); H_p^7((0,L) \times (\mathbb{R}))$. On the outflow boundary the prescribed data are $-\psi_{ytt} + \frac{\partial}{\partial y}(\phi_y \psi_{xt} - \phi_x \psi_{yt})$ and $\psi_{xtt} - \phi_y \psi_{xxt} + \phi_x \psi_{xyt} + \psi_{yt} \phi_{xx} - \psi_{xt} \phi_{xy}$. To find a function ψ satisfying these boundary conditions we arbitrarily set $\psi = 0$ on the boundary and then integrate the first boundary condition with respect to y. This leaves us with having to satisfy a given condition for $\psi_{xt} \in H_p^{9/2}((0,T) \times \mathbb{R})$, and from the second boundary condition, we can find $\psi_{xxt} \in H_p^{7/2}((0,T) \times \mathbb{R})$. Using the inverse trace theorem, we can find $\psi_t \in H_p^6((0,T) \times (0,L) \times \mathbb{R})$ so that the given boundary conditions are satisfied. We can thus construct a velocity field $\mathbf{v}^*$, satisfying the given boundary conditions, which has a time derivative of class H^5. We can now replace $\mathbf{v}$ in (33) by $\tilde{\mathbf{v}} + \mathbf{v}^*$ and absorb the terms resulting from $\mathbf{v}^*$ into the forcing function $\tilde{\mathbf{f}}$. For the rest of the section, we shall therefore assume homogeneous boundary conditions.

In order to construct solutions of (33), we shall take derivatives of the equation and reformulate it as a problem for higher derivatives of $\mathbf{v}$. In the process, we shall successively express lower-order time derivatives of $\mathbf{v}$ in terms of higher-order time derivatives. As a first step, let us take a time derivative of (33). We get

(39)
$$\begin{aligned}\rho(\mathbf{v}_{ttt} + 2(\mathbf{w} \cdot \nabla)\mathbf{v}_{tt} + 2(\mathbf{w}_t \cdot \nabla)\mathbf{v}_t + (\mathbf{w} \cdot \nabla)^2 \mathbf{v}_t) &= \mu \Delta \mathbf{v}_t + \mathbf{T} : \partial^2 \mathbf{v}_t - \nabla q_t + \tilde{\mathbf{f}}_t \\ &+ \mathbf{T}_t : \partial^2 \mathbf{v} - \rho(\mathbf{w} \cdot \nabla)(\mathbf{w}_t \cdot \nabla)\mathbf{v} - \rho(\mathbf{w}_t \cdot \nabla)(\mathbf{w} \cdot \nabla)\mathbf{v}.\end{aligned}$$

Moreover, $\mathbf{v}_t$ must be divergence free, vanish on the inflow boundary, satisfy the outflow condition (35), and satisfy the appropriate initial conditions. We would like to convert (39) to a problem which involves only $\mathbf{v}_t$ but not $\mathbf{v}$ itself. Of course, we could think of $\mathbf{v}$ as the time-integral of $\mathbf{v}_t$, but if we do so, then the spatial regularity of $\mathbf{v}$ implied by that of $\mathbf{v}_t$ is not sufficient to deal with the terms in (39). The remedy is to replace $\mathbf{v}$ in (39) by a new quantity $\hat{\mathbf{v}}$, which would equal $\mathbf{v}$ for a solution of (33).

Note that we can think of (33)–(35) as an elliptic boundary value problem for $\mathbf{v}$ if $\mathbf{v}_t$ is given. To make the outflow condition into a boundary condition for $\mathbf{v}$, we integrate it with respect to time:

$$(40)\quad \mathbf{v}_t + (\mathbf{w}\cdot\nabla)\mathbf{v} = (\mathbf{v}_t + (\mathbf{w}\cdot\nabla)\mathbf{v})(y,0) + \int_0^t [(\mathbf{w}_t\cdot\nabla)\mathbf{v} + (\mathbf{v}_t\cdot\nabla)\mathbf{w}](y,\tau)\,d\tau.$$

The idea now is to set $\hat{\mathbf{v}}$ equal to the solution of this elliptic boundary value problem. A minor problem is that the elliptic boundary value problem might not be uniquely solvable; however, it can be made uniquely solvable by a finite-rank perturbation. Let P be a finite-rank operator achieving this. Then we define $\hat{\mathbf{v}}$ as the solution of the problem

$$(41)\quad \rho(\mathbf{v}_{tt} + 2(\mathbf{w}\cdot\nabla)\mathbf{v}_t + (\mathbf{w}\cdot\nabla)^2)\hat{\mathbf{v}}) = \mu\Delta\hat{\mathbf{v}} + \mathbf{T} : \nabla^2\hat{\mathbf{v}} + P(\hat{\mathbf{v}} - \mathbf{v}) - \nabla\hat{q} + \tilde{\mathbf{f}},$$

with the additional conditions that $\hat{\mathbf{v}}$ is divergence free, vanishes at the inflow boundary, and satisfies

$$(42)\quad \mathbf{v}_t + (\mathbf{w}\cdot\nabla)\hat{\mathbf{v}} = (\mathbf{v}_t + (\mathbf{w}\cdot\nabla)\mathbf{v})(y,0) + \int_0^t [(\mathbf{w}_t\cdot\nabla)\hat{\mathbf{v}} + (\mathbf{v}_t\cdot\nabla)\mathbf{w}](y,\tau)\,d\tau$$

at the outflow boundary. Equation (39) is now modified to

$$(43)\quad \begin{aligned}\rho(\mathbf{v}_{ttt} + 2(\mathbf{w}\cdot\nabla)\mathbf{v}_{tt} + 2(\mathbf{w}_t\cdot\nabla)\mathbf{v}_t + (\mathbf{w}\cdot\nabla)^2\mathbf{v}_t) &= \mu\Delta\mathbf{v}_t + \mathbf{T} : \partial^2\mathbf{v}_t - \nabla q_t + \tilde{\mathbf{f}}_t\\ &+ \mathbf{T}_t : \partial^2\hat{\mathbf{v}} - \rho(\mathbf{w}\cdot\nabla)(\mathbf{w}_t\cdot\nabla)\hat{\mathbf{v}} - \rho(\mathbf{w}_t\cdot\nabla)(\mathbf{w}\cdot\nabla)\hat{\mathbf{v}}.\end{aligned}$$

We can choose P such that $P\mathbf{v}$ is of class $H_p^5((0,L)\times\mathbb{R})$, regardless of whether $\mathbf{v}$ itself has this regularity. For the solution of (41), we then have the estimate

$$(44)\quad \sum_{i=0}^{3}\|\hat{\mathbf{v}}\|_{i,5-i,2} + \|\hat{q}\|_{i,4-i,2} \le C\left(\sum_{i=0}^{4}\|\mathbf{v}_t\|_{i,4-i,2} + \sum_{i=0}^{3}\|\tilde{\mathbf{f}}\|_{i,3-i,2} + K\right),$$

where K is a term depending only on the initial data. We must make sure that the new problem (43) is equivalent to the original one. To this end, we take the time derivative of (41) and subtract (43). This yields

$$(45)\quad \mu\Delta(\hat{\mathbf{v}}_t - \mathbf{v}_t) + \mathbf{T} : \partial^2(\hat{\mathbf{v}}_t - \mathbf{v}_t) + P(\hat{\mathbf{v}}_t - \mathbf{v}_t) - \rho(\mathbf{w}\cdot\nabla)^2(\hat{\mathbf{v}}_t - \mathbf{v}_t) - \nabla(\hat{q}_t - q_t) = \mathbf{0}.$$

Moreover, the difference $\hat{\mathbf{v}}_t - \mathbf{v}_t$ has zero divergence and vanishes at the inflow boundary, and at the outflow boundary we have

$$(46)\quad (\mathbf{w}\cdot\nabla)(\hat{\mathbf{v}}_t - \mathbf{v}_t) = \mathbf{0}.$$

It now follows from the unique solvability of the elliptic boundary value problem that $\hat{\mathbf{v}}_t - \mathbf{v}_t = \mathbf{0}$, and since the initial data for $\mathbf{v}$ and $\hat{\mathbf{v}}$ agree, we have $\mathbf{v} = \hat{\mathbf{v}}$.

We shall solve (43) iteratively as follows:

$$(47)\quad \begin{aligned}\rho(\mathbf{v}^n_{ttt} + 2(\mathbf{w}\cdot\nabla)\mathbf{v}^n_{tt} + 2(\mathbf{w}_t\cdot\nabla)\mathbf{v}^{n-1}_t + (\mathbf{w}\cdot\nabla)^2\mathbf{v}^n_t) &= \mu\Delta\mathbf{v}^n_t + \mathbf{T} : \partial^2\mathbf{v}^n_t - \nabla q^n_t\\ &+ \tilde{\mathbf{f}}_t + \mathbf{T}_t : \partial^2\hat{\mathbf{v}}^{n-1} - \rho(\mathbf{w}\cdot\nabla)(\mathbf{w}_t\cdot\nabla)\hat{\mathbf{v}}^{n-1} - \rho(\mathbf{w}_t\cdot\nabla)(\mathbf{w}\cdot\nabla)\hat{\mathbf{v}}^{n-1}.\end{aligned}$$

At each step, this problem is of a similar form as (33), namely, if we write $\mathbf{v}^*$ instead of $\mathbf{v}_t^n$, then (47) has the form

$$\rho(\mathbf{v}_{tt}^* + 2(\mathbf{w}\cdot\nabla)\mathbf{v}_t^* + (\mathbf{w}\cdot\nabla)^2\mathbf{v}^*) = \mu\Delta\mathbf{v}^* + \mathbf{T} : \nabla^2\mathbf{v}^* - \nabla q^* + \mathbf{f}^*, \tag{48}$$

with a new forcing function $\mathbf{f}^*$. The boundary conditions are $\mathbf{v}^* = \mathbf{0}$ at the inflow boundary and $\mathbf{v}_t^* + (\mathbf{w}\cdot\nabla)\mathbf{v}^* - (\mathbf{v}^*\cdot\nabla)\mathbf{w} = \mathbf{0}$ at the outflow boundary. We shall now consider (48) with these boundary conditions and prove the following estimate, which is analogous to (37):

$$\begin{aligned}&\sum_{i=0}^{4}\|\mathbf{v}^*\|_{i,4-i,2} + \sum_{i=0}^{2}\|q^*\|_{i,3-i,2}\\ &\quad\le C\Big(\sum_{i=0}^{2}\|\tilde{\mathbf{f}}^*\|_{i,2-i,\infty} + \sum_{i=0}^{2}\Big\|\frac{\partial\mathbf{f}^*}{\partial t} + (\mathbf{w}\cdot\nabla)\mathbf{f}^*\Big\|_{i,2-i,1} + \|\mathbf{v}_0^*\|_4 + \|\mathbf{v}_1^*\|_3\Big).\end{aligned} \tag{49}$$

In conjuction with (44), this estimate can be used to show convergence of the iteration (47), and the solution then satisfies (33). Moreover, we shall show that the bound (49) also holds for

$$\sum_{i=0}^{4}\|\mathbf{v}^*\|_{i,4-i,\infty} + \sum_{i=0}^{2}\|q^*\|_{i,3-i,\infty}$$

in any subregion away from the outflow boundary. We can then use local elliptic estimates for (41) to get bounds on L^∞-type (in time) norms of $\hat{\mathbf{v}}$ and $\hat{q}$. Henceforth, we shall therefore consider (48) and aim to prove (49).

For the following, we need to introduce two functions $\alpha(x)$ and $\beta(x)$, which will be assumed to have the following properties for some positive value of ϵ:

$$\begin{aligned}&\alpha,\beta\in C^\infty([0,L];[0,1]),\quad \alpha(x)=1,\quad \beta(x)=0\quad\text{for}\quad x\in[L-\epsilon,L],\\ &\beta(x)=1,\quad \alpha(x)=0\quad\text{for}\quad x\in[0,\epsilon],\quad (1-\alpha)(1-\beta)=0.\end{aligned} \tag{50}$$

We now apply the operation $\frac{\partial}{\partial t} + (\alpha\mathbf{w}\cdot\nabla) + (\nabla(\alpha\mathbf{w}))^T$ to (48), and we introduce the new variable $\mathbf{u} = \mathbf{v}_t^* + \alpha(\mathbf{w}\cdot\nabla)\mathbf{v}^* - (\mathbf{v}^*\cdot\nabla)(\alpha\mathbf{w}) + \mathbf{v}^*\text{div }(\alpha\mathbf{w})$. The resulting equations are complicated, and we shall not write them out in full. However, we shall indicate their structure and emphasize the essential points. First, the left-hand side of (48) can be written in the form

$$\rho\big(\mathbf{u}_t + (2-\alpha)(\mathbf{w}\cdot\nabla)\mathbf{u} + (1-\alpha(2-\alpha))(\mathbf{w}\cdot\nabla)^2\mathbf{v}^* + \mathbf{h}_1(\mathbf{v}^*,\nabla\mathbf{v}^*,\mathbf{u})\big), \tag{51}$$

where we do not write out the form of the term $\mathbf{h}_1$; the essential point is that it does not involve second derivatives of $\mathbf{v}^*$ or any derivatives of $\mathbf{u}$. We use (51) and apply the operation $\frac{\partial}{\partial t} + \alpha(\mathbf{w}\cdot\nabla) + (\nabla(\alpha\mathbf{w}))^T$ to (48). Wherever we encounter a time derivative of $\mathbf{v}^*$, we reexpress it in terms of $\mathbf{u}$ and spatial derivatives of $\mathbf{v}^*$. Doing so, we find an equation of the form

$$\begin{aligned}\rho(\mathbf{u}_{tt} + 2(\mathbf{w}\cdot\nabla)\mathbf{u}_t + (\mathbf{w}\cdot\nabla)^2\mathbf{u}) &= \mu\Delta\mathbf{u} + \mathbf{T} : \partial^2\mathbf{u} - \nabla s\\ &\quad + \mathbf{h}_2(\mathbf{u},\mathbf{u}_t,\nabla\mathbf{u},\mathbf{v}^*,\nabla\mathbf{v}^*,\nabla^2\mathbf{v}^*,\mathbf{f}^*,\mathbf{f}_t^* + \alpha(\mathbf{w}\cdot\nabla)\mathbf{f}^*).\end{aligned} \tag{52}$$

Here $s = q_t^* + \alpha(\mathbf{w}\cdot\nabla)q^*$. Note that $\mathbf{u}$ has been defined in such a way that it is divergence free and vanishes on the boundary. Following similar ideas as above, we shall reduce our problem to showing that (52) has a unique solution satisfying an estimate of the form

(53)
$$\sum_{i=0}^{3} \|\mathbf{u}\|_{i,3-i,\infty} + \sum_{i=0}^{1} \|s\|_{i,2-i,\infty} \\ \le C \left(\sum_{i=0}^{1} \|\mathbf{h}_2\|_{i,1-i,\infty} + \sum_{i=0}^{1} \left\| \frac{\partial \mathbf{h}_2}{\partial t} + \beta(\mathbf{w}\cdot\nabla)\mathbf{h}_2 \right\|_{i,1-i,1} + \|\mathbf{u}_0\|_3 + \|\mathbf{u}_1\|_2 \right).$$

Here, as usual, $\mathbf{u}_0$ and $\mathbf{u}_1$ are the initial conditions for $\mathbf{u}$ and $\mathbf{u}_t$. In order to use (53) to solve (52) in an iterative fashion, we must, in an analogous fashion as above, replace $\mathbf{v}^*$ by a different quantity $\hat{\mathbf{v}}^*$. However, this step is now more complicated because we cannot simply use (48) to solve for $\mathbf{v}^*$ in terms of $\mathbf{u}$. This is because we lack a condition for $\mathbf{v}^*$ at the outflow boundary.

To construct $\hat{\mathbf{v}}^*$, we shall, of course, start with (48), which we can rewrite in the form

$$(54)\qquad \rho\big(\mathbf{u}_t + (2-\alpha)(\mathbf{w}\cdot\nabla)\mathbf{u} + (1-\alpha(2-\alpha))(\mathbf{w}\cdot\nabla)^2\mathbf{v}^* + \mathbf{h}_1(\mathbf{v}^*,\nabla\mathbf{v}^*,\mathbf{u})\big) \\ = \mu\Delta\mathbf{v}^* + \mathbf{T}:\partial^2\mathbf{v}^* - \nabla q^* + \mathbf{f}^*.$$

We have the inflow boundary condition $\mathbf{v}^* = \mathbf{0}$ but no outflow condition. To create an outflow condition, we consider (54) for $t = T'$. In addition, we consider (54) for $y = L$, and make the substitution

$$(55)\qquad \mathbf{v}_x^* = \frac{1}{w_1}(\mathbf{u} - \mathbf{v}_t^* - w_2\mathbf{v}_y^* + (\mathbf{v}^*\cdot\nabla)\mathbf{w}).$$

That is, the x-derivative of $\mathbf{v}^*$ has been expressed in terms of $\mathbf{u}$ and derivatives of $\mathbf{v}^*$ which are tangential to the (y,t)-plane. By differentiating (55), we can also reexpress the second derivative with respect to x.

For the pressure, we need to proceed in a slightly more complicated fashion. The reason is that (53) only provides estimates for spatial derivatives of s and not for the second time derivative of s. To circumvent this problem, let A be a positive definite self-adjoint operator in $L_p^2((0,L)\times\mathbb{R})$ which generates the interpolation scale of Sobolev spaces, i.e., the domain of A^n is $H_p^n((0,L)\times\mathbb{R})$. Let now r be the solution of

$$(56)\qquad r_t + Ar = s,\ r(0) = 0$$

and set $q^* = r + \phi$. Then ϕ satisfies the equation

$$(57)\qquad \phi_t + \alpha(\mathbf{w}\cdot\nabla)\phi = -\alpha(\mathbf{w}\cdot\nabla)r + Ar.$$

Moreover, we have the estimate

$$(58)\qquad \sum_{i=0}^{2} \|r\|_{i,3-i,2} \le C\sum_{i=0}^{1} \|s\|_{i,2-i,2}.$$

We now consider r as given in terms of s and introduce ϕ as a new variable. On the outflow boundary, we use the relationship

$$(59)\qquad \phi_x = \frac{1}{w_1}(Ar - (\mathbf{w}\cdot\nabla)r - \phi_t - w_2\phi_y)$$

to eliminate the derivative normal to the boundary. On the outflow boundary, we can therefore rewrite (54) into an equation in the (y,t)-plane which has the following form

$$\begin{aligned}\mu\Big(\mathbf{v}^*_{yy} + \frac{1}{w_1^2}(\mathbf{v}^*_{tt} + 2w_2\mathbf{v}^*_{yt} + w_2^2\mathbf{v}^*_{yy})\Big) + T_{yy}\mathbf{v}^*_{yy}\\ -\frac{2}{w_1}T_{xy}(\mathbf{v}^*_{yt} + w_2\mathbf{v}^*_{yy}) + \frac{1}{w_1^2}T_{xx}(\mathbf{v}^*_{tt} + 2w_2\mathbf{v}^*_{yt} + w_2^2\mathbf{v}^*_{yy})\\ + \frac{1}{w_1}(\phi_t - w_2\phi_y)\mathbf{e}_1 - \phi_y\mathbf{e}_2 = \mathbf{h}_3(\mathbf{u}, \mathbf{u}_t, \nabla\mathbf{u}, \mathbf{v}^*, \mathbf{v}^*_t, \nabla\mathbf{v}^*, Ar, \nabla r, \mathbf{f}^*).\end{aligned}\tag{60}$$

The divergence condition can be rewritten as

$$-\frac{1}{w_1}\left(\frac{\partial v_1^*}{\partial t} + w_2\frac{\partial v_1^*}{\partial y}\right) + \frac{\partial v_2^*}{\partial y} = -\frac{1}{w_1}(u_1 + (\mathbf{v}^*\cdot\nabla)w_1).\tag{61}$$

We now consider an elliptic boundary value problem which consists of (54) and the condition div $\mathbf{v}^* = 0$ on the set $\{T'\}\times(0,L)\times\mathbb{R}$, (60) and (61) on the set $(0,T')\times\{L\}\times\mathbb{R}$, the boundary condition $\mathbf{v}^* = \mathbf{0}$ on the set $\{T'\}\times\{0\}\times\mathbb{R}$, the boundary condition $\mathbf{v}^* = \mathbf{v}^*(L,y,0)$ (given in terms of the initial data) on the set $\{0\}\times\{L\}\times\mathbb{R}$, and interface conditions on the set $\{T'\}\times\{L\}\times\mathbb{R}$. These interface conditions are the continuity of $\mathbf{v}^*$ and ϕ and the equation

$$\mathbf{u} = \mathbf{v}^*_t + (\mathbf{w}\cdot\nabla)\mathbf{v}^* - (\mathbf{v}^*\cdot\nabla)\mathbf{w}.\tag{62}$$

It can be checked that the equations are elliptic and the boundary and interface conditions satisfy the complementing condition. We change $\mathbf{v}^*$ to $\hat{\mathbf{v}}^*$, add a finite-rank perturbation to ensure unique solvability, and solve for $\hat{\mathbf{v}}^*$. This yields an estimate of the form

$$\begin{aligned}\|\hat{\mathbf{v}}^*\|_{7/2} + \|\phi\|_{5/2} \le C(\|\mathbf{u}\|_{5/2} + \|\nabla\mathbf{u}\|_{3/2} + \|\mathbf{u}_t\|_{3/2}\\ + \|Ar\|_{3/2} + \|\nabla r\|_{3/2} + \|\mathbf{f}^*\|_{3/2} + \|\mathbf{v}^*(L,\cdot,0)\|_3).\end{aligned}\tag{63}$$

Here the Sobolev norms refer to the sets $\{T'\}\times(0,L)\times\mathbb{R}$ and $(0,T')\times\{L\}\times\mathbb{R}$.

We now choose $\hat{\mathbf{v}}^*$ at the outflow boundary to be the value just obtained and then obtain $\hat{\mathbf{v}}^*$ elsewhere by solving (54) with inflow and outflow conditions. Again we modify by a finite-rank operator if necessary. The $\hat{\mathbf{v}}^*$ thus obtained now replaces $\mathbf{v}^*$ in (52). Over the whole domain $(0,T')\times(0,L)\times\mathbb{R}$, we now have an estimate of the form

$$\sum_{i=0}^{2}\|\hat{\mathbf{v}}^*\|_{i,4-i,2} \le C\left(\sum_{i=0}^{3}\|\mathbf{u}\|_{i,3-i,2} + \sum_{i=0}^{1}\|s\|_{i,2-i,2} + \sum_{i=0}^{2}\|\mathbf{f}^*\|_{i,2-i,2} + \|\mathbf{v}^*(\cdot,\cdot,0)\|_{7/2}\right),\tag{64}$$

and this suffices to solve (52) iteratively. (In any region away from the outflow boundary, we can use local elliptic estimates to get bounds on norms of $\mathbf{v}^*$ and q^* which are L^∞ with respect to time.) The proof that the new problem is equivalent to the old one proceeds similarly as above; one integrates (52) (in which $\mathbf{v}^*$ has been replaced by $\hat{\mathbf{v}}^*$) along lines $d\mathbf{x}/dt = \alpha\mathbf{w}$, and from the result one can derive an elliptic equation for $\mathbf{v}^*$. By comparing this equation with that for $\hat{\mathbf{v}}^*$, one can derive an elliptic problem satisfied by the difference $\mathbf{v}^* - \hat{\mathbf{v}}^*$, which is of an analogous form as the one from which we obtained $\hat{\mathbf{v}}^*$ above. Then one uses a uniqueness argument.

We can now concentrate on solving (52) and establishing the estimate (53). We first take two more time derivatives, and we continue with the practice of expressing lower-order time derivatives by higher-order time derivatives in terms of elliptic boundary value problems. Denoting $\mathbf{u}_{tt} = \mathbf{g}$, $s_{tt} = \pi$, we end up having to solve a problem of the form

$$\rho(\mathbf{g}_{tt} + 2(\mathbf{w}\cdot\nabla)\mathbf{g}_t + (\mathbf{w}\cdot\nabla)^2\mathbf{g}) = \mu\Delta\mathbf{g} + \mathbf{T} : \partial^2\mathbf{g} - \nabla\pi + \tilde{\mathbf{h}}_t + \hat{\mathbf{h}}, \tag{65}$$

subject to the conditions that $\mathbf{g}$ is divergence free, vanishes on the boundary, and satisfies given initial conditions $\mathbf{g} = \mathbf{g}_0 \in H^1_p((0,L)\times\mathrm{I\!R})$, $\mathbf{g}_t = \mathbf{g}_1 \in L^2_p(\mathrm{I\!R})$. Compatibility between initial and boundary conditions still holds. We need to prove an estimate of the form

(66)
$$\|\mathbf{g}\|_{0,1,\infty} + \|\mathbf{g}\|_{1,0,\infty} \leq C(\|\mathbf{g}_0\|_1 + \|\mathbf{g}_1\|_0 + \|\tilde{\mathbf{h}}\|_{0,0,\infty} + \|\tilde{\mathbf{h}}_t + \beta(\mathbf{w}\cdot\nabla)\tilde{\mathbf{h}}\|_{0,0,1} + \|\hat{\mathbf{h}}\|_{0,0,1}).$$

We shall first derive a formal energy estimate for (65) and then show how this energy estimate can be used to show existence when used in conjunction with a Galerkin argument. To obtain an energy estimate, we multiply (65) by $\mathbf{z} := \mathbf{g}_t + \beta(\mathbf{w}\cdot\nabla)\mathbf{g} - (\mathbf{g}\cdot\nabla)(\beta\mathbf{w}) + \mathbf{g}\,\mathrm{div}\,(\beta\mathbf{w})$ and integrate over $(0,t)\times(0,L)\times(0,M)$. We shall use the notation

$$[\mathbf{a},\mathbf{b}](t) := \int_0^t\int_0^L\int_0^M \mathbf{a}(x,y,\tau)\cdot\mathbf{b}(x,y,\tau)\,dy\,dx\,d\tau. \tag{67}$$

For the various terms in (63), we find the following, after a number of integrations by parts:

$$\begin{aligned}
[\mathbf{g}_{tt},\mathbf{z}] &= \frac{1}{2}\frac{d}{dt}\int_0^L\int_0^M |\mathbf{g}_t(x,y,t)|^2\,dy\,dx \\
&+ \frac{d}{dt}\int_0^L\int_0^M (\beta(\mathbf{w}\cdot\nabla)\mathbf{g})\cdot\mathbf{g}_t(x,y,t)\,dy\,dx \\
&+ \frac{1}{2}[\mathrm{div}\,(\beta\mathbf{w})\mathbf{g}_t,\mathbf{g}_t] - [\beta(\mathbf{w}_t\cdot\nabla)\mathbf{g},\mathbf{g}_t] \\
&- \frac{d}{dt}\int_0^L\int_0^M [(\mathbf{g}\cdot\nabla)(\beta\mathbf{w})]\cdot\mathbf{g}_t(x,y,t)\,dy\,dx \\
&+ [(\mathbf{g}_t\cdot\nabla)(\beta\mathbf{w}),\mathbf{g}_t] + [(\mathbf{g}\cdot\nabla)(\beta\mathbf{w}_t),\mathbf{g}_t] \\
&+ \frac{d}{dt}\int_0^L\int_0^M \mathrm{div}\,(\beta\mathbf{w})\mathbf{g}\cdot\mathbf{g}_t(x,y,t)\,dy\,dx \\
&- [\mathrm{div}\,(\beta\mathbf{w})\mathbf{g}_t,\mathbf{g}_t] - [\mathrm{div}\,(\beta\mathbf{w}_t)\mathbf{g},\mathbf{g}_t].
\end{aligned} \tag{68}$$

$$
\begin{aligned}
[(\mathbf{w}\cdot\nabla)\mathbf{g}_t,\mathbf{z}] = {} & \frac{1}{2}\frac{d}{dt}\int_0^L\int_0^M \beta|(\mathbf{w}\cdot\nabla)\mathbf{g}|^2(x,y,t)\,dy\,dx \\
& -[\beta(\mathbf{w}_t\cdot\nabla)\mathbf{g},(\mathbf{w}\cdot\nabla)\mathbf{g}] \\
& -\frac{d}{dt}\int_0^L\int_0^M((\mathbf{g}\cdot\nabla))(\beta\mathbf{w})\cdot((\mathbf{w}\cdot\nabla)\mathbf{g})\,dy\,dx \\
& +[(\mathbf{g}_t\cdot\nabla)(\beta\mathbf{w}),(\mathbf{w}\cdot\nabla)\mathbf{g}]+[(\mathbf{g}\cdot\nabla)(\beta\mathbf{w}_t),(\mathbf{w}\cdot\nabla)\mathbf{g}] \\
& +[(\mathbf{g}\cdot\nabla)\beta\mathbf{w},(\mathbf{w}_t\cdot\nabla)\mathbf{g}] \\
& +\frac{d}{dt}\int_0^L\int_0^M \operatorname{div}\,(\beta\mathbf{w})(\mathbf{g}\cdot(\mathbf{w}\cdot\nabla)\mathbf{g})(x,y,t)\,dy\,dx \\
& -[\operatorname{div}\,(\beta\mathbf{w}_t)\mathbf{g},(\mathbf{w}\cdot\nabla)\mathbf{g}]-[\operatorname{div}\,(\beta\mathbf{w})\mathbf{g}_t,(\mathbf{w}\cdot\nabla)\mathbf{g}] \\
& -[\operatorname{div}\,(\beta\mathbf{w})\mathbf{g},(\mathbf{w}_t\cdot\nabla)\mathbf{g}].
\end{aligned}
\tag{69}
$$

$$
[\nabla\pi,\mathbf{z}]=0. \tag{70}
$$

(71)

$$
\begin{aligned}
[\tilde{\mathbf{h}}_t,\mathbf{z}] = {} & [\tilde{\mathbf{h}}_t+\beta(\mathbf{w}\cdot\nabla)\tilde{\mathbf{h}},\mathbf{g}_t]-[\tilde{\mathbf{h}},\beta(\mathbf{w}_t\cdot\nabla)\mathbf{g}+\mathbf{g}\operatorname{div}\,(\beta\mathbf{w}_t)] \\
& +\frac{d}{dt}\int_0^L\int_0^M \tilde{\mathbf{h}}\cdot(\beta(\mathbf{w}\cdot\nabla)\mathbf{g}-(\mathbf{g}\cdot\nabla)(\beta\mathbf{w})+\mathbf{g}\operatorname{div}\,(\beta\mathbf{w}))(x,y,t)\,dy\,dx \\
& +[\tilde{\mathbf{h}},(\mathbf{g}_t\cdot\nabla)(\beta\mathbf{w})+(\mathbf{g}\cdot\nabla)(\beta\mathbf{w}_t)].
\end{aligned}
$$

To deal with the remaining terms, we use the notation

$$
\mathbf{A}=\mu+\mathbf{T}-\rho\mathbf{w}\mathbf{w}^T. \tag{72}
$$

Since the expression for $[\mathbf{A}:\partial^2\mathbf{g},\mathbf{z}]$ is somewhat lengthy, we list it in several parts. We find

(73)

$$
\begin{aligned}
[\mathbf{A}:\partial^2\mathbf{g},\mathbf{g}_t] = {} & -\frac{1}{2}\frac{d}{dt}\int_0^L\int_0^M \operatorname{tr}((\nabla\mathbf{g})\mathbf{A}(\nabla\mathbf{g})^T)\,dy\,dx \\
& +\frac{1}{2}\int_0^t\int_0^L\int_0^M \operatorname{tr}((\nabla\mathbf{g})\mathbf{A}_t(\nabla\mathbf{g})^T)\,dy\,dx\,d\tau-[((\operatorname{div}\,\mathbf{A})\cdot\nabla)\mathbf{g},\mathbf{g}_t],
\end{aligned}
$$

$$
\begin{aligned}
& [\mathbf{A}:\partial^2\mathbf{g},-(\mathbf{g}\cdot\nabla)(\beta\mathbf{w})+\mathbf{g}\operatorname{div}\,(\beta\mathbf{w})]= \\
& \qquad -[((\operatorname{div}\,\mathbf{A})\cdot\nabla)\mathbf{g},-(\mathbf{g}\cdot\nabla)(\beta\mathbf{w})+\mathbf{g}\operatorname{div}\,(\beta\mathbf{w})] \\
& \qquad -\int_0^t\int_0^L\int_0^M \operatorname{tr}[(\nabla\mathbf{g})\mathbf{A}\nabla(-(\mathbf{g}\cdot\nabla)(\beta\mathbf{w})+\mathbf{g}\operatorname{div}\,(\beta\mathbf{w}))^T]\,dy\,dx\,d\tau,
\end{aligned}
\tag{74}
$$

Finally, we consider the term

$$
[\mathbf{A}:\partial^2\mathbf{g},\beta(\mathbf{w}\cdot\nabla)\mathbf{g}]=-[\beta(\mathbf{w}\cdot\nabla)(\mathbf{A}:\partial^2\mathbf{g}),\mathbf{g}]-[\mathbf{A}:\partial^2\mathbf{g},\operatorname{div}\,(\beta\mathbf{w})\mathbf{g}]. \tag{75}
$$

In the second term, we integrate by parts again to find

$$
\begin{aligned}
[\mathbf{A}:\partial^2\mathbf{g},\operatorname{div}\,(\beta\mathbf{w})\mathbf{g}] = {} & -\sum_{i,j}\left[A_{ij}\frac{\partial\mathbf{g}}{\partial x_j},\frac{\partial}{\partial x_i}(\operatorname{div}\,(\beta\mathbf{w})\mathbf{g})\right] \\
& -\sum_{i,j}\left[\frac{\partial A_{ij}}{\partial x_i}\frac{\partial\mathbf{g}}{\partial x_j},\operatorname{div}\,(\beta\mathbf{w})\mathbf{g}\right].
\end{aligned}
\tag{76}
$$

For the first term, integration by parts yields

$$(77)\quad \begin{aligned}-[\beta(\mathbf{w}\cdot\nabla)(\mathbf{A}:\partial^2\mathbf{g}),\mathbf{g}] &= \sum_{i,j}\left[\beta(\mathbf{w}\cdot\nabla)(A_{ij}\frac{\partial\mathbf{g}}{\partial x_j}),\frac{\partial\mathbf{g}}{\partial x_i}\right]\\ &\quad+\sum_{i,j}\left[\left(\frac{\partial(\beta\mathbf{w})}{\partial x_i}\cdot\nabla\right)\left(A_{ij}\frac{\partial\mathbf{g}}{\partial x_j}\right),\mathbf{g}\right]+\left[\beta(\mathbf{w}\cdot\nabla)\left(\frac{\partial A_{ij}}{\partial x_i}\frac{\partial\mathbf{g}}{\partial x_j}\right),\mathbf{g}\right].\end{aligned}$$

In the last two terms, we can do further integrations by parts to obtain expressions which are quadratic in the first derivatives of $\mathbf{g}$. For the first term on the right of (77), we note that

$$(78)\quad \frac{1}{2}\sum_{i,j}\beta(\mathbf{w}\cdot\nabla)\left(A_{ij}\frac{\partial\mathbf{g}}{\partial x_i}\cdot\frac{\partial\mathbf{g}}{\partial x_j}\right)=\sum_{i,j}\beta(\mathbf{w}\cdot\nabla)\left(A_{ij}\frac{\partial\mathbf{g}}{\partial x_j}\right)\cdot\frac{\partial\mathbf{g}}{\partial x_i}-\frac{1}{2}\frac{\partial\mathbf{g}}{\partial x_i}\cdot\frac{\partial\mathbf{g}}{\partial x_j}\beta(\mathbf{w}\cdot\nabla)A_{ij}.$$

From integrating (78), we therefore obtain an integral of

$$(79)\qquad \frac{1}{2}\sum_{i,j}\beta(\mathbf{w}\cdot\mathbf{n})A_{ij}\frac{\partial\mathbf{g}}{\partial x_i}\frac{\partial\mathbf{g}}{\partial x_j}$$

over the boundary. Since $\beta = 0$ on the outflow boundary, only the inflow boundary makes a contribution, and this contribution is negative.

From the formal energy estimate, we can easily derive a bound of the form (66). However, in deriving the energy estimate, we have of course done manipulations which are not justified by the regularity we expect of the solution. To make the argument rigorous, we use a Galerkin approximation. For technical reasons, we split (65) into two problems, one where the forcing term in the equation is as given and the initial data are zero and another where the forcing term is zero and the initial data are given. Obviously, the solution of (65) is then obtained by adding the solutions of the two subproblems. Let us first deal with the problem for zero initial data. We define a space

$$(80)\qquad X=\{\mathbf{g}\in H^1_p((0,L)\times\mathbb{R})\mid \operatorname{div}\mathbf{g}=0,\ \mathbf{g}(0,\cdot)=\mathbf{g}(L,\cdot)=\mathbf{0}\}.$$

Let ϕ^n, $n\in\mathbb{N}$, be a basis for X, i.e., a linearly independent set with a dense linear span. We can take the ϕ_n such that they vanish in a neighborhood of the boundary and are of class C^∞. We now seek Galerkin approximations such that

$$(81)\qquad \mathbf{g}^N_t+\beta(\mathbf{w}\cdot\nabla)\mathbf{g}_N-(\mathbf{g}^N\cdot\nabla)\mathbf{w}+\mathbf{g}^N\operatorname{div}(\beta\mathbf{w})=\sum_{i=1}^N c_i(t)\phi^i(x,y),$$

and $\mathbf{g}^N$ is determined from integrating (81) subject to zero initial and inflow conditions. We require the equation

$$(82)\qquad (\rho(\mathbf{g}^N_{tt}+2(\mathbf{w}\cdot\nabla)\mathbf{g}^N_t+(\mathbf{w}\cdot\nabla)^2\mathbf{g}^N)-\mu\Delta\mathbf{g}^N-\mathbf{T}:\partial^2\mathbf{g}^N-\tilde{\mathbf{h}}_t-\hat{\mathbf{h}},\phi^i)=0$$

to hold for every $i=1,\dots,N$. Here $(\cdot,\cdot)$ is the inner product in $L^2_p((0,L)\times\mathbb{R})$. The initial conditions are zero. For any fixed N, equation (82) is a system of Volterra integrodifferential equations which can be solved. By choosing the test function

$$(83)\qquad \mathbf{z}^N=\sum_{i=1}^N c_i(t)\phi^i(x,y),$$

one can repeat the energy estimates above (now there is sufficient regularity to justify the manipulations) and obtain uniform bounds for the approximate solutions $\mathbf{g}^N$. As usual, one can then extract a weakly-$*$ convergent subsequence, the limit of which yields the solution we seek.

For zero forcing terms and nonzero initial data, we need to do much less fancy footwork. We can obtain a much simpler energy estimate than above by simply multiplying (65) by $\mathbf{g}_t$. A standard Galerkin argument (along the lines of §8.2 in Chapter 3 of [1]) can then be used for a rigorous justification.

REFERENCES

[1] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications* I, Springer-Verlag, Berlin, Heidelberg, 1972.

[2] M. RENARDY, *Inflow boundary conditions for steady flows of viscoelastic fluids with differential constitutive laws,* Rocky Mountain J. Math., 18 (1988), pp. 445–453; Corrigendum, 19 (1989), p. 561.

[3] ———, *Local existence of solutions of the Dirichlet initial-boundary value problem for incompressible hypoelastic materials,* SIAM J. Math. Anal., 21 (1990), pp. 1369–1385.

[4] ———, *An alternative approach to inflow boundary conditions for Maxwell fluids in three space dimensions,* J. Non-Newtonian Fluid Mech., 36 (1990), pp. 419–425.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 932–958, July 1996

003

A GEOMETRIC INTERPRETATION OF THE HEAT EQUATION WITH MULTIVALUED INITIAL DATA*

LAWRENCE C. EVANS†

Abstract. We utilize the level-set method to interpret geometrically what it means to solve the heat equation with multivalued initial data. We prove that in one space dimension, the limits of "geometrically natural" approximations instantly unfold multivalued initial data, according to an equal-area rule. In higher dimensions, the limits of certain "analytically natural" approximations display similar effects.

Key words. heat equation, level-set method, viscosity solutions

AMS subject classification. 35K05

1. Introduction.

A formal analysis of Burgers' equation,

$$u_t + uu_x = 0 \quad \text{in } \mathbb{R} \times (0, \infty), \tag{1.1}$$

obtained by tracking classical characteristics, suggests that a solution will in general become multivalued after a time. If, for instance, the initial data represent a mass of fluid centered at the origin, the solution u corresponds to a wave whose velocity at each point equals its height. As the higher parts of the wave consequently move faster than the lower, the wave will later "break" and "fold over." It is customary to reject such multivaluedness on physical grounds (cf. Whitham [25]) by accepting as true solutions u of (1.1) only those arising as limits when $\varepsilon \to 0^+$ of solutions u^ε to the "viscous" approximations

$$u_t^\varepsilon + u^\varepsilon u_x^\varepsilon = \varepsilon u_{xx}^\varepsilon \quad \text{in } \mathbb{R} \times (0, \infty). \tag{1.2}$$

The term $\varepsilon u_{xx}^\varepsilon$ forces the approximate solutions u^ε to remain smooth, and so any limit u is single valued, although in general discontinuous in regions of shock formation. If we imagine $\varepsilon > 0$ as fixed and think of the nonlinear term in (1.2) as a lower-order perturbation, the function u^ε is well behaved since the linear heat equation (a) smooths irregular initial data and, in particular, (b) keeps solutions from becoming multivalued.

This paper provides a further analysis and geometric interpretation of the effect (b) for the linear heat equation. We will show that a solution of the heat equation is single valued at times $t > 0$, even if the initial function is multivalued, with a graph that admits folds, complicated topology, etc. Otherwise stated, we assert that property (b) is so pronounced that not only will single-valued initial data remain so, but also multivalued data will instantly become single valued under the heat evolution.

Making sense of this claim requires that we first of all devise a way to solve the heat equation with multivalued starting data. This problem we will approach using the so-called "level-set method." The idea is to think of $\mathbb{R}^n \times (0, \infty)$ as being completely filled up with hypersurfaces, each of which represents a solution to the heat equation on $\mathbb{R}^{n-1} \times (0, \infty)$. We then regard these surfaces as being the level sets of a function $v : \mathbb{R}^n \times [0, \infty) \to \mathbb{R}$ and write down a nonlinear partial differential equation (PDE) that v verifies. We next attempt to analyze this nonlinear PDE rigorously and,

* Received by the editors October 12, 1994; accepted for publication February 10, 1995. This research was supported in part by National Science Foundation grant DMS-9203440.

† Department of Mathematics, University of California at Berkeley, Berkeley, CA 94720.

in particular, to prove that various approximate solutions v^ε converge to a limit v. We can then regard the various level sets of v as determining solutions of the heat equation, even starting from possibly multivalued initial data.

The formal calculations are as follows. Think of $v : \mathbb{R}^n \times [0,\infty) \to \mathbb{R}$ as a smooth function each of whose level sets, thought of as graphs in the x_n-direction, solve the heat equation. Suppose a portion of some such level set is—locally, at least—represented by the smooth graph

$$x_n = u(x',t) \qquad (t \geq 0, \ \ x' \in \mathbb{R}^{n-1}). \tag{1.3}$$

Then

$$v(x', u(x',t), t)$$

is constant in the variables $x' \in \mathbb{R}^{n-1}$ and $t \geq 0$. Differentiating, we deduce

$$v_{x_n} u_t + v_t = 0, \tag{1.4}$$

$$v_{x_i} + v_{x_n} u_{x_i} = 0 \qquad (1 \leq i \leq n-1), \tag{1.5}$$

$$\Delta' v + 2\sum_{i=1}^{n-1} v_{x_i x_n} u_{x_i} + v_{x_n x_n}|D'u|^2 + v_{x_n}\Delta' u = 0, \tag{1.6}$$

where $D' = (\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_{n-1}})$ and $\Delta' = \sum_{i=1}^{n-1} \frac{\partial^2}{\partial x_i^2}$ denote, respectively, the gradient and the Laplacian in the x'-variables, $x' = (x_1, \ldots, x_{n-1})$. Assuming that $v_{x_n} \neq 0$ and that u solves the heat equation

$$u_t = \Delta' u \qquad \text{in } \mathbb{R}^{n-1} \times (0,\infty),$$

we can simplify (1.6), using (1.4) and (1.5) to conclude

$$v_t = \Delta' v - \frac{2v_{x_i}}{v_{x_n}} v_{x_i x_n} + \frac{|D'v|^2}{v_{x_n}^2} v_{x_n x_n} \tag{1.7}$$

along the given level set of v, the implicit summation being for $i = 1$ to $n-1$. We now suppose that each level set of v represents the graph of a solution of the heat equation. The nonlinear PDE (1.7) then holds—formally, at least—everywhere in $\mathbb{R}^n \times (0,\infty)$.

Let us refer to (1.7) as the *level-surface heat equation.* Observe that (1.7) is degenerate parabolic and is undefined wherever $v_{x_n} = 0$.

Our plan hereafter is to study the initial value problem for the PDE (1.7) and to try to understand the behavior in time of the level sets of solutions; we informally regard these as defining the generalized heat flow. We will therefore be particularly interested in a "geometric" interpretation of the level surface heat equation, as this viewpoint will suggest natural approximation schemes.

This paper is structured so that §§2 and 3 recall the general theory of weak (that is, viscosity) solutions of "geometric" parabolic PDEs, following Chen, Giga, and Goto [8]. In §§4–6, we focus our attention on the case $n = 2$ and study the motion of the approximating level curves in the limit as $\varepsilon \to 0$. We prove in §5 that these curves rapidly unfold to become graphs. This assertion allows us to interpret the level-surface heat equation as "instantly unfolding" multivalued initial data, although the precise nature of this transformation remains unclear in general. We do, however, manage in §6 to show rigorously that in certain circumstances this process follows an equal-area construction.

The behavior of the level sets in the general case ($n \geq 3$) appears much more complicated. We are, in particular, unable to analyze carefully the asymptotic behavior of the level sets of the geometrically natural approximations (4.3_ε), and we propose instead in §7 to analyze certain analytically natural approximations (cf. (7.4_ε)). We prove that in the limit as $\varepsilon \to 0$, the level sets again become graphs. Further work is indicated here, as the limiting behavior of the level solution presumably entails some kind of interesting higher-dimensional analogue of the equal-area construction.

The key point of §§4–7 is the identification in the PDE (1.7) and in the approximations (2.5_ε) and (7.4_ε) of geometric and/or analytic mechanisms that force the instantaneous unfolding of level sets.

The general technique of studying nonlinear PDEs whose level sets evolve according to various geometric laws has in recent years proved extremely fruitful; see Osher and Sethian [23] for numerics and Chen, Giga, and Goto [8], Evans and Spruck [9]–[12], Soner [24], etc., for theory. The guiding insight for this paper, that the level-set method is also useful for other, nongeometric PDEs, is, in fact, very old: Caratheodory in his book [7, §49] describes a related method of Jacobi for investigating Hamilton–Jacobi PDEs. Recently, S. Osher has revived the technique in [22] within the context of image processing. His work is inspired by the paper by Bruckstein and Kimmel [5], and this paper is inspired by his. We hope also that the following study will be relevant in an investigation of a PDE/viscosity-solution approach to crystalline curvature motion, a la J. Taylor (cf. Cahn, Handwerker, and Taylor [6] and the references therein). The formal PDEs describing crystalline curvature motion involve various strong singularities, and the hope is that the level sets of solutions to smoother, approximating PDEs will in the limit instantly develop faces, evolving thereafter according to certain ordinary differential equations (ODE). This conjectured effect is presumably some kind of more complicated variant of the instant unfolding established here.

2. The level-surface heat equation: Geometrically natural approximations.

We commence our study of the initial value problem for the level-surface heat equation

$$(2.1)\qquad \begin{cases} v_t = \Delta' v - \frac{2v_{x_i}}{v_{x_n}} v_{x_i x_n} + \frac{|D'v|^2}{v_{x_n}^2} v_{x_n x_n} & \text{in } \mathbb{R}^n \times (0,\infty), \\ v = g & \text{in } \mathbb{R}^n \times \{t=0\} \end{cases}$$

by first noting that the PDE is of the general form

$$(2.2)\qquad v_t = F(D^2 v, Dv)$$

for F defined by

$$(2.3)\qquad F(R,p) = p_n^{-2}\left(p_n^2 \sum_{i=1}^{n-1} r_{ii} - 2\sum_{i=1}^{n-1} p_i p_n r_{in} + |p'|^2 r_{nn}\right),$$

where $R = ((r_{ij})) \in M^{n\times n}$, the space of $n \times n$ real matrices, and $p = (p', p_n) \in \mathbb{R}^n$, $p' = (p_1, \ldots, p_{n-1}) \in \mathbb{R}^{n-1}$, $p_n \neq 0$. The nonlinear term satisfies the structural identity

$$(2.4)\qquad F(\lambda R + \mu(p \otimes p), \lambda p) = \lambda F(R,p)$$

for all $\lambda, \mu \in \mathbb{R}$, $R \in M^{n\times n}$, $p \in \mathbb{R}^n$ with $p_n \neq 0$. The level surface heat equation is consequently "geometric" in the terminology of Chen, Giga, and Goto [8]. This means that the evolution in time of each level set of v depends only upon the geometry of that set and is unaffected by the behavior of neighboring level sets of v. (In studying parametric—that is, "geometric"—integrals in the calculus of variations, it is often useful to study associated nonparametric integrals, as, for instance, in Federer [13, §5.1.9]. Here we are in some sense reversing the customary procedure by transforming the (nongeometric) heat equation into the (geometric) level-surface heat equation.)

Although our PDE (2.1) verifies (2.4), it nevertheless does not fall within the scope of Chen, Giga, and Goto [8] since the nonlinear term is singular along the plane $\{p_n = 0\}$. Ishii [17] has recently extended the general theory to allow for singularities on such a set, but the level-set heat equation again fails to be covered because the coefficients $p_i p_n^{-1}$ and $|p'|^2 p_n^{-2}$ are unbounded near $\{p_n = 0\}$. See also Ohnuma and Sato [20]. We are, in fact, not able to devise a satisfactory notion of weak solution for (2.1). We will see, however, that this is not really the key issue: the point is that limits of solutions to approximating PDE evolve so that the level sets become graphs in the x_n-direction. The strong singularity, which precludes any invocation of [8], [17], or [20], forces this simplified behavior of the level sets.

It is therefore appropriate now to turn our attention to a well-behaved approximation scheme which is "geometrically natural."

Let us fix $\varepsilon > 0$ and consider instead of (2.1) the problem

$$(2.5_\varepsilon) \qquad \begin{cases} v_t^\varepsilon = \frac{(v_{x_n}^\varepsilon)^2 \Delta' v^\varepsilon - 2 v_{x_i}^\varepsilon v_{x_n}^\varepsilon v_{x_i x_n}^\varepsilon + |D' v^\varepsilon|^2 v_{x_n x_n}^\varepsilon}{(v_{x_n}^\varepsilon)^2 + \varepsilon^2 |D' v^\varepsilon|^2} & \text{in } \mathbb{R}^n \times (0,\infty), \\ v^\varepsilon = g & \text{on } \mathbb{R}^n \times \{t=0\}. \end{cases}$$

The PDE (2.5_ε) has the structure

$$v_t^\varepsilon = F^\varepsilon(D^2 v^\varepsilon, D v^\varepsilon)$$

for

$$F^\varepsilon(R,p) = (p_n^2 + \varepsilon^2 |p'|^2)^{-1} \left(p_n^2 \sum_{i=1}^{n-1} r_{ii} - 2 \sum_{i=1}^{n-1} p_i p_n r_{in} + |p'|^2 r_{nn} \right),$$

where $R = ((r_{ij})) \in M^{n \times n}$, $p = (p', p_n) \in \mathbb{R}^n$, $p \neq 0$. A computation verifies the identity

$$F^\varepsilon(\lambda R + \mu(p \otimes p), \lambda p) = \lambda F^\varepsilon(R,p) \qquad (\mu, \lambda \in \mathbb{R}, \ p \neq 0),$$

and so (2.5_ε) is geometric. And although the nonlinearity F^ε is not defined at $\{p = 0\}$, it, unlike F, is bounded on compact subsets of $M^{n \times n} \times (\mathbb{R}^n - \{0\})$. In particular, (2.5_ε) is included in the existence and uniqueness theory of Chen, Giga, and Goto [8]. We recall the relevant definitions.

DEFINITION 2.1. *A bounded, uniformly continuous function v^ε is a* weak subsolution (supersolution) *of (2.5_ε) provided that for each smooth $\phi \in C^\infty(\mathbb{R}^n \times (0,\infty))$ such that*

$$(2.6) \qquad \begin{aligned} &v^\varepsilon - \phi \textit{ attains a local maximum (minimum)} \\ &\textit{at a point } (x_0, t_0) \in \mathbb{R}^n \times (0,\infty), \end{aligned}$$

(a) *if*

$$(2.7) \qquad D\phi(x_0, t_0) \neq 0,$$

then

$$(2.8) \quad \phi_t - (\phi_{x_n}^2 + \varepsilon^2 |D'\phi|^2)^{-1} (\phi_{x_n}^2 \Delta' \phi - 2 \phi_{x_i} \phi_{x_n} \phi_{x_i x_n} + |D'\phi|^2 \phi_{x_n x_n}) \leq 0 \ \ (\geq 0)$$

at the point (x_0, t_0); and

(b) *if*

$$D\phi(x_0,t_0)=0, \qquad D^2\phi(x_0,t_0)=0, \tag{2.9}$$

then

$$\phi_t \le 0 \quad (\ge 0) \tag{2.10}$$

at the point (x_0,t_0).

DEFINITION 2.2. *We call* v^ε *a* weak solution *if* v^ε *is both a weak subsolution and supersolution.*

Observe that there is no requirement if

$$D\phi(x_0,t_0)=0, \qquad D^2\phi(x_0,t_0)\neq 0. \tag{2.11}$$

Let us now suppose $g:\mathbb{R}^n\to\mathbb{R}$ is smooth, $\sup_{\mathbb{R}^n}|g|$, $|Dg|<\infty$, and for some $R>0$,

$$\begin{cases} D'g=0 & \text{if } |x'|\ge R \text{ or } |x_n|\ge R, \\ g_{x_n}>0 & \text{if } |x_n|\le R,\ |x'|\ge R, \\ g_{x_n}=0 & \text{if } |x_n|\ge R. \end{cases} \tag{2.12}$$

In particular, the level sets of g are flat graphs in $\{|x_n|\ge R\}$ and $\{|x'|\ge R\}$.

THEOREM 2.3 (existence of approximate solutions).

(i) *For each* $\varepsilon>0$, *there exists a unique weak solution of* (2.5_ε).

(ii) *In addition,* v^ε *is Lipschitz continuous, with the bounds*

$$\sup_{\mathbb{R}^n\times(0,\infty)}|Dv^\varepsilon| \le \sup_{\mathbb{R}^n}|Dg|, \tag{2.13}$$

$$\sup_{\mathbb{R}^n\times(0,\infty)}|v^\varepsilon_t| \le \frac{C}{\varepsilon^2}\sup_{\mathbb{R}^n}|D^2g|. \tag{2.14}$$

(iii) *Furthermore, the mapping* $g\mapsto v^\varepsilon(\cdot\ ,t)$ *is a contraction in* $L^\infty(\mathbb{R}^n)$.

Proof. For $\delta>0$, we further approximate by the PDE

$$\begin{cases} v^{\varepsilon,\delta}_t = ((v^{\varepsilon,\delta}_{x_n})^2+\varepsilon^2|D'v^{\varepsilon,\delta}|^2+\delta)^{-1}((v^{\varepsilon,\delta}_{x_n})^2\Delta' v^{\varepsilon,\delta} - 2v^{\varepsilon,\delta}_{x_i}v^{\varepsilon,\delta}_{x_n}v^{\varepsilon,\delta}_{x_ix_n} \\ \qquad\qquad + |D'v^{\varepsilon,\delta}|^2 v^{\varepsilon,\delta}_{x_nx_n}) + \delta\Delta v^{\varepsilon,\delta} \quad \text{in } \mathbb{R}^n\times(0,\infty), \\ v^{\varepsilon,\delta}=g \qquad \text{on } \mathbb{R}^n\times\{t=0\}. \end{cases} \tag{$2.15_{\delta,\varepsilon}$}$$

This equation has the form

$$v^{\varepsilon,\delta}_t = a^{\varepsilon,\delta}_{ij}(Dv^{\delta,\varepsilon})v^{\varepsilon,\delta}_{x_ix_j} \qquad \text{in } \mathbb{R}^n\times(0,\infty), \tag{2.16}$$

the implicit summation in (2.16) being for $1\le i,j\le n$. The coefficients $a^{\varepsilon,\delta}_{ij}$ are smooth, bounded, and uniformly elliptic, and consequently the quasi-linear PDE (2.15) has a unique, smooth, bounded solution (cf. Ladyzhenskaya, Solonnikov, and Ural'ceva [18]).

Differentiating $(2.15_{\delta,\varepsilon})$ in the unit direction ξ, we find

$$v^{\varepsilon,\delta}_{\xi t} = a^{\varepsilon,\delta}_{ij}v^{\varepsilon,\delta}_{\xi x_ix_j} + a^{\varepsilon,\delta}_{ij,p_l}v^{\varepsilon,\delta}_{x_ix_j}v^{\varepsilon,\delta}_{\xi x_l}.$$

The maximum principle therefore implies that $v_\xi^{\varepsilon,\delta}$ attains its maximum at $t = 0$; whence

$$\sup_{\substack{0<\varepsilon,\delta\leq 1\\ \mathbb{R}^n\times(0,\infty)}} |Dv^{\varepsilon,\delta}| = \sup_{\substack{0<\varepsilon,\delta\leq 1\\ \mathbb{R}^n}} |Dv^{\varepsilon,\delta}|\,|_{t=0} = \sup_{\mathbb{R}^n} |Dg|, \tag{2.17}$$

$$\sup_{\substack{0<\delta\leq 1\\ \mathbb{R}^n\times(0,\infty)}} |v_t^{\varepsilon,\delta}| = \sup_{\substack{0<\delta\leq 1\\ \mathbb{R}^n}} |v_t^{\varepsilon,\delta}|\,|_{t=0} \leq \frac{C}{\varepsilon^2}\sup_{\mathbb{R}^n} |D^2 g|. \tag{2.18}$$

Owing to the bounds (2.17) and (2.18), there exists a sequence $\delta_j \to 0$ and a Lipschitz function v^ε such that

$$v^{\varepsilon,\delta_j} \longrightarrow v^\varepsilon \quad \text{locally uniformly in } \mathbb{R}^n \times [0,\infty). \tag{2.19}$$

Routine viscosity-solution arguments (cf. [8], [9]) prove v^ε to be a weak solution of (2.5_ε). Estimates (2.13) and (2.14) follow from (2.17) and (2.18). The uniqueness of the weak solution follows from Chen, Giga, and Goto [8], as does the contraction assertion (iii). □

We next derive a bound on v_t^ε which does not depend on $\varepsilon > 0$. This will later be useful in investigating the instantaneous unfolding of level sets.

LEMMA 2.4. *For each compact set $K \subset \mathbb{R}^n$, there exists a constant $C = C(K)$ such that*

$$\operatorname*{ess\,sup}_{\substack{0<\varepsilon\leq 1\\ x\in K}} |v_t^\varepsilon(x,t)| \leq C\left(1+\frac{1}{t}\right) \tag{2.20}$$

for a.e. $t > 0$.

Proof. 1. We will employ a scaling argument. For each $\lambda > 1$, set

$$w(x,t) = w^{\varepsilon,\lambda}(x,t) = v^\varepsilon(\lambda x, \lambda^2 t) \qquad (x \in \mathbb{R}^n, \quad t > 0). \tag{2.21}$$

2. We assert the following:

$$w_t = (w_{x_n}^2 + \varepsilon^2|D'w|^2)^{-1}(w_{x_n}^2\Delta' w - 2w_{x_i}w_{x_n}w_{x_ix_n} + |D'w|^2 w_{x_nx_n}) \tag{2.22}$$

in $R^n \times (0,\infty)$ in the weak sense. To verify this claim, let us suppose that $\phi \in C^\infty(\mathbb{R}^n \times (0,\infty))$ is given and $w - \phi$ has a local maximum at a point $(x_0,t_0) \in \mathbb{R}^n \times (0,\infty)$ with $D\phi(x_0,t_0) \neq 0$. Define

$$\psi(x,t) = \phi(\lambda^{-1}x, \lambda^{-2}t). \tag{2.23}$$

Then $v^\varepsilon - \psi$ has a local maximum at the point $(x_\lambda, t_\lambda) = (\lambda x_0, \lambda^2 t_0)$. Now

$$\begin{cases} \psi_t(x_\lambda,t_\lambda) = \lambda^{-2}\phi_t(x_0,t_0), \\ D\psi(x_\lambda,t_\lambda) = \lambda^{-1}D\phi(x_0,t_0), \\ D^2\psi(x_\lambda,t_\lambda) = \lambda^{-2}D^2\phi(x_0,t_0), \end{cases} \tag{2.24}$$

and consequently $D\psi(x_\lambda,t_\lambda) \neq 0$. Because v^ε is a weak solution of (2.5_ε), we have

$$\psi_t \leq (\psi_{x_n}^2 + \varepsilon^2|D'\psi|^2)^{-1}(\psi_{x_n}^2\Delta'\psi - 2\psi_{x_i}\psi_{x_n}\psi_{x_ix_n} + |D'\psi|^2\psi_{x_nx_n}) \leq 0$$

at (x_λ, t_λ). Employing (2.24), we therefore deduce

$$\phi_t \leq (\phi_{x_n}^2 + \varepsilon^2|D'\phi|^2)^{-1}(\phi_{x_n}^2\Delta'\phi - 2\phi_{x_i}\phi_{x_n}\phi_{x_ix_n} + |D'\phi|^2\phi_{x_nx_n}) \leq 0$$

at (x_0, t_0). We similarly deduce $\phi_t \le 0$ at (x_0, t_0) if $D\phi(x_0, t_0) = D^2\phi(x_0, t_0) = 0$. The opposite inequalities obtain should $w - \phi$ have a local minimum at (x_0, t_0). The claim (2.22) is proved.

3. In view of the contraction property for solutions of (2.5_ε) (Theorem 2.3(iii)), we have for each $x_0 \in \mathbb{R}^n$, $t_0 > 0$ that

$$|w^{\varepsilon,\lambda}(x_0, t_0) - v^\varepsilon(x, t_0)| \le \|g^\lambda - g\|_{L^\infty(\mathbb{R}^n)},$$

where $g^\lambda(x) = g(\lambda x)$, $x \in \mathbb{R}^n$. Consequently, taking $x_0 = 0$ above and recalling (2.21), we find

$$\frac{|v^\varepsilon(0, \lambda^2 t_0) - v^\varepsilon(0, t_0)|}{\lambda - 1} \le \frac{\|g^\lambda - g\|}{\lambda - 1} L^\infty(\mathbb{R}^n). \tag{2.25}$$

Now for a.e. $t_0 > 0$, $t \mapsto v^\varepsilon(0, t)$ is differentiable at $t = t_0$. For such a time t_0, we let $\lambda \to 1^+$:

$$2t_0 |v_t^\varepsilon(0, t_0)| \le \|Dg \cdot x\|_{L^\infty(\mathbb{R}^n)}.$$

Since $\sup_{\mathbb{R}^n} |Dg| < \infty$ and since g satisfies (2.12), the last term is finite. Consequently,

$$|v_t^\varepsilon(0, t_0)| \le \frac{C}{t_0},$$

assuming v^ε is differentiable at $(0, t_0)$. Shifting in space, we can replace 0 in the above calculation by any point $x_0 \in K$, K a given compact subset of $\mathbb{R}^n$. Thus

$$|v_t^\varepsilon(x_0, t_0)| \le \frac{C(K)}{t_0},$$

provided v^ε is differentiable in t at (x_0, t_0). But for a.e. x_0, $t \mapsto v^\varepsilon(x_0, t)$ is differentiable in t for a.e. t_0, according to Rademacher's theorem. □

3. Level sets solve the heat equation. It remains to interpret the level sets of any limit v of the approximate solutions v^ε as solving the heat equation. We modify a technique from [11].

THEOREM 3.1. *Assume $\varepsilon_j \to 0$ and*

$$v^{\varepsilon_j} \longrightarrow v \text{ locally uniformly in } \mathbb{R}^n \times (0, \infty). \tag{3.1}$$

Suppose that some level set of v can be represented in some region $N \subset \mathbb{R}^n \times (0, \infty)$ as a graph in the x_n-direction. That is, assume

$$N \cap \{(x, t) | v(x, t) = C\} = N \cap \{(x, t) | x_n = u(x', t)\} \tag{3.2}$$

for some constant C and some continuous function $u : \mathbb{R}^{n-1} \times (0, \infty) \to \mathbb{R}$. Then the height function u is smooth, and u solves the heat equation

$$u_t = \Delta' u \text{ within } N. \tag{3.3}$$

Proof. 1. We will show the heat equation (3.3) is satisfied in the weak (i.e., viscosity) sense. Let $\psi \in C^\infty(\mathbb{R}^{n-1} \times (0, \infty))$ and suppose

$$\begin{aligned} &u - \psi \text{ has a maximum (minimum) at a point} \\ &(x_0', t_0) \text{ such that } (x_0', u(x_0', t_0), t_0) = (x_0, t_0) \in N. \end{aligned} \tag{3.4}$$

We must show

$$\psi_t - \Delta'\psi \leq 0 \;\; (\geq 0) \quad \text{at} \quad (x_0', t_0). \tag{3.5}$$

Let us first assume

$$\begin{aligned} &u - \psi \;\text{ has a strict maximum at } (x_0', t_0) \\ &\text{with } \; u(x_0', t_0) = \psi(x_0', t_0). \end{aligned} \tag{3.6}$$

2. Without loss of generality, we may take $C = 0$. We may as well also suppose that

$$v^\varepsilon, v \leq 0 \quad \text{in} \quad \mathbb{R}^n \times (0, \infty);$$

otherwise, we note that $\tilde{v}^\varepsilon = -(v^\varepsilon)^2$ is also a weak solution of (2.5$_\varepsilon$) and $\tilde{v}^\varepsilon \to -v^2 \leq 0$. Also, we can assume $\lim_{|x'| \to \infty} \psi = +\infty$. Next, set

$$\phi(x,t) = \psi(x',t) - x_n. \tag{3.7}$$

Since $\psi(x',t) \geq u(x',t)$ for x' near x_0' and t near t_0,

$$\phi(x,t) \geq u(x',t) - x_n = 0$$

for all points on the level surface $\{v = 0\}$. Furthermore, $\phi(x_0, t_0) = v(x_0, t_0)$ when $(x_0, t_0) = (x_0', u(x_0', t_0), t_0)$. Define

$$\Phi(z) = \begin{cases} z & (z \geq 0), \\ \inf\{\phi(x,t) | v(x,t) \geq z\} & (z < 0); \end{cases}$$

then

$$\begin{cases} \Phi(0) = 0, & \Phi \text{ is lower semicontinuous,} \\ \Phi(z) \leq 0 & \text{if } \; z \leq 0, \\ \phi \geq \Phi(v). \end{cases}$$

Select a continuous function $\Psi : \mathbb{R} \to \mathbb{R}$ so that $\Psi \leq \Phi$, $\Psi(z) = \Phi(z)$ for $z \geq 0$. Consequently,

$$\phi \geq \Psi(v) = \overline{v}$$

with equality only at (x_0, t_0). Now

$$\overline{v}^\varepsilon = \Psi(v^\varepsilon)$$

is a weak solution of

$$\overline{v}^\varepsilon_t = ((\overline{v}^\varepsilon_{x_n})^2 + \varepsilon^2 |D\overline{v}^\varepsilon|^2)^{-1}((\overline{v}^\varepsilon_{x_n})^2 \Delta' \overline{v}^\varepsilon - 2\overline{v}^\varepsilon_{x_i} \overline{v}^\varepsilon_{x_n} \overline{v}^\varepsilon_{x_i x_n} + |D'\overline{v}^\varepsilon|^2 \overline{v}^\varepsilon_{x_n x_n}).$$

Since $\overline{v}^{\varepsilon_j} \to \overline{v}$ locally uniformly,

$$\overline{v}^{\varepsilon_j} - \phi \;\text{ has a maximum at a point } (x_{\varepsilon_j}, t_{\varepsilon_j})$$

with $(x_{\varepsilon_j}, t_{\varepsilon_j}) \to (x_0, t_0)$. As $\phi_{x_n} \neq 0$, we conclude

$$\phi_t \leq (\phi^2_{x_n} + \varepsilon_j^2 |D'\phi|^2)^{-1}(\phi^2_{x_n} \Delta'\phi - 2\phi_{x_i}\phi_{x_n}\phi_{x_i x_n} + |D'\phi|^2 \phi_{x_n x_n})$$

at $(x_{\varepsilon_j}, t_{\varepsilon_j})$. Recalling the definition (3.7) of ϕ, we obtain

$$\psi_t \leq (1 + \varepsilon_j^2 |D'\psi|^2)^{-1} \Delta'\psi$$

at $(x'_{\varepsilon_j}, t_{\varepsilon_j})$. Let $\varepsilon_j \to 0$ to discover

$$\psi_t \leq \Delta'\psi \quad \text{at} \quad (x_0', t_0).$$

The opposite inequality similarly holds if $u-\psi$ has a strict minimum at (x_0', t_0). Thus

$$u_t = \Delta' u \quad \text{in} \quad N$$

in the weak sense.

3. Finally, choose any cylinder $C' = B'(x_0', r) \times [t_0, t_0 + r^2] \subset N$ and consider the PDE

$$\begin{cases} \tilde{u}_t = \Delta' \tilde{u} & \text{in } C', \\ \tilde{u} = u & \text{on } B'(x_0', r) \times \{t = t_0\}, \\ \tilde{u} = u & \text{on } \partial B'(x_0', r) \times [t_0, t_0 + r^2]. \end{cases}$$

There exists a unique solution $\tilde{u}$, $\tilde{u}$ continuous on C' and C^∞ in the interior of C'. By uniqueness of viscosity solutions,

$$\tilde{u} = u,$$

and so u is C^∞ within N. □

4. Motion of level curves in the plane: Geometric interpretation. As noted in §2, the limit v of the solutions v^ε of (2.5_ε) is not in general a solution to the initial value problem (2.1), as v does not always continuously take on the initial value g. An informal interpretation of what happens is as follows. Assuming that the limit v continuously takes on the initial value $\tilde{g}$, we expect the approximations v^ε to develop an "initial layer," during which the level sets of v^ε move very rapidly, from those of g to approximately the level sets of $\tilde{g}$. Thereafter, the level sets of v^ε approximate the level sets of v and move slowly. Our intention is to substantiate this picture as rigorously as possible.

To simplify matters, for this and the next two sections, let us suppose $n = 2$, so that, as we shall see, a typical level set of v^ε is a curve evolving in the plane $\mathbb{R}^2$. We write $(x_1, x_2) = (x, y)$ to denote a typical point. We return now to our PDEs (2.1) and (2.5_ε) and explicitly display the geometric meaning. Let us temporarily suppose that v is a smooth solution of (2.1), with $v_y \neq 0$ in some region, to which we turn our attention. Then

$$\boldsymbol{\nu} = \frac{Dv}{|Dv|} = \frac{(v_x, v_y)}{|Dv|} = (\nu^1, \nu^2)$$

is a unit normal vector field to any given level curve. The normal velocity of this curve is $v_t/|Dv|$ and its curvature is

$$\begin{aligned} \kappa = \operatorname{div}(\boldsymbol{\nu}) &= \frac{1}{|Dv|} \frac{(\Delta v - v_{x_i} v_{x_i} v_{x_i x_j})}{|Dv|^2} \\ &= \frac{1}{|Dv|^3}(v_y^2 v_{xx} - 2 v_x v_y v_{xy} + v_x^2 v_{yy}). \end{aligned} \tag{4.1}$$

Since v solves (2.1) and thus

$$v_t = v_{xx} - \frac{2v_x}{v_y} v_{xy} + \frac{v_x^2}{v_y^2} v_{yy},$$

we can utilize (4.1) to compute that

$$\frac{v_t}{|Dv|} = \frac{|Dv|^2}{(v_y)^2} \kappa.$$

Thus the geometric law for the motion of the level curves of v is

$$\text{normal velocity} = \frac{\kappa}{(\nu^2)^2}. \tag{4.2}$$

In particular, (4.2) provides a geometric interpretation of the one-dimensional heat equation. Analogously, the PDE (2.5_ε) implies that the level curves of v^ε evolve according to the geometric law that

$$\text{normal velocity} = \frac{\kappa}{(\nu^2)^2 + \varepsilon^2(\nu^1)^2}. \tag{4.3_ε}$$

Thus the approximation $\varepsilon = 1$ corresponds to classical curvature motion and $\varepsilon = 0$ to the heat equation. It is clear that (4.3_ε) is in some sense a natural approximation to (4.2). The law of motion (4.2) ordains—at least formally—an infinite propagation velocity whenever the normal ν is horizontal and $\kappa \neq 0$. The approximate law of motion (4.3_ε), on the other hand, is a smooth nonisotropic modification of classical flow by curvature in the plane (cf. Oaks [19], Grayson [16], Gage and Hamilton [15], etc.).

We illustrate the effects of such a motion upon a given curve Γ_0 in Figure 1.

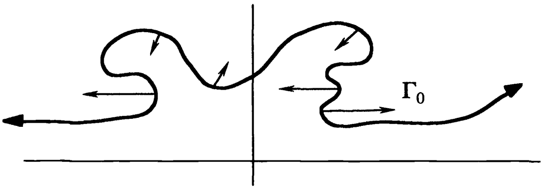


FIG. 1.

If we regard Γ_0 as an initial level set of v^ε, the law of motion (4.3_ε) forces a large horizontal velocity, as illustrated, along the folds and a small velocity in the regions where the curve is approximately flat and approximately horizontal.

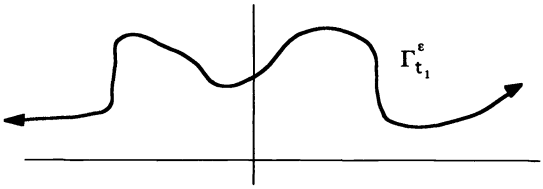


FIG. 2. $\Gamma^\varepsilon_{t_1}$ $(0 < t_1 << 1)$.

We consequently expect the level set to evolve quickly into a shape approximately like that in Figure 2: there is no longer a large horizontal velocity since the curvature is small where the curve is close to vertical. Thereafter, the motion should be approximated by the usual heat flow (see Figure 3).

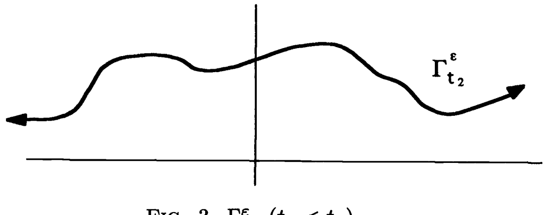


FIG. 3. $\Gamma^\varepsilon_{t_2}$ $(t_1 < t_2)$.

This example suggests that the level sets of v^ϵ, if not initially graphs, will rapidly unfold and become graphs.

The following formal calculation reinforces this belief. Suppose v is a smooth solution of (2.1) for $n = 2$,

$$v_t = v_{xx} - \frac{2v_x}{v_y}v_{xy} + \frac{v_x^2}{v_y^2}v_{yy} \quad \text{in } \mathbb{R}^2 \times (0, \infty) \tag{4.4}$$

with $|Dv| \neq 0$. Let Γ_t denote some level curve of v at a fixed time $t > 0$. Consider the ODE

$$\begin{cases} \dot{x}(s) = v_y(x(s), y(s), t), \\ \dot{y}(s) = -v_x(x(s), y(s), t) \qquad (\dot{} = \frac{d}{ds}). \end{cases} \tag{4.5}$$

Then

$$\frac{d}{ds}\, v(x(s), y(s), t) = 0, \tag{4.6}$$

and so if $(x(0), y(0)) \in \Gamma_t$, we have $(x(s), y(s)) \in \Gamma_t$ as well for all $s \in \mathbb{R}$. Differentiating (4.6) with respect to s, we compute

$$v_x\ddot{x} + v_y\ddot{y} + v_{xx}(\dot{x})^2 + 2v_{xy}\dot{x}\dot{y} + v_{yy}(\dot{y})^2 = 0.$$

Thus (4.5) implies

$$\begin{aligned} |\dot{x}\ddot{y} - \dot{y}\ddot{x}| &= |v_y^2 v_{xx} - 2v_x v_y v_{xy} + v_x^2 v_{yy}| \\ &= |\dot{x}|^2 |v_t| \qquad \text{by (4.4)}. \end{aligned}$$

According to (2.20), we have the estimate

$$|v_t| \leq \frac{C}{t},$$

and thus

$$|\dot{x}\ddot{y} - \dot{y}\ddot{x}| \leq \frac{C}{t}|\dot{x}|^2. \tag{4.7}$$

But

$$\frac{d}{ds}\left(\frac{\dot{x}}{((\dot{x})^2 + (\dot{y})^2)^{\frac{1}{2}}}\right) = \frac{\dot{y}(\dot{y}\ddot{x} - \dot{x}\ddot{y})}{((\dot{x})^2 + (\dot{y})^2)^{\frac{3}{2}}} = \dot{y}\kappa.$$

Consequently, (4.7) implies

$$\begin{aligned} \left|\frac{d}{ds}\left(\frac{\dot{x}}{((\dot{x})^2 + (\dot{y})^2)^{\frac{1}{2}}}\right)\right| &\leq \frac{C}{t}\,\frac{|\dot{y}||\dot{x}|^2}{((\dot{x})^2 + (\dot{y})^2)^{\frac{3}{2}}} \\ &\leq \frac{C}{t}\,\frac{|\dot{x}|}{((\dot{x})^2 + (\dot{y})^2)^{\frac{1}{2}}}. \end{aligned}$$

Applying Gronwall's inequality, we see that $\dot{x} = v_y$ either never vanishes along Γ_t or else is identically zero. The later possibility is excluded, as the level sets of v are horizontal lines for large $|x|$. Hence Γ_t is a graph, $y = u(x, t)$, for some function $u(\cdot\,, t) : \mathbb{R} \to \mathbb{R}$, which necessarily solves the heat equation.

The point of the foregoing calculation is that the curvature $\kappa = (\dot{y}\ddot{x} - \dot{x}\ddot{y})((\dot{x})^2 + (\dot{y})^2)^{-\frac{3}{2}}$ of the local curve Γ_t can be computed in terms of the right-hand side of the PDE (4.4), which in turn is bounded at each time $t > 0$. However, the proof is *not* geometrically intrinsic inasmuch as the bound on v_t implies an estimate only in $\frac{\kappa|Dv|}{|\nu^2|^2}$

and not $\frac{\kappa}{(\nu^2)^2}$. Note in particular that the parameter s is not arclength. It is consequently difficult to modify the foregoing computation to make a rigorous assertion: we lack in particular any good lower bounds on $\{|Dv^\varepsilon|\}_{0<\varepsilon\leq 1}$. Some different tools are called for, entailing a closer look at the level curves of the approximations.

5. Unfolding of the level curves of v^ε in the plane. We focus our attention on the approximations (2.5_ε), which for $n = 2$ read

$$(5.1_\varepsilon)\qquad \begin{cases} v_t^\varepsilon = \frac{(v_y^\varepsilon)^2 v_{xx}^\varepsilon - 2v_x^\varepsilon v_y^\varepsilon v_{xy}^\varepsilon + (v_x^\varepsilon)^2 v_{yy}^\varepsilon}{(v_y^\varepsilon)^2 + \varepsilon^2 (v_x^\varepsilon)^2} & \text{in } \mathbb{R}^2 \times (0,\infty), \\ v^\varepsilon = g & \text{on } \mathbb{R}^2 \times \{t = 0\}. \end{cases}$$

As observed in §4, this PDE says that the level curves of v^ε move with normal velocity $((\nu^2)^2 + \varepsilon^2(\nu^1)^2)^{-1}\kappa$.

Let us first note that each level set of v^ε is in fact a smooth, embedded curve for all times $t \geq 0$ provided it is so at $t = 0$. This fact is a variant of M. Grayson's important theorem that an embedded curve moving under curvature motion in the plane remains smooth and embedded until (and if) it collapses to a point. See Oaks [9] for proofs.

Let us hereafter look at a given level set $\Gamma_t^\varepsilon = \{(x,y) \in \mathbb{R}^n | v^\varepsilon(x,y,t) = C\}$ for some given constant C. We assume

$$(5.2)\qquad \begin{cases} \Gamma_0^\varepsilon \text{ is a smooth, embedded curve with} \\ p(\Gamma_0^\varepsilon) = \mathbb{R}, \end{cases}$$

$p(\,\cdot\,)$ denoting the projection on $\mathbb{R}^2$ onto the x-axis. In view of hypothesis (2.12), Γ_0^ε is a horizontal line within the region $\{|x| \geq R\}$ (see Figure 4).

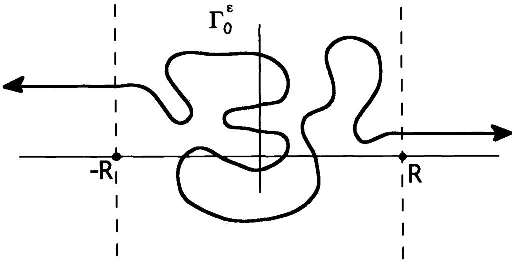


FIG. 4

First, we study the level sets $\{\Gamma_t^\varepsilon\}_{t\geq 0}$ in the region $\{|x| \geq R\}$.

LEMMA 5.1.

(i) *For each time $t \geq 0$, Γ_t^ε is a single-valued graph in the region $\{|x| \geq R\}$; that is,*

$$(5.3)\qquad \Gamma_t^\varepsilon \cap \{|x| \geq R\} = \{(x,y) \in \mathbb{R}^n \,||x| \geq R,\ u^\varepsilon(x,t) = y\}$$

for some smooth function $u^\varepsilon : \mathbb{R} \times [0,\infty) \to \mathbb{R}$.

(ii) *The height function u^ε solves the PDE*

$$(5.4)\qquad u_t^\varepsilon = \frac{u_{xx}^\varepsilon}{1 + \varepsilon^2 (u_x^\varepsilon)^2} \qquad \textit{in } \{|x| \geq R\} \times (0,\infty).$$

(iii) *There exists a constant C depending only on R such that*

$$\sup_{0<\varepsilon\leq 1} \sup_{\substack{|x|\geq R+1 \\ t\geq 0}} |u_x^\varepsilon| \leq C < \infty. \tag{5.5}$$

Proof. 1. At time $t = 0$, the curve Γ_0^ε and the vertical line $\{x = R\}$ meet transversely (in fact, perpendicularly) at a single point. According to results of Angenent [1], [2], the curves Γ_0^ε and $\{x = R\}$ intersect at most once for each time $t > 0$. (Indeed, near any point p_0 where the normal to Γ_t^ε is bounded away from $e_2 = (0, 1)$, we can locally represent Γ_t^ε as a graph in the y-direction.) Thus for some neighborhood N of p_0,

$$\Gamma_t^\varepsilon \cap N = \{(x, y, t) \in N | w^\varepsilon(y, t) = x\},$$

where $w^\varepsilon : \mathbb{R} \times (0, \infty) \to \mathbb{R}$ is smooth and solves the PDE

$$w_t^\varepsilon = \frac{w_{yy}^\varepsilon}{(w_y^\varepsilon)^2 + \varepsilon^2} \qquad \text{near } p_0$$

(cf. (6.2_ε) below). According to Angenent [1, Thm. C], the number of crossings of the line $\{x = \mathbb{R}\}$ is nonincreasing in time. (See also Angenent [2, §5] for a related assertion for isotropic curvature flow.)

As $\Gamma_t^\varepsilon \cap \{x > R\} \neq \emptyset$ and $\Gamma_t^\varepsilon \cap \{x < -R\} \neq \emptyset$, in fact, Γ_t^ε intersects the line $\{x = R\}$ precisely once. Let $y(t)$ denote the point of intersection; then $t \mapsto y(t)$ is smooth. Furthermore, $\{y(t)\}_{t\geq 0}$ is bounded since $\{\Gamma_t^\varepsilon\}_{t\geq 0}$ cannot intersect the lines $y = \pm R$.

2. Now consider the PDE

$$u_t^\varepsilon = \frac{u_{xx}^\varepsilon}{1 + \varepsilon^2 (u_x^\varepsilon)^2} \qquad \text{in } \{x > R\} \times (0, \infty) \tag{5.5_ε}$$

with the initial condition

$$u^\varepsilon(x, 0) \equiv y(0) \tag{5.6}$$

and boundary conditions

$$u^\varepsilon(0, t) = y(t), \qquad \lim_{x\to\infty} u^\varepsilon(x, t) = y(0). \tag{5.7}$$

This initial/boundary-value problem has a unique, smooth solution u^ε (cf. [18]) whose graph moves according to the geometric motion (4.3_ε). By uniqueness, $\Gamma_t^\varepsilon \cap \{x > R\}$ is the graph of u^ε.

3. We next employ routine Bernstein-type methods to derive uniform interior estimates for u_x^ε in $\{x > R + 1\} \times [0, \infty)$. For simplicity of notation, we suppress the superscript ε.

Let $\zeta \in C_c^\infty(\mathbb{R})$ be a cutoff function such that

$$0 \leq \zeta \leq 1, \qquad \operatorname{spt}(\zeta) \subset (R, \infty), \tag{5.8}$$

and set

$$z = \zeta^2 u_x^2 + \lambda u^2,$$

the constant $\lambda > 0$ to be selected later. Then

$$\text{(5.9)}\qquad \begin{cases} z_t &= 2\zeta^2 u_x u_{xt} + 2\lambda u u_t, \\ z_x &= 2\zeta^2 u_x u_{xx} + 2\zeta\zeta_x u_x^2 + 2\lambda u u_x, \\ z_{xx} &= 2\zeta^2 u_x u_{xxx} + 2\zeta^2 u_{xx}^2 + 8\zeta\zeta_x u_x u_{xx} \\ & \quad + 2(\zeta\zeta_{xx} + \zeta_x^2)u_x^2 + 2\lambda u u_{xx} + 2\lambda u_x^2. \end{cases}$$

Differentiate the PDE (5.5_ε) with respect to x:

$$\text{(5.10)}\qquad u_{xt} = \frac{u_{xxx}}{1+\varepsilon^2 u_x^2} - \frac{2\varepsilon^2 u_x u_{xx}^2}{(1+\varepsilon^2 u_x^2)^2}.$$

Fix $T > 0$. If z attains its maximum over $\{x \geq R\} \times [0, T]$ at some point $x_0 > R$, $0 < t_0 \leq T$, where $u_x(x_0, t_0) \neq 0$, then

$$0 \leq z_t - \frac{z_{xx}}{1+\varepsilon^2 u_x^2} \qquad \text{at } (x_0, t_0).$$

Utilizing (5.9), we compute

$$\begin{aligned} 0 \leq{}& 2\zeta^2 u_x \left(u_{xt} - \frac{u_{xxx}}{1+\varepsilon^2 u_x^2} \right) + 2\lambda u \left(u_t - \frac{u_{xx}}{1+\varepsilon^2 u_x^2} \right) \\ & - \frac{2\zeta^2 u_{xx}^2}{1+\varepsilon^2 u_x^2} - \frac{2\lambda u_x^2}{1+\varepsilon^2 u_x^2} - \frac{8\zeta\zeta_x u_x u_{xx}}{1+\varepsilon^2 u_x^2} - \frac{2(\zeta\zeta_{xx} + \zeta_x^2)u_x^2}{1+\varepsilon^2 u_x^2}. \end{aligned}$$

We simplify, making use of (5.5_ε) and (5.10):

$$\begin{aligned} 0 &\leq 2\zeta^2 u_x \left(\frac{-2\varepsilon^2 u_x u_{xx}^2}{(1+\varepsilon^2 u_x^2)^2} \right) - \frac{2\zeta^2 u_{xx}^2}{1+\varepsilon^2 u_x^2} - \frac{2\lambda u_x^2}{1+\varepsilon^2 u_x^2} \\ &\quad + \frac{C\zeta |u_x||u_{xx}|}{1+\varepsilon^2 u_x^2} + \frac{C u_x^2}{1+\varepsilon^2 u_x^2} \\ &\leq \frac{C u_x^2}{1+\varepsilon^2 u_x^2} - \frac{2\lambda u_x^2}{1+\varepsilon^2 u_x^2} \\ &< 0 \end{aligned}$$

if $\lambda > 0$ is fixed large enough. Thus z cannot in fact attain its maximum at a point $(x_0, t_0) \in \{x > R\} \times (0, T]$, with $u_x(x_0, t_0) \neq 0$.

Now restore the superscript ε. Since $z^\varepsilon = \zeta^2 (u_x^\varepsilon)^2 + \lambda (u^\varepsilon)^2$ cannot attain an interior maximum unless $u_x = 0$, and since u^ε is bounded, we deduce that

$$\sup_{\varepsilon > 0} \sup_{\substack{x \geq R \\ t \geq 0}} |z^\varepsilon| \leq C < \infty,$$

the constant C depending only on R and $\sup |\zeta_x|, |\zeta_{xx}|$. Given any point (x, t) with $x > R + 1$, we select ζ so that $\zeta(x) = 1$ and conclude

$$\sup_{\varepsilon > 0} |u_x^\varepsilon(x, t)| \leq C < \infty.$$

The constant C depends only on R.

Similar arguments apply for the region $\{x < -R\}$. □

Next we study the evolution of $\{\Gamma_t^\varepsilon\}_{t \geq 0}$ within the set $\{|x| \leq R\}$.

LEMMA 5.2. *There exists a constant* C *such that for each* $0 < \varepsilon \le 1$ *and* $T > 0$,

(5.11)
$$\sup_{0\le t\le T} H^1(\Gamma_t^\varepsilon \cap \{|x| \le R+1\}) + \int_0^T \int_{\Gamma_t^\varepsilon\cap\{|x|\le R+1\}} \frac{(\kappa^\varepsilon)^2}{(\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2}\, dH^1\, dt$$
$$\le C(H^1(\Gamma_0^\varepsilon \cap \{|x| \le R+1\}) + T).$$

Here H^1 denotes the one-dimensional Hausdorff measure and $\nu^\varepsilon = (\nu^{\varepsilon,1}, \nu^{\varepsilon,2})$ is the unit normal to Γ_t^ε (taken to be pointing upwards in $\{|x| \ge R\}$).

Proof. Select a cutoff function $\zeta \in C_c^\infty(\mathbb{R}^2)$ such that

$$\text{(5.12)}\qquad \begin{cases} 0 \le \zeta \le 1, & \zeta \equiv 1 \quad \text{if } |x| \le R+1,\ |y| \le R+1,\\ & \zeta \equiv 0 \quad \text{if } |x| \ge R+2 \ \text{ or } |y| \ge R+2,\\ |D\zeta| \le C. & \end{cases}$$

Then

$$\frac{d}{dt}\left(\int_{\Gamma_t^\varepsilon} \zeta^2 dH^1\right) = \int_{\Gamma_t^\varepsilon} \zeta^2 \kappa^\varepsilon \boldsymbol{\nu}^\varepsilon \cdot \mathbf{v}^\varepsilon dH^1 + \int_{\Gamma_t^\varepsilon} 2\zeta \mathbf{v}^\varepsilon \cdot D\zeta\, dH^1,$$

$\mathbf{v}^\varepsilon$ denoting the normal velocity vector. Now, according to (4.3_ε),

$$\mathbf{v}^\varepsilon = \frac{-\kappa^\varepsilon}{(\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2}\, \nu^\varepsilon.$$

Thus

$$\text{(5.13)}\qquad \begin{aligned} \frac{d}{dt}\left(\int_{\Gamma_t^\varepsilon} \zeta^2 dH^1\right) &\le -\int_{\Gamma_t^\varepsilon} \frac{\zeta^2(\kappa^\varepsilon)^2}{(\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2}\, dH^1 \\ &\quad + 2\int_{\Gamma_t^\varepsilon} \frac{\zeta|D\zeta||\kappa^\varepsilon|}{(\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2}\, dH^1 \\ &\le -\tfrac{1}{2}\int_{\Gamma_t^\varepsilon} \frac{\zeta^2(\kappa^\varepsilon)^2}{(\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2}\, dH^1 \\ &\quad + C\int_{\Gamma_t^\varepsilon} \frac{|D\zeta|^2}{(\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2}\, dH^1. \end{aligned}$$

Since $\Gamma_t^\varepsilon \cap \{|y| > R\} = \emptyset$ and $|D\zeta| \ne 0$ only if $R+1 \le |x| \le R+2$ or $R+1 \le |y| \le R+2$, we can compute

$$\text{(5.14)}\qquad \begin{aligned} \int_{\Gamma_t^\varepsilon} \frac{|D\zeta|^2}{(\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2}\, dH^1 &\le C\int_{\Gamma_t^\varepsilon\cap\{R+1\le|x|\le R+2\}} \frac{1}{(\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2}\, dH^1 \\ &= \int_{\{R+1\le|x|\le R+2\}} \frac{(1+(u_x^\varepsilon)^2)^{\frac{3}{2}}}{1+\varepsilon^2(u_x^\varepsilon)^2}\, dx \\ &\le C \qquad \text{by Lemma 5.1(iii).} \end{aligned}$$

Integrating (5.13), we obtain inequality (5.11). □

Finally, we estimate a time T_ε by which the level sets Γ_t^ε have become graphs. The method is to devise a kind of Harnack inequality for $\nu^{2,\varepsilon}$.

THEOREM 5.3 (unfolding of level curves). *There exists a time $T_\varepsilon > 0$ such that Γ_t^ε is a smooth graph in the y-direction for all times $t \geq T_\varepsilon$. Furthermore, we have the estimate*

$$0 < T_\varepsilon \leq \frac{C(H^1(\Gamma_0^\varepsilon \cap \{|x| \leq R+1\})+1)^2}{(\log\varepsilon)^2}. \tag{5.15}$$

Proof. In the following calculations, s denotes arclength and $\dot{} = \frac{d}{ds}$. Then for each fixed time $t \geq 0$,

$$\frac{d}{ds}\log((\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2) = \frac{2\nu^{2,\varepsilon}\dot{\nu}^{2,\varepsilon} + 2\varepsilon^2\nu^{1,\varepsilon}\dot{\nu}^{1,\varepsilon}}{(\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2},$$

and so

$$\left|\frac{d}{ds}\log((\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2)\right| \leq \frac{C|\kappa^\varepsilon|}{((\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2)^{\frac{1}{2}}}.$$

Let p_0 denote the point where Γ_t^ε intersects the line $x = R+1$ and take p_1 to be any point on $\Gamma_t^\varepsilon \cap \{|x| \leq R\}$. Choose arclength parametrization so that p_0 corresponds to s_0 and p_1 to $s_1 > s_0$. Then

$$\begin{aligned}\log((\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2)|_{p_1} &\geq \log((\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2)|_{p_0} \\ &\quad - C\int_{s_0}^{s_1}\frac{|\kappa^\varepsilon|}{((\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2)^{\frac{1}{2}}}\,ds \\ &\geq \log((\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2)|_{p_0} \\ &\quad - C\int_{\Gamma_t^\varepsilon\cap\{|x|\leq R+1\}}\frac{|\kappa^\varepsilon|}{((\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2)^{\frac{1}{2}}}\,dH^1.\end{aligned}$$

Now $\nu^{2,\varepsilon}|_{p_0}$ is bounded away from 0—say $(\nu^{2,\varepsilon})^2 \geq \theta > 0$—uniformly in ε and t according to Lemma 3.1(iii). Consequently,

(5.16)

$$\begin{aligned}\inf_{\Gamma_t^\varepsilon}\log((\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2) \geq \log(\theta) - C&\left(\int_{\Gamma_t^\varepsilon\cap\{|x|\leq R+1\}}\frac{|\kappa^\varepsilon|^2}{(\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2}\,dH^1\right)^{\frac{1}{2}} \\ &\times H^1(\Gamma_t^\varepsilon \cap \{|x| \leq R+1\})^{\frac{1}{2}}.\end{aligned}$$

Now return to estimate (5.11) with $T = T_\varepsilon \leq 1$ as selected below. There exists a time $0 < t_\varepsilon \leq T^\varepsilon$ such that

$$\int_{\Gamma_{t_\varepsilon}^\varepsilon\cap\{|x|\leq r+1\}}\frac{(\kappa^\varepsilon)^2}{(\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2}\,dH^1 \leq \frac{C}{T_\varepsilon}(H^1(\Gamma_0^\varepsilon \cap \{|x| \leq R+1\})+1).$$

Set $t = t_\varepsilon$ in (5.16):

$$\inf_{\Gamma_{t_\varepsilon}^\varepsilon}\log((\nu^{2,\varepsilon})^2 + \varepsilon^2(\nu^{1,\varepsilon})^2) \geq \log\theta - \frac{C}{T_\varepsilon^{\frac{1}{2}}}(H^1(\Gamma_0^\varepsilon \cap \{|x| \leq R+1\})+1).$$

Consequently,

$$\inf_{\Gamma_{t_\varepsilon}^\varepsilon}(\nu^{2,\varepsilon})^2 \geq \theta e^{-\frac{C}{T_\varepsilon^{\frac{1}{2}}}(H^1(\Gamma_0^\varepsilon\cap\{|x|\leq R+1\})+1)} - \varepsilon > 0 \tag{5.17}$$

provided

$$-\frac{C}{T_\varepsilon^{\frac{1}{2}}}(H^1(\Gamma_0^\varepsilon \cap \{|x| \leq R+1\})+1) > \log\left(\frac{\varepsilon}{\theta}\right). \tag{5.18}$$

But if

$$T_\varepsilon \geq C\frac{H^1((\Gamma_0^\varepsilon \cap \{|x| \leq R+1\})+1)^2}{(\log \varepsilon)^2}$$

for some constant C, estimates (5.18) and thus (5.17) obtain. Consequently, there exists a time $0 \leq t_\varepsilon \leq T_\varepsilon$ for which estimate (5.17) holds. The curve $\Gamma_{t_\varepsilon}^\varepsilon$ is therefore a graph in the e_2-direction. The evolution maintains the property of being a graph, and thus Γ_t^ε is a graph for all times $t \geq T_\varepsilon$. □

6. Equal-area construction. The calculations in §5 demonstrate that the level curves of v^ε become graphs in the y-direction after a time of order $(\log \varepsilon)^{-2}$; thus, given a curve Γ_0 in the plane which is a horizontal line in the region $\{|x| > R\}$ for some $R > 0$, we can envision defining the operator

$$\mathcal{G}(\Gamma_0) = \lim_{t \to 0^+} \lim_{\varepsilon \to 0} \Gamma_t^\varepsilon. \tag{6.1}$$

The effect of $\mathcal{G}$ would be to replace the curve Γ_0 with a graph $\mathcal{G}(\Gamma_0)$, presumably in some natural, canonical way. This is mostly speculation since it is not known in general that the limits in (6.1) exist.

It is possible in one simple case to give a rigorous analysis. Let us assume Γ_0 is the backward-S-shaped curve in Figure 5.

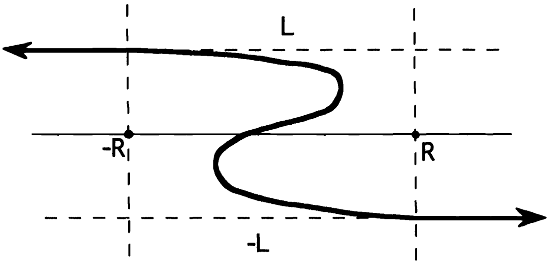


FIG. 5.

We will prove that in this case $\mathcal{G}(\Gamma_0)$ exists and is the graph (with a jump) obtained by the Maxwell equal-area construction (see Figure 6).

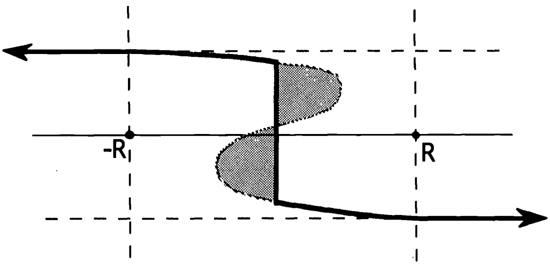


FIG. 6.

Consequently, the heat equation "instantly converts Γ_0 into $\mathcal{G}(\Gamma_0)$."

To make this assertion rigorous, let us suppose that Γ_0 is flat outside $\{|x| \leq R\}$ for some $R > 0$, say $\Gamma_0 = \{y = -L\}$ on $\{x > R\}$ and $\Gamma_0 = \{y = L\}$ on $\{x < -R\}$, where $L > 0$. We propose to interpret Γ_0 and its subsequent evolution $\{\Gamma_t^\varepsilon\}_{t\geq 0}$ under the geometric motion (4.3_ε) as the graph in the x-direction of a function $w^\varepsilon(y,t)$. We easily deduce that the PDE w^ε verifies by noting the graph of w^ε is a level curve of v^ε, v^ε solving (5.1_ε). For the following computation, suppose v^ε is smooth, with $v_x^\varepsilon \neq 0$. Then

$$v^\varepsilon(w^\varepsilon(y,t),y,t) = C \qquad (y \in \mathbb{R},\ \ t \geq 0)$$

for some constant C. Consequently,

$$\begin{gathered} v_x^\varepsilon w_t^\varepsilon + v_t^\varepsilon = 0, \qquad v_x^\varepsilon w_y^\varepsilon + v_y^\varepsilon = 0, \\ v_{xx}^\varepsilon (w_y^\varepsilon)^2 + 2v_{xy}^\varepsilon w_y^\varepsilon + v_x^\varepsilon w_{yy}^\varepsilon + v_{yy}^\varepsilon = 0. \end{gathered}$$

Substituting into (5.1_ε), we deduce after some easy computations that

$$w_t^\varepsilon = \frac{w_{yy}^\varepsilon}{(w_y^\varepsilon)^2 + \varepsilon^2} \qquad \text{in } (-L,L)\times(0,\infty). \tag{6.2}$$

Remark. Note in particular the case $\varepsilon = 0$. If a function $y = u(x,t)$ solves the heat equation and its graph can be written as function $x = w(y,t)$, then w solves

$$w_t = \frac{w_{yy}}{(w_y)^2}. \tag{6.3}$$

This PDE and the corresponding divergence-structure PDE

$$z_t = \left(\frac{z_y}{z^2}\right)_y \tag{6.4}$$

satisfied by $z = w_y$ have been identified in several papers as being exactly solvable; see Bluman and Kumai [3], [4], Fokas and Yortos [14], and Olver [21]. The point is that (6.3) is tranformed into the linear heat equation by a simple rotation through $\pi/2$. □

We propose to study the limits of the function w^ε as $\varepsilon \to 0$ provided the initial data is

$$w^\varepsilon = h \qquad \text{on } (-L,L)\times\{t=0\}, \tag{6.5}$$

where $h : (-L,L) \to \mathbb{R}$ has the graph obtained by rotating Γ_0 clockwise through $\pi/2$.

We transform (6.2) and (6.5) to read

$$\begin{cases} w_t^\varepsilon = (\Phi^\varepsilon(w_y^\varepsilon))_y & \text{in } (-L,L)\times(0,\infty), \\ w^\varepsilon = -\infty & \text{on } \{-L\}\times(0,\infty), \\ w^\varepsilon = +\infty & \text{on } \{L\}\times(0,\infty), \\ w^\varepsilon = h & \text{on } \mathbb{R}\times\{t=0\} \end{cases} \tag{6.6_ε}$$

for

$$\Phi^\varepsilon(r) = \frac{1}{\varepsilon}\left(\arctan\left(\frac{r}{\varepsilon}\right) - \frac{\pi}{2}\right) \qquad (r \in \mathbb{R}). \tag{6.7}$$

Consider next the initial-value boundary problem

$$\begin{cases} W_t^\varepsilon = \Phi^\varepsilon(W_{yy}^\varepsilon) & \text{in } (-L,L)\times(0,\infty), \\ W^\varepsilon = A & \text{on } \{-L\}\times[0,\infty), \\ W^\varepsilon = B & \text{on } \{L\}\times[0,\infty), \\ W^\varepsilon = H & \text{on } \mathbb{R}\times\{t=0\}, \end{cases} \tag{6.8_ε}$$

where

$$H' = h \quad \text{in } |y| < L, \quad \min H = 0, \qquad H(-L) = A, \quad H(L) = B. \tag{6.9}$$

Thus H has two wells and $H(y) = +\infty$ for $|y| > L$ (see Figure 7).

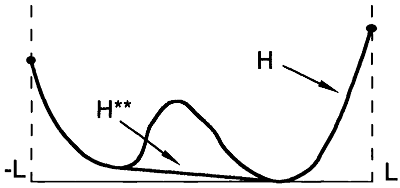


FIG. 7.

LEMMA 6.1. *There exists for each $\varepsilon > 0$ a unique smooth solution W^ε of* (6.8_ε).

Proof. 1. The curve Γ_0 intersects each horizontal line $\{y = C\}$ precisely once for each $-L < C < L$. Furthermore, this intersection is transverse. Consequently, Angenent's theorem [1] implies Γ_t^ε intersects each line $\{y = C\}$ precisely once for each time $t > 0$, and so Γ_t^ε is a graph in the x-direction—say $x = w^\varepsilon(y,t)$, where $|y| < L$, $t \geq 0$. The function w^ε is a smooth solution of $w_t^\varepsilon = (\Phi^\varepsilon(w_y^\varepsilon))_y$. Define

$$W^\varepsilon(y,t) = A + \int_{-L}^{y} w^\varepsilon(z,t)dz \qquad \text{for } |y| < L.$$

Then W^ε is a smooth solution of the PDE $W_t^\varepsilon = \Phi^\varepsilon(W_{yy}^\varepsilon)$, with $W^\varepsilon = H$ at $t = 0$. Now, near the lines $y = \pm L$, the graph of w^ε is approximately the graph of a solution to the heat equation rotated through $\pi/2$. Consequently,

$$\lim_{y \to \pm L} w_y^\varepsilon = \lim_{y \to \pm L} W_{yy}^\varepsilon = +\infty,$$

and so

$$\lim_{y \to \pm L} w_t^\varepsilon = \lim_{y \to \pm L} \Phi^\varepsilon(W_{yy}^\varepsilon) = 0 \tag{6.10}$$

uniformly in $[0,T]$ for each $T > 0$. Thus $w^\varepsilon(-L,t) = A$, $w^\varepsilon(L,t) = B$ for $t \geq 0$.

2. If W^ε and $\tilde{W}^\varepsilon$ are two solutions, the graphs of $w^\varepsilon = W_y^\varepsilon$ and $\tilde{w}^\varepsilon = \tilde{W}_y^\varepsilon$ both evolve according to the geometric motion (4.3_ε) starting from Γ_0. By uniqueness, $w^\varepsilon \equiv \tilde{w}^\varepsilon$. Since $W^\varepsilon(L,t) = \tilde{W}^\varepsilon(L,t)$ for all $t \geq 0$, $W^\varepsilon \equiv \tilde{W}^\varepsilon$. □

Next we study the asymptotic behavior of W^ε as $\varepsilon \to 0$. Let us write H^{**} to denote the convex regularization of H; that is, H^{**} is the largest convex function less than or equal to H.

THEOREM 6.2 (equal-area construction). *We have*

$$\lim_{t \to 0^+} \lim_{\varepsilon \to 0} W^\varepsilon(\cdot,t) = H^{**} \tag{6.11}$$

locally uniformly in $(-L,L)$, *and*

$$\lim_{t \to 0^+} \lim_{\varepsilon \to 0} w^\varepsilon(\cdot,t) = (H^{**})' \tag{6.12}$$

in $L^p_{\text{loc}}(-L,L)$ $(1 \leq p < \infty)$. *In particular, the limits in* (6.11) *and* (6.12) *exist.*

Proof. 1. Fix $t > 0$. Inasmuch as $\Phi^\varepsilon \leq 0$, we deduce from (6.8_ε) that

$$W^\varepsilon(y,t) \leq H(y) \qquad (|y| < L, \quad t \geq 0). \tag{6.13}$$

Recall also that the graph Γ_t^ε of $w^\varepsilon = W_y^\varepsilon$ moves following (4.3_ε). According to Theorem 5.3, Γ_t^ε is a graph in the y-direction for times $t \geq T_\varepsilon = O((\log \varepsilon)^{-2})$. The mapping $y \mapsto w^\varepsilon(y,t)$ is therefore strictly increasing for times $t \geq T_\varepsilon$. Consequently, $y \mapsto W^\varepsilon(y,t)$ is uniformly convex.

Fix a small time $\tau > 0$. Then, replacing τ_ε by τ in (5.17), we deduce for some time $0 \leq t_\varepsilon \leq \tau$ that

$$\begin{aligned}\inf_{\Gamma_{t_\varepsilon}^\varepsilon}(\nu^{2,\varepsilon})^2 &\geq \theta\, e^{-C/\tau^{\frac{1}{2}}} - \varepsilon \\ &\geq \gamma\end{aligned}$$

for $\gamma = \gamma(\tau) > 0$ and all ε sufficiently small. In particular, therefore,

$$\sup_{\substack{t \geq \tau \\ x \in \mathbb{R}}} |u_x^\varepsilon(x,t)| < M$$

for ε sufficiently small, $M = M(\tau)$. Consequently,

$$W_{yy}^\varepsilon(y,t) \geq \mu > 0 \tag{6.14}$$

for some constant $\mu > 0$ and all $t \geq \tau$, $|y| < L$. Since

$$\lim_{\varepsilon \to 0} \Phi^\varepsilon(r) = -\frac{1}{r} \qquad \text{if } r > 0 \tag{6.15}$$

uniformly on compact subsets of $[\mu, \infty)$, we deduce from the PDE (6.8_ε) that

$$\sup_{\varepsilon > 0} \sup_{\substack{t \geq \tau > 0 \\ |y| < L}} |W_t^\varepsilon(y,t)| < \infty.$$

But also

$$\sup_{\varepsilon > 0} \sup_{\substack{t \geq 0 \\ |y| \leq L - \sigma}} |W_y^\varepsilon(y,t)| < \infty$$

for each $\sigma > 0$. Hence we can extract a subsequence $\varepsilon_j \to 0$ and find a function $W : \mathbb{R} \times (0,\infty) \to \mathbb{R}$ such that

$$W^{\varepsilon_j} \longrightarrow W \qquad \text{locally uniformly in } (-L,L) \times (0,\infty). \tag{6.16}$$

In view of (6.13) and (6.14), for each $t > 0$,

$$y \mapsto W(y,t) \quad \text{is convex and} \quad W(y,t) \leq H^{**}(y) \quad (|y| < L). \tag{6.17}$$

2. Next, assume

$$\hat{H} \text{ is uniformly convex}, \quad \hat{H} \leq H. \tag{6.18}$$

Let $\hat{W}^\varepsilon$ denote the solution of (6.8_ε) with $\hat{H}$ replacing H. Then

$$\hat{W}^\varepsilon \leq W^\varepsilon \quad \text{on } [-L,L] \times (0,\infty) \tag{6.19}$$

by the maximum principle. Passing to a further subsequence if necessary, we may assume

$$\hat{W}^{\varepsilon_j} \longrightarrow \hat{W} \quad \text{locally uniformly in } (-L,L) \times (0,\infty). \tag{6.20}$$

Since $\hat{H}'' \geq \theta > 0$ for some $\theta > 0$, we see from (6.15) that $\hat{W}$ is a smooth solution of

$$\hat{W}_t = -(\hat{W}_{yy})^{-1} \quad \text{in } (-L,L) \times (0,\infty).$$

Therefore,

$$\sup_{\substack{t\geq 0\\ |y|\leq L}} |\hat{W}_t| < \infty,$$

and so

$$\sup_{|y|\leq L} |\hat{W}(y,t) - \hat{H}(y)| = O(t) \quad \text{as } t \to 0. \tag{6.21}$$

From (6.17)–(6.21), we deduce

$$H^{**}(y) \geq W(y,t) \geq \hat{H}(y) + O(t) \quad (|y| \leq L). \tag{6.22}$$

Since the mappings $y \mapsto W(y,t)$ are convex and bounded for $0 < t \leq 1$, the derivatives $\{W_y(\,\cdot, t))\}_{0<t\leq 1}$ are uniformly locally bounded. Hence there exists a sequence $t_k \to 0$ and a convex function G such that

$$W(\cdot, t_k) \longrightarrow G$$

locally uniformly. Owing to (6.21),

$$H^{**}(y) \geq G(y) \geq \hat{H}(y) \qquad (|y| \leq L)$$

for each uniformly convex $\hat{H} \leq H$, as above. We conclude that $G = H^{**}$. This deduction holds for any such sequence $t_k \to 0$, and so

$$\lim_{t\to 0^+} W(\cdot, t) = H^{**} \tag{6.23}$$

locally uniformly.

As $y \mapsto W(y,t)$ is convex for each $t > 0$, the limit

$$\lim_{t\to 0^+} W_y(\cdot, t) = (H^{**})' \tag{6.24}$$

exists in L^p_{loc}, $1 \leq p < \infty$. Furthermore, for each fixed $t > 0$,

$$w^{\varepsilon_j}(\cdot, t) = W^{\varepsilon_j}_y(\cdot, t) \longrightarrow W_y(\cdot, t) \quad \text{in } L^p_{\text{loc}}.$$

Consequently, we may rewrite (6.23) and (6.24) to read

$$\begin{cases} \lim_{t\to 0^+} \lim_{\varepsilon_j \to 0} W^{\varepsilon_j}(\cdot, t) = H^{**}, \\ \lim_{t\to 0^+} \lim_{\varepsilon_j \to 0} w^{\varepsilon_j}(\cdot, t) = (H^{**})'. \end{cases} \tag{6.25}$$

3. It remains to replace the subsequence $\varepsilon_j \to 0$ in (6.25) with the full limit $\varepsilon \to 0^+$. Suppose therefore that we have another sequence $\varepsilon_i \to 0$, and

$$W^{\varepsilon_i} \longrightarrow \tilde{W}$$

locally uniformly in $(-L, L) \times (0, \infty)$. Then, as above,

$$H^{**} \geq \tilde{W} \geq \hat{H} + O(t)$$

for any uniformly convex $\hat{H} \leq H$. Thus given any $\delta > 0$, there exists a time $t_\delta > 0$ such that $t_\delta \to 0$ as $\delta \to 0$ and

$$\|\tilde{W}(\cdot, t_\delta) - W(\cdot, t_\delta)\|_{L^\infty(-L,L)} \leq \delta.$$

Now W and $\tilde{W}$ are uniformly convex and smooth in $\mathbb{R} \times [t_\delta, \infty)$, and

$$W_t = -(W_{yy})^{-1}, \quad \tilde{W}_t = -(\tilde{W}_{yy})^{-1} \quad \text{in } (-L, L) \times [t_\delta, \infty).$$

Furthermore,

$$W = \tilde{W} \quad \text{on} \quad \{-L\} \times (0,\infty) \quad \text{and} \quad \{L\} \times (0,\infty).$$

Thus the maximum principle implies

$$\sup_{\substack{|y|\leq L \\ t\geq t_\delta}} |\tilde{W}(y,t) - W(y,t)| < \delta.$$

This estimate obtains for each $\delta > 0$, and consequently $W = \tilde{W}$ on $[-L, L]\times [0,\infty)$. □

7. Analytically natural approximations: Motion of level sets in $\mathbb{R}^n$. It seems difficult to extend to $\mathbb{R}^n$ most of the analysis in §§4–6 concerning the unfolding of the level curves in the plane. Indeed, even the laws of evolution (4.2) and (4.3_ε) become considerably more complicated in higher dimensions.

To deduce the geometric law of motion in general, let us for the purposes of the following calculation suppose v is a smooth solution of the level-surface heat equation

$$v_t = \Delta' v - \frac{2v_{x_i}}{v_{x_n}} v_{x_i x_n} + \frac{|D'v|^2}{v_{x_n}^2} v_{x_n x_n} \quad \text{in } \mathbb{R}^n \times (0,\infty), \tag{7.1}$$

where, recall, the implicit summation is for $i = 1$ to $n-1$. We also assume $v_{x_n} \neq 0$ in some region, in which we analyze as follows the motion of the level surfaces of v. Let $\{\Gamma_t\}_{t\geq 0}$ denote some such level surface. Recall that

$$\boldsymbol{\nu} = \frac{Dv}{|Dv|} = (\nu^1, \ldots, \nu^n)$$

is a unit normal vector field to Γ_t and $v_t/|Dv|$ is the normal velocity. We also write $\boldsymbol{\nu} = (\nu', \nu^n)$ and denote by Γ_t' the surfaces obtained by intersecting Γ_t with the planes $\{x_n = \text{constant}\}$. Then $(n-1)$ times the mean curvature of Γ_t is

$$H = \operatorname{div}(\boldsymbol{\nu}) = \operatorname{div}\left(\frac{Dv}{|Dv|}\right)$$

and $(n-2)$ times the mean curvature of Γ_t' is

$$H' = \operatorname{div}\left(\frac{D'v}{|D'v|}\right),$$

the latter divergence computed in the variables $x' = (x_1, \ldots x_{n-1})$ with x_n held fixed.

Now

$$|Dv|H = \Delta v - \frac{v_{x_k} v_{x_l}}{|Dv|^2} v_{x_k x_l},$$

the summation for $1 \leq k, l \leq n$. Consequently,

$$\begin{aligned} |Dv|^3 H &= |Dv|^2 \Delta v - v_{x_k} v_{x_l} v_{x_k x_l} \\ &= |D'v|^2 \Delta' v + |D'v|^2 v_{x_n x_n} + v_{x_n}^2 \Delta' v - 2 v_{x_i} v_{x_n} v_{x_i x_n} - v_{x_i} v_{x_j} v_{x_i x_j}, \end{aligned}$$

summing $1 \leq i, j \leq n-1$. Thus (7.1) implies

$$\begin{aligned} |Dv|^3 H &= v_{x_n}^2 v_t + |D'v|^2 \Delta' v - v_{x_i} v_{x_j} v_{x_i x_j} \\ &= v_{x_n}^2 v_t + |D'v|^3 H'. \end{aligned}$$

Hence

$$\frac{v_t}{|Dv|} = \frac{|Dv|^2}{v_{x_n}^2} H - \frac{|D'v|^3}{|Dv|v_{x_n}^2} H'.$$

Consequently the level sets of v evolve so that

$$\text{normal velocity} = \frac{1}{(\nu^n)^2} H - \frac{|\nu'|^3}{(\nu^n)^2} H'. \tag{7.2}$$

For the approximations v^ε, which solve (3.2_ε), the law of motion of the level sets is

$$\text{normal velocity} = \frac{1}{(\nu^n)^2 + \varepsilon^2|\nu'|^2} H - \frac{|\nu'|^3}{(\nu^n)^2 + \varepsilon^2|\nu'|^2} H'. \tag{7.3_ε}$$

The evolution given by the PDE (2.5_ε), or equivalently (7.3_ε), is "geometrically natural" but nonetheless hard to visualize and difficult to study rigorously. We therefore propose to study instead the solutions w^ε of the "analytically natural" approximations

(7.4_ε)
$$\begin{cases} w_t^\varepsilon = ((w_{x_n}^\varepsilon)^2 + \varepsilon^2)^{-1}(((w_{x_n}^\varepsilon)^2 + \varepsilon^2)\Delta' w^\varepsilon - 2w_{x_i}^\varepsilon w_{x_n}^\varepsilon w_{x_i x_n}^\varepsilon + |D'w^\varepsilon|^2 w_{x_n x_n}^\varepsilon) \\ \qquad \text{in } \mathbb{R}^n \times (0,\infty), \\ w^\varepsilon = g \qquad \text{on } \mathbb{R}^n \times \{t=0\}, \end{cases}$$

the summation for $i = 1, \dots, n-1$. This PDE does not verify the geometric scaling identity (3.3), (3.5) but on the other hand is defined pointwise even if $|Dw^\varepsilon| = 0$. In the rest of the paper, we provide some preliminary analysis of the behavior of w^ε as $\varepsilon \to 0$. Many questions remain open, however.

LEMMA 7.1. *There exists a unique weak solution to* (7.4_ε). *In addition, we have the estimates*

$$\sup_{0<\varepsilon\le 1} \operatorname*{ess\,sup}_{\mathbb{R}\times[0,\infty)} |w^\varepsilon|, |Dw^\varepsilon| < \infty, \tag{7.5}$$

$$\operatorname*{ess\,sup}_{\substack{0<\varepsilon\le 1\\ x\in K}} |w_t^\varepsilon(x,t)| \le C\left(1 + \frac{1}{t}\right). \tag{7.6}$$

Proof. Since (7.4_ε) is not uniformly parabolic, we approximate further by considering the PDE

$$\begin{cases} w_t^{\varepsilon,\delta} = ((w_{x_n}^{\varepsilon,\delta})^2 + \varepsilon^2)^{-1}((w_{x_n}^{\varepsilon,\delta})^2 + \varepsilon^2)\Delta' w^{\varepsilon,\delta} \\ \qquad -2w_{x_i}^{\varepsilon,\delta} w_{x_n}^{\varepsilon,\delta} w_{x_i x_n}^{\varepsilon,\delta} + (|D'w|^2 + \delta) w_{x_n x_n}^{\varepsilon,\delta}) & \text{in } \mathbb{R}^n \times (0,\infty), \\ w^{\varepsilon,\delta} = g & \text{on } \mathbb{R}^n \times \{t=0\}. \end{cases} \tag{$7.7_{\varepsilon,\delta}$}$$

This PDE is uniformly parabolic for any solution with bounded gradient and so possesses a unique, bounded, smooth solution $w^{\varepsilon,\delta}$. We have the estimates

$$\sup_{\delta>0} \sup_{\mathbb{R}^n\times[0,\infty)} |w^{\varepsilon,\delta}|, |Dw^{\varepsilon,\delta}|, |w_t^{\varepsilon,\delta}| < \infty,$$

and thus there exists a subsequence $\delta_j \to 0$ and a Lipschitz function w^ε such that

$$w^{\varepsilon,\delta_j} \longrightarrow w^\varepsilon \quad \text{locally uniformly on } \mathbb{R}^n \times [0,\infty).$$

Then w^ε is the unique weak (i.e., viscosity) solution of (7.4_ε).

Estimate (7.5) is clear from the contraction property of the mapping $t \mapsto w^\varepsilon(\cdot, t)$. To prove estimate (7.6), we modify the method of Lemma 2.4 by noting that

$$z(x,t) = z^{\varepsilon,\lambda}(x,t) = w^\varepsilon(\lambda x', x_n, \lambda^2 t)$$

is also a weak solution of

$$z_t = (z_{x_n}^2 + \varepsilon^2)^{-1}((z_{x_n}^2 + \varepsilon^2)\Delta' z - 2z_{x_i} z_{x_n} z_{x_i x_n} + |D'z|^2 z_{x_n x_n}) \quad \text{in } \mathbb{R}^n \times (0,\infty). \ \square$$

In view of the bounds (7.5), (7.6), there exists a sequence $\varepsilon_j \to 0$ and a locally Lipschitz function v on $\mathbb{R}^n \times (0,\infty)$ such that

$$w^{\varepsilon_j} \longrightarrow v \quad \text{locally uniformly on } \mathbb{R} \times (0,\infty).$$

It is straightforward to verify that v is a weak solution of the level-surface heat equation, although in general v will not continuously attain the initial values presented by g. We expect instead that the level sets of v will be graphs in the x_n-direction, the point being that the PDE (7.4$_\varepsilon$) contains some kind of mechanism forcing the derivatives $w_{x_n}^\varepsilon$ to become nonnegative as $\varepsilon \to 0$, even though this is not necessarily true at time $t = 0$.

THEOREM 7.2. *We have*

$$v_{x_n} \geq 0 \quad a.e.\ in\ \mathbb{R}^n \times (0,\infty).$$

Proof. 1. We consider the PDE (7.7$_{\varepsilon,\delta}$) and for notational simplicity suppress the superscripts ε, δ. Thus $w = w^{\varepsilon,\delta}$ satisfies

(7.8)
$$w_t = (w_{x_n}^2 + \varepsilon^2)^{-1}((w_{x_n}^2 + \varepsilon^2)\Delta' w - 2w_{x_i} w_{x_n} w_{x_i x_n} + (|D'w|^2 + \delta) w_{x_n x_n}) \quad \text{in } \mathbb{R}^n \times (0,\infty).$$

Differentiate (7.8) with respect to x_n and (selectively) set $z = z^{\varepsilon,\delta} = w_{x_n}$:

$$\begin{aligned} &z_t - (w_{x_n}^2 + \varepsilon)^{-1}((w_{x_n}^2 + \varepsilon)\Delta' z - 2w_{x_i} w_{x_n} z_{x_i x_n} + (|D'w|^2 + \delta) z_{x_n x_n}) \\ &= -\frac{2z|D'z|^2}{z^2 + \varepsilon^2} + \left(-2\, \frac{w_{x_i} w_{x_i x_n}}{w_{x_n}^2 + \varepsilon^2} + 4\, \frac{w_{x_i} w_{x_n}^2 w_{x_i x_n}}{(w_{x_n}^2 + \varepsilon^2)^2} \right. \\ &\quad \left. + 2\, \frac{w_{x_i} w_{x_i x_n}}{w_{x_n}^2 + \varepsilon^2} - 2\, \frac{(|D'w|^2 + \delta) w_{x_n x_n} w_{x_n}}{(w_{x_n}^2 + \varepsilon^2)^2} \right) z_{x_n}. \end{aligned} \tag{7.9}$$

This PDE has the form

$$z_t - a_{kl} z_{x_k x_l} = -\frac{2z|D'z|^2}{z^2 + \varepsilon^2} + b z_{x_n} \quad \text{in } \mathbb{R}^n \times (0,\infty), \tag{7.10}$$

where the coefficients $((a_{kl}))_{1 \leq k,l \leq n}$ are uniformly positive definite.

2. We propose to compare $z = z(x,t)$ with $\tilde{z} = \tilde{z}^\varepsilon = \tilde{z}(x',t)$, a solution of

$$\begin{cases} \tilde{z}_t - \Delta' \tilde{z} = -\frac{2\tilde{z}|D'\tilde{z}|^2}{\tilde{z}^2 + \varepsilon^2} & \text{in } \mathbb{R}^{n-1} \times (0,\infty), \\ \tilde{z} = \tilde{g} & \text{on } \mathbb{R}^{n-1} \times \{t = 0\}. \end{cases} \tag{7.11$_\varepsilon$}$$

Here

$$\tilde{g}(x') \equiv \inf_{x_n} g_{x_n}(x', x_n) \qquad (x' \in \mathbb{R}^{n-1}). \tag{7.12}$$

In view of hypothesis (2.12),

$$\tilde{g} = 0 \quad \text{if} \quad |x'| \geq R. \tag{7.13}$$

Set

$$\min_{\mathbb{R}^{n-1}} \tilde{g} \equiv m. \tag{7.14}$$

We may assume $m < 0$, as otherwise we at once deduce that $\tilde{z} \geq 0$ in $\mathbb{R}^{n-1} \times [0, \infty)$.

Let us transform the PDE (7.11_ε) into the heat equation by writing

$$\phi(\tilde{z}) = \tilde{u}, \tag{7.15}$$

where

$$\phi(x) = \phi^\varepsilon(x) = \frac{1}{\varepsilon}\left[\arctan\left(\frac{x}{\varepsilon}\right) - \arctan\left(\frac{m}{\varepsilon}\right)\right] \qquad (x \in \mathbb{R}). \tag{7.16}$$

A calculation verifies

$$\tilde{u}_t = \Delta' \tilde{u} \quad \text{ in } \quad \mathbb{R}^{n-1} \times (0, \infty). \tag{7.17}$$

Now in view of (7.14), we have the normalization

$$\min_{\mathbb{R}^{n-1}} \tilde{u}(\,\cdot, 0) = 0, \tag{7.18}$$

and furthermore

$$\tilde{u}(\cdot, 0)|_{\mathbb{R}^{n-1} - B(0,R)} = -\varepsilon^{-1} \arctan\left(\frac{m}{\varepsilon}\right) \geq C\varepsilon^{-1} \tag{7.19}$$

since $m < 0$. Now,

$$\begin{aligned}
\tilde{u}(x', t) &= \frac{1}{(4\pi t)^{\frac{n-1}{2}}} \int_{\mathbb{R}^{n-1}} e^{-\frac{|x'-y'|^2}{4t}} \tilde{u}(y', 0) dy' \\
&\geq \frac{C}{\varepsilon (4\pi t)^{\frac{n-1}{2}}} \int_{\mathbb{R}^{n-1} - B(0,R)} e^{-\frac{|x'-y'|^2}{4t}} \, dy' \\
&\geq \frac{C}{\varepsilon t^{\frac{n-1}{2}}} \int_{\mathbb{R}^{n-1} - B(0,R)} e^{-\frac{|y'|^2}{4t}} \, dy' \\
&= \frac{C}{\varepsilon} \int_{\mathbb{R}^{n-1} - B(0, \frac{R}{t^{\frac{1}{2}}})} e^{-|z'|^2/4} \, dz'.
\end{aligned}$$

Fix any time $t_0 > 0$. Then if $x' \in \mathbb{R}^{n-1}$, $t \geq t_0$, there exists $\gamma = \gamma(t_0) > 0$ such that

$$\tilde{u}(x', t) \geq \gamma \varepsilon^{-1}.$$

Owing then to (7.15) and (7.16),

$$\arctan\left(\frac{\tilde{z}^\varepsilon(x', t)}{\varepsilon}\right) - \arctan\left(\frac{m}{\varepsilon}\right) \geq \gamma > 0, \tag{7.20}$$

and we have now restored the superscript ε on $\tilde{z}^\varepsilon$, the solution of (7.11_ε). Since $m < 0$,

$$\arctan\left(\frac{m}{\varepsilon}\right) = -\frac{\pi}{2} + o(1) \quad \text{ as } \varepsilon \to 0.$$

Consequently, (7.20) implies

$$\tilde{z}^\varepsilon(x, t) \geq \varepsilon \tan\left(\gamma - \frac{\pi}{2} + o(1)\right) \geq -C\varepsilon \qquad (x' \in \mathbb{R}^{n-1},\ t \geq t_0). \tag{7.21}$$

3. Now return to the PDE (7.10) and note that $\tilde{z} = \tilde{z}^\varepsilon$, regarded now as a function of $x \in \mathbb{R}^n$, $t \geq 0$, solves the PDE

$$\tilde{z}_t - a_{kl}\tilde{z}_{x_k x_l} = -\frac{2\tilde{z}}{\tilde{z}^2 + \varepsilon^2}|D'z|^2 + b\tilde{z}_{x_n} \quad \text{in } \mathbb{R}^n \times (0,\infty) \tag{7.22}$$

with

$$\tilde{z} \leq z \quad \text{on } \mathbb{R}^n \times \{t = 0\}. \tag{7.23}$$

We claim that, in fact,

$$\tilde{z} \leq z \text{ in } \mathbb{R}^n \times [0,\infty). \tag{7.24}$$

To see this, write $\tilde{u} = \phi(\tilde{z})$ (as in (7.15)) and $u = \phi(z)$. We invoke (7.10) and (7.16) to calculate

$$\begin{aligned} u_t - a_{kl}u_{x_k x_l} &= bu_{x_n} + 2a_{in}\frac{\phi''}{(\phi')^2}u_{x_i}u_{x_n} - \frac{a_{nn}\phi''}{(\phi')^2}\,u_{x_n}^2 \quad \text{in } \mathbb{R}^n \times (0,\infty) \\ &= cu_{x_n}, \end{aligned}$$

for

$$c = c(x,t) = b(x,t) + 2a_{in}\frac{\phi''(z)}{\phi'(z)^2}u_{x_i} - a_{nn}\frac{\phi''(z)}{\phi'(z)^2}u_{x_n}.$$

Now since $\tilde{u}$ solves the heat equation (7.17) and does not depend on the variable x_n, we have

$$\tilde{u}_t - a_{kl}\tilde{u}_{x_k x_l} = c\tilde{u}_{x_n} \qquad \text{in } \mathbb{R}^n \times (0,\infty)$$

as well. Because ϕ is strictly monotone,

$$\tilde{u} \leq u \quad \text{on} \quad \mathbb{R}^n \times \{t = 0\}.$$

Consequently, the maximum principle implies

$$\tilde{u} \leq u \quad \text{in} \quad \mathbb{R}^n \times [0,\infty);$$

whence (7.24) follows.

4. We finally restore the superscripts to the notation and rewrite (7.24) as

$$z^\varepsilon \leq z^{\varepsilon,\delta} = w_{x_n}^{\varepsilon,\delta} \qquad \text{in } \mathbb{R}^n \times [0,\infty).$$

Letting $\delta \to 0$, we deduce

$$z^\varepsilon \leq w_{x_n}^\varepsilon \qquad \text{a.e. in } \mathbb{R}^n \times [0,\infty), \tag{7.25}$$

w^ε the weak solution of the approximation (7.4$_\varepsilon$). Observe also that we have

$$w_{x_n}^\varepsilon \overset{*}{\rightharpoonup} v_{x_n} \tag{7.26}$$

weakly star in $L^\infty(\mathbb{R}^n \times (0,\infty))$. Finally, fix $t_0 > 0$. Then from (7.21) and (7.25), we deduce that

$$v_{x_n} \geq 0 \qquad \text{a.e. in } \mathbb{R}^n \times [t_0,\infty).$$

This conclusion is valid for each time $t_0 > 0$. □

REFERENCES

[1] S. Angenent, *The zero set of a solution of a parabolic equation*, J. Reine Angew. Math., 390 (1988), pp. 79–96.

[2] ———, *Nodal properties of solutions of parabolic equations*, Rocky Mountain J. Math., 21 (1991), pp. 585–592.

[3] G. W. Bluman and S. Kumei, *On the remarkable nonlinear diffusion equation* $(\partial/\partial x)[a(u+b)^{-2}(\partial u/\partial x)] - (\partial u/\partial t) = 0$, J. Math. Phys., 21 (1980), pp. 1019–1023.

[4] ———, *Symmetrics and Differential Equations*, Springer-Verlag, Berlin, New York, Heidelberg, 1989.

[5] A. M. Bruckstein and R. Kimmel, *Shape from shading in level sets*, Report CIS-9209, Computer Science Department, Technion, Haifa, Israel, 1992.

[6] J. Cahn, C. Handwerker, and J. Taylor, *Geometric models of crystal growth*, preprint.

[7] C. Caratheodory, *Calculus of Varieties of Partial Differential Equations of the First Order*, Chelsea, New York, 1982.

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

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

[10] ———, *Motion of level sets by mean curvature* II, Trans. Amer. Math. Soc., 330 (1992), pp. 321–332.

[11] ———, *Motion of level sets by mean curvature* III, J. Geom. Anal., 2 (1992), pp. 121–150.

[12] ———, *Motion of level sets by mean curvature* IV, J. Geom. Anal., 5 (1995), pp. 77–114.

[13] H. Federer, *Geometric Measure Theory*, Springer-Verlag, Berlin, New York, Heidelberg, 1969.

[14] A. S. Fokas and Y. C. Yortos, *On the exactly solvable equation* $S_t = [(\beta S+\gamma)^{-2}S_x]_x + \alpha(\beta S+\gamma)^{-2}S_x$ *occurring in two-phase flow in porous media*, SIAM J. Appl. Math., 42 (1982), pp. 318–332.

[15] M. Gage and R. Hamilton, *The heat equation shrinking convex plane curves*, J. Differential Geom., 23 (1986), pp. 69–96.

[16] M. Grayson, *The heat equation shrinks embedded plane curves to round points*, J. Differential Geom., 26 (1987), pp. 285–314.

[17] H. Ishii, *Degenerate parabolic PDE's with discontinuities and generalized evolutions of surfaces*, Adv. Differential Equations , 1 (1996), pp. 51–72.

[18] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[19] J. Oaks, *Singularities and self-intersections of curves evolving on surfaces*, Indiana U. Math J., 43 (1994), pp. 959–981.

[20] M. Ohnuma and M.-H. Sato, *Singular degenerate parabolic equations with applications to geometric evolutions*, Differential Integral Equations, 6 (1993), pp. 1265–1280.

[21] P. Olver, *Applications of Lie Groups to Differential Equations*, Springer-Verlag, Berlin, New York, Heidelberg, 1986.

[22] S. Osher, *A level set formulation for the solution of the Dirichlet problem for Hamilton–Jacobi equations*, SIAM J. Math. Anal., 24 (1993), pp. 1145–1153.

[23] S. Osher and J. Sethian, *Fronts propagation with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.

[24] M. Soner, *Motion of a set by the curvature of its boundary*, J. Differential Equations, 101 (1993), pp. 313–372.

[25] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley Interscience, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 959–976, July 1996

INVERTIBILITY AND A TOPOLOGICAL PROPERTY OF SOBOLEV MAPS*

STEFAN MÜLLER†, SCOTT J. SPECTOR‡, AND QI TANG§

Abstract. Let Ω be a bounded domain in $\mathbb{R}^n$, let $\mathbf{d} : \overline{\Omega} \to \overline{\Omega}$ be a homeomorphism, and consider a function $\mathbf{u} : \overline{\Omega} \to \mathbb{R}^n$ that agrees with $\mathbf{d}$ on $\partial\Omega$. If $\mathbf{u}$ is continuous and injective then $\mathbf{u}(\Omega) = \mathbf{d}(\Omega)$. Motivated by problems in nonlinear elasticity the relationship between $\mathbf{u}(\Omega)$ and $\mathbf{d}(\Omega)$ is analyzed when the continuity and invertibility assumptions are weakened. Specifically maps that are continuous on almost every line and maps that lie in the Sobolev space $W^{1,p}$ with $n-1<p<n$ are considered.

Key words. Sobolev spaces, elasticity, cavitation, singular minimizers, injectivity almost everywhere

AMS subject classifications. 73G05, 73C50, 46E35, 26B99

1. Introduction.

In this paper, we investigate a topological property of maps $\mathbf{u} : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ that lie in certain Sobolev classes. In view of applications to nonlinear elasticity, we are specifically interested in the extent to which the boundary data $\mathbf{u}|_{\partial\Omega}$ determines the image of Ω under $\mathbf{u}$, provided that $\mathbf{u}$ satisfies certain hypotheses of invertibility.

In the context of continuous maps, one has the following classical result. Let $\Omega \subset \mathbb{R}^n$ be a bounded domain (a nonempty, connected, open set). Suppose that $\mathbf{u}$ and $\mathbf{d}$ are continuous maps from $\overline{\Omega}$ into $\mathbb{R}^n$ and that the restriction of each map to Ω is injective. If $\mathbf{u}$ and $\mathbf{d}$ agree on the boundary, $\partial\Omega$, then $\mathbf{u}(\Omega) = \mathbf{d}(\Omega)$. For a continuously differentiable map $\mathbf{u}$, the global invertibility condition can be replaced by the pointwise condition $\det D\mathbf{u} \neq 0$. Ball has generalized this result to Sobolev maps.

THEOREM (see [BA 81]). *Let* $\mathbf{d}$ *and* Ω *be as above. Assume that* Ω *has Lipschitz boundary and that* $\mathbf{d}(\Omega)$ *satisfies the cone condition. Suppose that, for some* $p > n$ *and* $q > n$, $\mathbf{u} \in W^{1,p}(\Omega;\mathbb{R}^n) \cap C^0(\overline{\Omega};\mathbb{R}^n)$, $\det D\mathbf{u} > 0$ *a.e. and*

$$\int_\Omega |D\mathbf{u}(\mathbf{x})|^p + \frac{|\operatorname{adj} D\mathbf{u}(\mathbf{x})|^q}{(\det D\mathbf{u}(\mathbf{x}))^{q-1}} d\mathbf{x} < \infty. \tag{1.1}$$

If $\mathbf{u} = \mathbf{d}$ *on* $\partial\Omega$ *then* $\mathbf{u}|_\Omega$ *is a homeomorphism and* $\mathbf{u}(\Omega) = \mathbf{d}(\Omega)$.

Here adj $\mathbf{F}$ denotes the transpose of the matrix of cofactors of the matrix $\mathbf{F}$ and (1.1) is chosen so that the inverse of $\mathbf{u}$ (if it exists) lies in the space $W^{1,q}$. Ball's proof relies on subtle arguments that involve the Brouwer degree and approximation of a tentative inverse.

* Received by the editors February 25, 1994; accepted for publication (in revised form) October 21, 1994. This research was begun while the authors participated in the Program on Phase Transitions and Free Boundary Value Problems at the Institute for Mathematics and Its Applications in Minneapolis.

† Mathematisches Institut, Universität Freiburg, 79104 Freiburg, Germany. The research of this author was partially supported by University of Bonn grant SFB256 and National Science Foundation grant DMS-9002679.

‡ Department of Mathematics, Southern Illinois University, Carbondale, IL 62901. The research of this author was partially supported by the National Science Foundation and the U.S. Air Force Office of Scientific Research.

§ Department of Mathematics, University of Sussex, Brighton BN1 9QH, United Kingdom. The research of this author was partially supported by Nuffield Foundation grant SCI/180/91/220/G.

We note that the motivation for studying such properties of Sobolev maps arises in nonlinear elasticity. In order that such a map correspond to a reasonable physical notion of a deformation, it must satisfy some invertibility hypotheses. (For example, one might require that $\det D\mathbf{u} > 0$ a.e. and that $\mathbf{u}|_{\Omega\backslash N}$ be one-to-one, where N has measure zero.) Questions then arise as to the proper formulation of, the relationships among, and the topological implications of such hypotheses. For many materials, the expression on the left-hand side of (1.1) gives a lower bound (up to a constant) for the energy of an elastic body undergoing a deformation $\mathbf{u}$. However, for some materials, one only has the lower bound $\int_\Omega |D\mathbf{u}(\mathbf{x})|^p d\mathbf{x}$ for some $p < n$. Such materials allow discontinuous deformations that correspond to the formation of a hole or cavity in the material (see, e.g., [Ba 82]). Moreover, the formation of such cavities has been observed in experiments on such materials (see, e.g., [GL 58]). In this case, the above theorem of Ball does not apply and we propose to study the situation in more detail.

Our first result concerns maps that are continuous on almost every line segment. This is the case, for example, for (the precise representative of) maps in $W^{1,1}(\Omega;\mathbb{R}^n)$ (see [Mo 66]). In the following, we denote the k-dimensional Hausdorff measure by $\mathcal{H}^k$.

THEOREM TL (topological location theorem). *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain and let $\mathbf{d} : \overline{\Omega} \to \mathbb{R}^n$ be a homeomorphism. Suppose that $\mathbf{u} : \overline{\Omega} \to \mathbb{R}^n$ satisfies the following:*

(i) *$\mathbf{u}(\mathbf{x}) = \mathbf{d}(\mathbf{x})$ for every $\mathbf{x} \in \partial\Omega$;*

(ii) *there is a set $N \subset \Omega$ with $\mathcal{H}^{n-1}(N) = 0$ such that $\mathbf{u}|_{\overline{\Omega}\backslash N}$ is injective; and*

(iii) *each of the n functions $t \mapsto \mathbf{u}(t, x_2, \ldots, x_n), \ldots, t \mapsto \mathbf{u}(x_1, \ldots, x_{n-1}, t)$ is continuous on each line segment in Ω for $\mathcal{L}^{n-1}$ almost every value of the other independent variables. Then either*

$$\mathbf{u}(\mathbf{x}) \in \mathbf{d}(\Omega) \quad \textit{for a.e. } \mathbf{x} \in \Omega$$

or

$$\mathbf{u}(\mathbf{x}) \in \mathbb{R}^n\backslash\mathbf{d}(\overline{\Omega}) \quad \textit{for a.e. } \mathbf{x} \in \Omega.$$

Remarks. 1. If $\mathbf{u}$ is the precise representative (see (2.12) and Proposition 2.7) of a map in $W^{1,1}(\Omega;\mathbb{R}^n)$, then the conclusion of this theorem actually holds for $\mathcal{H}^{n-1}$ a.e. $\mathbf{x} \in \Omega$ (see the proof of Theorem AL in §4).

2. We have not been able to determine if the conclusion of the theorem is also valid if one replaces hypothesis (i) with the more natural hypothesis that $\mathbf{u} = \mathbf{d}$ $\mathcal{H}^{n-1}$ a.e. on $\partial\Omega$.

3. In §5, we give an example that shows that the conclusion of the theorem may not be valid if the set N satisfies $\mathcal{H}^{n-1}(N) > 0$.

4. The proof of this theorem, which is given in §3, exploits the fact that the image of "most" line segments cannot intersect $\mathbf{d}(\partial\Omega)$ and hence must remain in $\mathbf{d}(\Omega)$ or $\mathbb{R}^n\backslash\mathbf{d}(\overline{\Omega})$. The use of line segments in different directions yields the same conclusion for little cubes (minus a null set) and the fact that Ω is connected finishes the argument.

In contrast to the case of continuous $\mathbf{u}$, one cannot, in general, conclude that $\mathbf{u}(\Omega) = \mathbf{d}(\Omega)$ or even that $\mathbf{u}(\Omega) \subset \mathbf{d}(\Omega)$. For example, take Ω to be the unit ball and consider the maps

$$\mathbf{f}(\mathbf{x}) = \frac{1 + |\mathbf{x}|}{2|\mathbf{x}|}\,\mathbf{x} \tag{1.2}$$

and

$$\mathbf{g}(\mathbf{x}) = \frac{2 - |\mathbf{x}|}{|\mathbf{x}|}\,\mathbf{x}, \tag{1.3}$$

each of which is equal to the identity on the boundary of the unit ball, $B(\mathbf{0},1)$. However,

$$\mathbf{f}(B(\mathbf{0},1)) \subset B(\mathbf{0},1),$$
$$\mathbf{g}(B(\mathbf{0},1)) \subset \mathbb{R}^n \backslash \overline{B(\mathbf{0},1)}.$$

We next aim for conditions that distinguish between the two alternatives given in Theorem TL. To this end, we note that the map $\mathbf{f}$ given by (1.2) satisfies $\det D\mathbf{f} > 0$ a.e. and $\mathbf{f}(\Omega) \subset \mathbf{d}(\Omega)$, while the map $\mathbf{g}$ given by (1.3) satisfies $\det D\mathbf{g} < 0$ a.e. and $\mathbf{g}(\Omega) \subset \mathbb{R}^n \backslash \mathbf{d}(\overline{\Omega})$. The following result shows that the sign of the Jacobian does indeed distinguish between the two alternatives. For simplicity, we focus on the case when $\mathbf{d}$ is the identity and Ω is the unit ball. We refer to §2 for the definition of the precise representative.

THEOREM AL (analytic location theorem). *Let Ω be the unit ball in $\mathbb{R}^n$. Suppose that $\mathbf{u} \in W^{1,p}(\Omega; \mathbb{R}^n)$, with $p > n-1$. Let $\mathbf{u}^*$ denote the precise representative. Suppose that*

(i) $\mathbf{u}^*(\mathbf{x}) = \mathbf{x}$ *for $\mathcal{H}^1$ a.e. $\mathbf{x} \in \partial\Omega$;*

(ii) *there is a set $N \subset \Omega$ with $\mathcal{H}^1(N) = 0$ such that $\mathbf{u}|_{\overline{\Omega}\backslash N}$ is injective; and*

(iii) $\det D\mathbf{u} \neq 0$ *a.e. and* $\det D\mathbf{u} > 0$ *on a set of positive measure.*

Then

$$\mathbf{u}(\mathbf{x}) \in \Omega \text{ for } \mathcal{H}^1 \text{ a.e. } \mathbf{x} \in \Omega$$

and

$$\det D\mathbf{u} > 0 \text{ a.e.}$$

Remarks. 1. Various generalizations are possible. The unit ball can be replaced by a more general smooth (or even Lipschitz) domain. It suffices that the restriction to the boundary of the precise representative agree $\mathcal{H}^1$ a.e. with the restriction of a continuously differentiable diffeomorphism $\mathbf{d}$. (Weaker smoothness assumptions on $\mathbf{d}$ will also suffice.) These and other generalizations are left to the (technically) courageous reader.

2. In regard to hypothesis (ii), Theorem TL suggests that the condition $\mathcal{H}^{n-1}(N) = 0$ might be sufficient. However, we have been unable to prove or disprove such a statement. The heart of the matter seems to be whether a map $\mathbf{w} \in W^{1,p}(S^{n-1}; \mathbb{R}^n)$ that is injective $\mathcal{H}^{n-2}$ a.e. has a degree that is nonzero on exactly one component. This seems plausible since $\mathcal{H}^{n-2}$ null sets do not disconnect S^{n-1}.

3. In view of applications to nonlinear elasticity, it would be useful to determine conditions that imply hypothesis (ii) (as well as hypothesis (ii) of Theorem TL). In this regard, Ciarlet and Nečas [CN 87] give a criterion based upon the change of variables formula that ensures injectivity $\mathcal{L}^n$ a.e. (see Šverák [Šv 88] and Tang [Ta 88] for further extensions). No such criterion seems to be known for invertibility $\mathcal{H}^1$ a.e. or $\mathcal{H}^{n-1}$ a.e.

The proof of Theorem AL is simple in spirit but requires some technical preparation, which we have summarized in §2. The main idea is as follows. The first step is to show that

$$\det D\mathbf{u} > 0 \text{ a.e.} \quad \text{or} \quad \det D\mathbf{u} < 0 \text{ a.e.} \tag{1.4}$$

To this end, let $B(\mathbf{a}, r) \subset \Omega$ be the ball of radius r centered at $\mathbf{a}$. For $\mathcal{L}^1$ a.e. r, $\mathbf{u}^*|_{\partial B(\mathbf{a},r)}$ is injective and in $W^{1,p}(\partial B(\mathbf{a}, r); \mathbb{R}^n)$ (and hence continuous). By the Jordan separation theorem, $\mathbb{R}^n \backslash \mathbf{u}^*(\partial B(\mathbf{a}, r))$ consists of two components, one of which is bounded. Let U be the bounded component. Moreover, the Brouwer degree of $\mathbf{u}^*|_{\partial B(\mathbf{a},r)}$ in U is either $+1$ or -1. Assume the former for definiteness. In this case, the outward unit normal $\mathbf{n}$ to U can be expressed in terms of the cofactors of the tangential derivatives as

$$\mathbf{n} = \frac{(\Lambda_{n-1}D\mathbf{u})\boldsymbol{\nu}}{|(\Lambda_{n-1}D\mathbf{u})\boldsymbol{\nu}|} \tag{1.5}$$

(see Lemma 2.4), where $\boldsymbol{\nu}$ is the outward unit normal to $B(\mathbf{a}, r)$.

Next, by Theorem TL, either $\mathbf{u}(\mathbf{x}) \in U$ for a.e. $\mathbf{x} \in B(\mathbf{a}, r)$ or $\mathbf{u}(\mathbf{x}) \in \mathbb{R}^n \backslash \overline{U}$ for a.e. $\mathbf{x} \in B(\mathbf{a}, r)$. Suppose the latter. Now if $\mathbf{u}^*$ was C^1 at $\mathbf{z} \in \partial B(\mathbf{a}, r)$, it would follow from (1.5) that $\det D\mathbf{u}^*(\mathbf{z}) < 0$. This result can also be obtained in the current situation from an approximation argument (see Proposition 2.6 and [MS 95, Lem. 2.5]). In particular, $\det D\mathbf{u}$ is of constant sign on a.e. sphere and one easily deduces (1.4) (see Lemma 4.1).

For the second step, we argue by contradiction and assume that the second alternative in Theorem TL is satisfied. Extend $\mathbf{u}$ to an $\mathcal{H}^1$ a.e. injective map of $B(\mathbf{0}, 2)$ by letting $\mathbf{u}(\mathbf{x}) = \mathbf{x}/|\mathbf{x}|^2$ on $B(\mathbf{0}, 2) \backslash B(\mathbf{0}, 1)$. The first step applied to the extended map implies that $\det D\mathbf{u} < 0$ a.e. in $B(\mathbf{0}, 2)$, which is the desired contradiction.

Finally, we mention some related work. The idea of using degree-theoretic arguments on almost every sphere is due to Šverák [Šv 88] (see [MTY 94] for some extensions). This approach necessitates the assumption $p > n - 1$ in order that the maps be continuous on such spheres. Malý [Ma 93] recently used rather delicate estimates to show that in certain situations some information can be obtained about the borderline case $p = n - 1$. It is well known that for $p \leq n$ maps in $W^{1,p}$ may have very pathological behavior. Besicovitch [Be 50] constructed continuous maps from $\mathbb{R}^2$ into $\mathbb{R}^3$ that are in $W^{1,2}$ but that map a Lebesgue null set onto a set of positive $\mathcal{L}^3$-measure. Ponomarev [Po 87] obtained homeomorphisms in $W^{1,p}(p < n)$ that map null sets onto sets of positive measure and Malý and Martio [MM 92] solved a long-standing conjecture by exhibiting a continuous map in $W^{1,n}$ that satisfies $\det D\mathbf{u} = 0$ a.e. and maps a null set onto a set of positive measure. For $p > n$, such behavior is, of course, ruled out by the area formula of Marcus and Mizel [MM 73]. If $p = n$ and $\det D\mathbf{u} > 0$ a.e., then $\mathbf{u}$ has a continuous representative that maps null sets onto null sets (see [GV 76], [Šv 88], and [MZ 92]).

2. Preliminaries. In this section, we review the main analytical tools used in the paper. We first introduce some notation from multilinear algebra which will be useful in describing the behavior of surfaces under differentiable maps. One of our basic tools is the Brouwer degree, and we will make use of its representation as a boundary integral (Proposition 2.1) as well as the Jordan separation theorem (Proposition 2.2). We then review some facts about sets of finite perimeter. Together with the boundary representation for the Brouwer degree, these lead to a description of the generalized normal of the connected components of $\mathbb{R}^n/\mathbf{u}(\partial\Omega)$ (see Lemma 2.4). A second characterization of that normal is obtained in Proposition 2.6. Here the notion of approximate differentiability turns out to be very useful. The comparison of the characterizations in Lemma 2.4 and Proposition 2.6 is at the basis of our proof of Theorem AL. Finally, we recall the notion of the precise representative $\mathbf{u}^*$ of a Sobolev function $\mathbf{u} \in W^{1,p}$, which allows one to assign to $\mathbf{u}$ pointwise values off a set of Hausdorff dimension $n - p$.

We now introduce some notation from multilinear algebra (see, e.g., [Fe 69, Chap. 1] or [Sp 65, Chap. 4]). We denote by $\cdot$ the scalar product of vectors in $\mathbb{R}^n$ and by $\wedge$ their exterior product. For $n \geq 2$, we identify the space $\Lambda_{n-1}\mathbb{R}^n$ of alternating $n - 1$ tensors on $\mathbb{R}^n$ with $\mathbb{R}^n$ by means of the map

$$(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{n-1}) \mapsto \mathbf{a}_1 \wedge \mathbf{a}_2 \wedge \cdots \wedge \mathbf{a}_{n-1},$$

which is characterized by the following conditions:

(i) it is multilinear and alternating; and

(ii) if $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ is the canonical basis for $\mathbb{R}^n$, then $\mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_{i-1} \wedge \mathbf{e}_{i+1} \wedge \cdots \wedge \mathbf{e}_n = (-1)^{n-i}\mathbf{e}_i$.

If $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$, then $\mathbf{a} \wedge \mathbf{b}$ is the usual vector (cross) product.

We write $\mathrm{Lin}^{n\times m}$ for the set of all linear maps from $\mathbb{R}^m$ into $\mathbb{R}^n$ and $\mathrm{adj} : \mathrm{Lin}^{n\times n} \to \mathrm{Lin}^{n\times n}$ for the unique continuous function that satisfies

$$\mathbf{F}(\mathrm{adj}\,\mathbf{F}) = (\det \mathbf{F})\mathbf{Id} \tag{2.1}$$

for all $\mathbf{F} \in \mathrm{Lin}^{n\times n}$, where $\det \mathbf{F}$ is the determinant of $\mathbf{F}$ and $\mathbf{Id} \in \mathrm{Lin}^{n\times n}$ is the identity mapping. Thus, with respect to any orthonormal basis, the matrix corresponding to $\mathrm{adj}\,\mathbf{F}$ is the transpose of the cofactor matrix corresponding to $\mathbf{F}$. We note that $\mathrm{adj}\,\mathbf{F}$ satisfies

$$(\mathrm{adj}\,\mathbf{F})^T(\mathbf{a}_1 \wedge \mathbf{a}_2 \wedge \cdots \wedge \mathbf{a}_{n-1}) = \mathbf{F}\mathbf{a}_1 \wedge \mathbf{F}\mathbf{a}_2 \wedge \cdots \wedge \mathbf{F}\mathbf{a}_{n-1} \tag{2.2}$$

for all vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{n-1} \in \mathbb{R}^n$.

If V is an $(n-1)$-dimensional subspace of $\mathbb{R}^n$ and $\mathbf{L} : V \to \mathbb{R}^n$ is linear, then we define $\Lambda_{n-1}\mathbf{L} : \Lambda_{n-1}V \to \mathbb{R}^n$ by

$$(\Lambda_{n-1}\mathbf{L})(\mathbf{a}_1 \wedge \mathbf{a}_2 \wedge \cdots \wedge \mathbf{a}_{n-1}) = \mathbf{L}\mathbf{a}_1 \wedge \mathbf{L}\mathbf{a}_2 \wedge \cdots \wedge \mathbf{L}\mathbf{a}_{n-1}. \tag{2.3}$$

Let $\mathbf{v}$ be a unit vector normal to V. Then the one-dimensional subspace $\Lambda_{n-1}V$ can be identified with $\{\lambda\mathbf{v} : \lambda \in \mathbb{R}\}$. If $\tilde{\mathbf{L}} \in \mathrm{Lin}^{n\times n}$ is a linear extention of $\mathbf{L}$, then, by (2.2) and (2.3),

$$(\Lambda_{n-1}\mathbf{L})\mathbf{v} = (\mathrm{adj}\,\tilde{\mathbf{L}})^T\mathbf{v}. \tag{2.4}$$

In the following, Ω will denote a nonempty, bounded, open subset of $\mathbb{R}^n$, $n \geq 2$, whose boundary $\partial\Omega$ is smooth. We denote by $L^p(\Omega)$ and $W^{1,p}(\Omega)$ the spaces of p-summable and Sobolev functions, respectively. A vector-valued or matrix-valued function is in $L^p(W^{1,p})$ if all its components are, and we write, e.g., $L^p(\Omega; \mathbb{R}^n)$. The spaces $W^{1,p}(\partial\Omega)$ are defined by local charts (see, e.g., [Mo 66]).

We denote the n-dimensional Lebesgue measure by $\mathcal{L}^n$ and write $\mathcal{H}^k$ for the k-dimensional Hausdorff measure. We use the notation

$$B(\mathbf{a}, r) = \{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x} - \mathbf{a}| < r\}, \quad S(\mathbf{a}, r) := \partial B(\mathbf{a}, r).$$

We briefly recall some facts about the Brouwer degree (see, e.g., Schwartz [Sc 69] for more details). Let $\mathbf{u} : \overline{\Omega} \to \mathbb{R}^n$ be a C^∞ map. If $\mathbf{y}_0 \in R^n \backslash \mathbf{u}(\partial\Omega)$ is such that $\det D\mathbf{u}(\mathbf{x}) \neq 0$ for all $\mathbf{x} \in \mathbf{u}^{-1}(\mathbf{y}_0)$, one defines

$$\deg(\mathbf{u}, \Omega, \mathbf{y}_0) = \sum_{\mathbf{x} \in \mathbf{u}^{-1}(\mathbf{y})} \mathrm{sgn}\det D\mathbf{u}(\mathbf{x}).$$

If φ is a C^∞ function supported in the connected component of $R^n \backslash \mathbf{u}(\partial\Omega)$ that contains $\mathbf{y}_0$, one can show that

$$\int_\Omega (\varphi \circ \mathbf{u}) \det D\mathbf{u}\, d\mathbf{x} = \deg(\mathbf{u}, \Omega, \mathbf{y}_0) \int_{\mathbb{R}^n} \varphi\, d\mathbf{y}. \tag{2.5}$$

Using this formula and approximating by C^∞ functions, one can define $\deg(\mathbf{u}, \Omega, \mathbf{y})$ for any continuous function $\mathbf{u} : \overline{\Omega} \to \mathbb{R}^n$ and any $\mathbf{y} \in \mathbb{R}^n \backslash \mathbf{u}(\partial\Omega)$. Moreover, the degree only depends on $\mathbf{u}|_{\partial\Omega}$.

Indeed, if $\mathbf{u} \in C^\infty(\overline{\Omega}; R^n)$, then since Ω has smooth boundary, the degree can be expressed as a boundary integral as follows. First, recall the identity (see, e.g., [Mo

66, Lem. 4.4.6])

$$\operatorname{div}(\operatorname{adj} D\mathbf{u})^T = \mathbf{0}$$

or, in components,

$$\frac{\partial}{\partial x^j}(\operatorname{adj} D\mathbf{u})_i^j = 0 \quad \text{for } i = 1, \ldots, n.$$

Let φ be as above and let $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$ be C^∞ with $\operatorname{div}\mathbf{g} = \varphi$. Then the above identity in connection with (2.1) implies that

$$\operatorname{div}[(\operatorname{adj} D\mathbf{u})(\mathbf{g} \circ \mathbf{u})] = (\varphi \circ \mathbf{u}) \det D\mathbf{u}.$$

Thus by the divergence theorem and (2.4),

$$\int_\Omega (\varphi \circ \mathbf{u}) \det D\mathbf{u}\, d\mathbf{x} = \int_{\partial\Omega} (\mathbf{g} \circ \mathbf{u}) \cdot (\Lambda_{n-1} D\mathbf{u})\boldsymbol{\nu}\, d\mathcal{H}^{n-1},$$

and hence by (2.5),

$$\deg(\mathbf{u}, \partial\Omega, \mathbf{y}_0) \int_{\mathbb{R}^n} \varphi\, d\mathbf{y} = \int_{\partial\Omega} (\mathbf{g} \circ \mathbf{u}) \cdot (\Lambda_{n-1} D\mathbf{u})\boldsymbol{\nu}\, d\mathcal{H}^{n-1}. \tag{2.6}$$

Here $\boldsymbol{\nu}$ denotes the outward normal to $\partial\Omega$ and $D\mathbf{u}$ is viewed as a map from the tangent space of $\partial\Omega$ to $\mathbb{R}^n$.

PROPOSITION 2.1. *Let $p > n-1$. Suppose that $\overline{\mathbf{u}}$ is the continuous representative of a function in $W^{1,p}(\partial\Omega; \mathbb{R}^n)$. Then $\overline{\mathbf{u}}$ satisfies equation* (2.6). *Moreover, if $\mathbf{h} \in C^1(\mathbb{R}^n; \mathbb{R}^n)$ then*

$$\int_{\mathbb{R}^n} \deg(\overline{\mathbf{u}}, \partial\Omega, \mathbf{y})(\operatorname{div}\mathbf{h})(\mathbf{y})\, d\mathbf{y} = \int_{\partial\Omega} (\mathbf{h} \circ \overline{\mathbf{u}}) \cdot (\Lambda_{n-1} D\overline{\mathbf{u}})\boldsymbol{\nu}\, d\mathcal{H}^{n-1} \tag{2.7}$$

Remark. Note that $\overline{\mathbf{u}}$ is differentiable $\mathcal{H}^{n-1}$ a.e. on $\partial\Omega$ (see, e.g., [EG 92]) and $D\overline{\mathbf{u}}(\mathbf{x})$ is viewed as a linear map from the tangent space $T_\mathbf{x}(\partial\Omega)$ into $\mathbb{R}^n$.

Proof. For (2.6), see [Šv 88, Lem. 1]. In order to prove (2.7), we will use a slightly sharpened version of (2.6). Let U_i be a connected component of $\mathbb{R}^n \backslash \overline{\mathbf{u}}(\partial\Omega)$ and suppose that $\psi_i \in \mathcal{L}^q(\mathbb{R}^n)$, $q > n$, with support in U_i. Write $\mathbf{k}(\mathbf{z}) = c_n \mathbf{z}/|\mathbf{z}|^n$ and let $\mathbf{b}_i = \mathbf{k} * \psi_i$ so that $\operatorname{div}\mathbf{b}_i = \psi_i$. We claim that

$$\begin{aligned}\int_{\mathbb{R}^n} \deg(\overline{\mathbf{u}}, \partial\Omega, \mathbf{y})\psi_i\, d\mathbf{y} &= \deg(\overline{\mathbf{u}}, \partial\Omega, U_i) \int_{\mathbb{R}^n} \psi_i\, d\mathbf{y} \\ &= \int_{\partial\Omega} (\mathbf{b}_i \circ \overline{\mathbf{u}}) \cdot (\Lambda_{n-1} D\overline{\mathbf{u}})\boldsymbol{\nu}\, d\mathcal{H}^{n-1}.\end{aligned} \tag{2.8}$$

Indeed, this follows easily from (2.6) if we approximate ψ_i in $\mathcal{L}^q$ by $C_0^\infty(U_i)$ functions $\psi_i^{(j)}$ and use the fact that $\mathbf{b}_i^{(j)}$ converges in $W^{1,q}_{loc}(\mathbb{R}^n; \mathbb{R}^n)$ (and hence locally uniformly) and that $\overline{\mathbf{u}}(\partial\Omega)$ is compact.

We next observe that $\deg(\overline{\mathbf{u}}, \partial\Omega, \cdot)$ is integrable. To see this, it suffices to let $\psi_i = \chi_{U_i} \operatorname{sgn}(\deg(\overline{\mathbf{u}}, \partial\Omega, U_i))$ (with the convention $\operatorname{sgn}(0) = 0$) and to observe that $\mathbf{b} := \sum_i \mathbf{b}_i$ converges in $W^{1,q}_{loc}$ (and hence locally uniformly) since the degree vanishes on the unbounded component of $\mathbb{R}^n \backslash \overline{\mathbf{u}}(\partial\Omega)$.

Finally, let $\psi_i = \chi_{U_i} \operatorname{div}\mathbf{h}$ and note that $\mathcal{L}^n(\overline{\mathbf{u}}(\partial\Omega)) = 0$ by the area formula. Thus it follows from (2.8) and the dominated convergence theorem that

$$\int_{\mathbb{R}^n} \deg(\overline{\mathbf{u}}, \partial\Omega, \mathbf{y}) \operatorname{div}\mathbf{h}\, d\mathbf{y} = \int_{\partial\Omega} (\mathbf{b} \circ \overline{\mathbf{u}}) \cdot (\Lambda_{n-1} D\overline{\mathbf{u}})\boldsymbol{\nu}\, d\mathcal{H}^{n-1}.$$

Moreover, $\operatorname{div}\mathbf{b} = \operatorname{div}\mathbf{h}$ and hence (2.6) applied to $\mathbf{b}-\mathbf{h}$ implies that

$$\int_{\partial\Omega} (\mathbf{b}\circ\overline{\mathbf{u}})\cdot(\Lambda_{n-1}D\overline{\mathbf{u}})\boldsymbol{\nu}\,d\mathcal{H}^{n-1} = \int_{\partial\Omega} (\mathbf{h}\circ\overline{\mathbf{u}})\cdot(\Lambda_{n-1}D\overline{\mathbf{u}})\boldsymbol{\nu}\,d\mathcal{H}^{n-1},$$

which finishes the proof. □

We will use the following generalization of the Jordan curve theorem (see, e.g., [Sc 69, Thm. 3.21]).

PROPOSITION 2.2. *Assume that $\mathbb{R}^n\backslash\partial\Omega$ consists of exactly two connected components. Suppose that $\overline{\mathbf{u}} : \partial\Omega \to \mathbb{R}^n$ is continuous and injective. Then $\mathbb{R}^n\backslash\overline{\mathbf{u}}(\partial\Omega)$ consists of exactly two connected components U and V and $\partial U = \partial V = \overline{\mathbf{u}}(\partial\Omega)$. If we denote by U the bounded component, we have*

$$\deg(\overline{\mathbf{u}},\partial\Omega,\mathbf{y}) = \begin{cases} m & \text{if } \mathbf{y}\in U, \\ 0 & \text{if } \mathbf{y}\notin \mathbb{R}^n\backslash\overline{U}, \end{cases}$$

where $m\in\{-1,1\}$.

Next, we briefly recall some properties of functions of bounded variation and sets of finite perimeter. More details can be found in [EG 92], [Gi 84], or [Zi 89]. A function $\varphi \in L^1(\mathbb{R}^n)$ is said to be of bounded variation ($\varphi\in BV(\mathbb{R}^n)$) if its distributional derivatives are signed measures, i.e., if

$$\|D\varphi\|_{\mathcal{M}} := \sup\left\{\int_{\mathbb{R}^n} \varphi \operatorname{div}\mathbf{w}\,d\mathbf{y} : \mathbf{w}\in C_0^1(\mathbb{R}^n;\mathbb{R}^n), |\mathbf{w}|\le 1\right\} < \infty.$$

A measurable set A is said to have finite perimeter if its characteristic function χ_A is in $BV(\mathbb{R}^n)$. Sets of finite perimeter enjoy a surprising degree of regularity. Specifically, there exists a set $\partial^* A$ (the reduced boundary) and a function $\boldsymbol{\nu} : \partial^*A \to \mathbb{R}^n$ (the generalized outward unit normal, denoted $\boldsymbol{\nu}(\mathbf{a},A)$) such that the distributional derivative $D\chi_A$ satisfies

$$D\chi_A = -\boldsymbol{\nu}\mathcal{H}^{n-1}\lfloor\,\partial^*A$$

(where $\lfloor$ denotes the restriction) and

$$\mathcal{H}^{n-1}(\partial^*A) = \|D\chi_A\|_{\mathcal{M}}, \quad |\boldsymbol{\nu}| = 1 \quad \mathcal{H}^{n-1} \text{ a.e. on } \partial^*A.$$

Moreover, $\boldsymbol{\nu}$ agrees with the measure-theoretic normal $\mathcal{H}^{n-1}$ a.e. on ∂^*A, i.e., for $\mathcal{H}^{n-1}$ a.e. $\mathbf{a}\in\partial^*A$, one has

$$\lim_{r\to 0+} \frac{\mathcal{L}^n\{\mathbf{x}\in B(\mathbf{a},r)\cap A : (\mathbf{x}-\mathbf{a})\cdot\boldsymbol{\nu}(\mathbf{a},A)\ge 0\}}{r^n} = 0, \tag{2.9}$$

$$\lim_{r\to 0+} \frac{\mathcal{L}^n\{\mathbf{x}\in B(\mathbf{a},r)\backslash A : (\mathbf{x}-\mathbf{a})\cdot\boldsymbol{\nu}(\mathbf{a},A)\le 0\}}{r^n} = 0. \tag{2.10}$$

In order to compare the outward unit normals of Ω and $\mathbf{u}(\Omega)$, we will use the following version of the area formula.

PROPOSITION 2.3 (see [MM 73], [Fe 69, Cor. 3.2.20]). *Let Γ be an oriented, smooth, $(n-1)$-dimensional manifold with continuous unit normal field $\boldsymbol{\nu}$. Suppose that $\mathbf{u}\in W^{1,p}(\Gamma;\mathbb{R}^n)\cap C^0(\Gamma;\mathbb{R}^n)$ with $p > n-1$. Then for any $\mathcal{H}^{n-1}$-measurable $A\subset\Gamma$,*

$$\mathcal{H}^{n-1}(\mathbf{u}(A)) \le \int_A |(\Lambda_{n-1}D\mathbf{u})\boldsymbol{\nu}|\,d\mathcal{H}^{n-1} \le \int_A |D\mathbf{u}|^{n-1}\,d\mathcal{H}^{n-1}.$$

Moreover, if $\varphi : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{H}^{n-1}$-measurable and $\mathbf{u}$ *is injective then*

$$\int_A (\varphi \circ \mathbf{u}) |(\Lambda_{n-1} D\mathbf{u})\boldsymbol{\nu}|\, d\mathcal{H}^{n-1} = \int_{\mathbf{u}(A)} \varphi \, d\mathcal{H}^{n-1}$$

whenever either integrand is in L^1.

Remark. One consequence of the above result that we will need is that the image under $\mathbf{u}$ of the set $\{\mathbf{x} \in \Gamma : (\Lambda_{n-1} D\mathbf{u}(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x}) = \mathbf{0}\}$ is an $\mathcal{H}^{n-1}$ null set.

LEMMA 2.4. *Let Ω have smooth boundary with outward unit normal $\boldsymbol{\nu}$. Suppose that $\mathbb{R}^n \backslash \partial\Omega$ has exactly two connected components. Let $\mathbf{u} \in W^{1,p}(\partial\Omega; \mathbb{R}^n) \cap C^0(\partial\Omega; \mathbb{R}^n)$, with $p > n-1$, and assume that* $\mathbf{u}$ *is injective. Denote by U the bounded component of $\mathbb{R}^n \backslash \mathbf{u}(\partial\Omega)$ (cf. Proposition 2.2). Then U has finite perimeter, the reduced boundary $\partial^* U$ agrees with $\partial U = \mathbf{u}(\partial\Omega)$ up to an $\mathcal{H}^{n-1}$ null set, and the generalized normal $\tilde{\boldsymbol{\nu}}$ satisfies*

$$\tilde{\boldsymbol{\nu}}(\mathbf{y}, U) = m \frac{(\Lambda_{n-1} D\mathbf{u})\boldsymbol{\nu}}{|(\Lambda_{n-1} D\mathbf{u})\boldsymbol{\nu}|} (\mathbf{u}^{-1}(\mathbf{y})) \quad \textit{for } \mathcal{H}^{n-1} \textit{ a.e. } \mathbf{y} \in \mathbf{u}(\partial\Omega),$$

where $m = \deg(\mathbf{u}, \partial\Omega, U)$. If, moreover, $|(\Lambda_{n-1} D\mathbf{u})\boldsymbol{\nu}| > 0$ $\mathcal{H}^{n-1}$ a.e., then

$$\tilde{\boldsymbol{\nu}}(\mathbf{u}(\mathbf{x}), U) = m \frac{(\Lambda_{n-1} D\mathbf{u}(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x})}{|(\Lambda_{n-1} D\mathbf{u}(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x})|} \quad \textit{for } \mathcal{H}^{n-1} \textit{ a.e. } \mathbf{x} \in \partial\Omega.$$

Proof. Define

$$\tilde{\boldsymbol{\nu}}(\mathbf{y}) = \begin{cases} \mathbf{0} & \text{if } \Lambda_{n-1} D\mathbf{u}(\mathbf{u}^{-1}(\mathbf{y}))\boldsymbol{\nu}(\mathbf{u}^{-1}(\mathbf{y})) = \mathbf{0}, \\ \frac{(\Lambda_{n-1} D\mathbf{u})\boldsymbol{\nu}}{|(\Lambda_{n-1} D\mathbf{u})\boldsymbol{\nu}|} (\mathbf{u}^{-1}(\mathbf{y})) & \text{otherwise,} \end{cases}$$

and note that, by Proposition 2.3, $|\tilde{\boldsymbol{\nu}}(\mathbf{y})| = 1$ for $\mathcal{H}^{n-1}$ a.e. $\mathbf{y}$. Let $\mathbf{g} \in C^1(\mathbb{R}^n; \mathbb{R}^n)$. Then, by the area formula with $\varphi = \mathbf{g} \cdot \tilde{\boldsymbol{\nu}}$,

$$\int_{\partial\Omega} (\mathbf{g} \circ \mathbf{u}) \cdot (\Lambda_{n-1} D\mathbf{u})\boldsymbol{\nu} \, d\mathcal{H}^{n-1} = \int_{\mathbf{u}(\partial\Omega)} \mathbf{g} \cdot \tilde{\boldsymbol{\nu}} \, d\mathcal{H}^{n-1},$$

and hence, by Propositions 2.1 and 2.2,

$$\begin{aligned} \int_{\mathbf{u}(\partial\Omega)} \mathbf{g} \cdot \tilde{\boldsymbol{\nu}} \, d\mathcal{H}^{n-1} &= \int_{\mathbb{R}^n} \deg(\mathbf{u}, \partial\Omega, \mathbf{y})(\operatorname{div} \mathbf{g})(\mathbf{y}) \, d\mathbf{y} \\ &= m \int_{\mathbb{R}^n} \chi_U (\operatorname{div} \mathbf{g}) \, d\mathbf{y}, \end{aligned} \tag{2.11}$$

where $m \in \{-1, 1\}$.

Finally, it is clear from (2.11) that U has finite perimeter and

$$D\chi_U = -m\boldsymbol{\nu} \mathcal{H}^{n-1} \lfloor \mathbf{u}(\partial\Omega). \qquad \square$$

We next give a characterization of the (measure-theoretic) normal of the image $\mathbf{u}(D)$ of subsets $D \subset \Omega$. To this end, it is useful to consider the approximate derivative of $\mathbf{u}$.

DEFINITION 2.5. *Let $E \subset \mathbb{R}^n$ be a measurable set. For $\mathbf{x} \in \mathbb{R}^n$, we define the upper and lower density of E at $\mathbf{x}$ by*

$$D^+(\mathbf{x}, E) = \limsup_{r \to 0+} \frac{\mathcal{L}^n(B(\mathbf{x}, r) \cap E)}{\omega_n r^n},$$

$$D^-(\mathbf{x}, E) = \liminf_{r \to 0+} \frac{\mathcal{L}^n(B(\mathbf{x}, r) \cap E)}{\omega_n r^n},$$

where ω_n *denotes the volume of the unit ball in* $\mathbb{R}^n$. *If the two values agree, their common value is the* density, $D(\mathbf{x}, E)$, *of* E *at* $\mathbf{x}$. *We will sometimes say that* $\mathbf{x}$ *is a point of density* $D(\mathbf{x}, E)$ *of* E.

Let $\mathbf{u} : E \to \mathbb{R}^m$ be measurable. We say that $\mathbf{u}$ is *approximately differentiable* at $\mathbf{x}$ if there is an $\mathbf{F} \in \mathrm{Lin}^{m\times n}$ such that for every $\varepsilon > 0$,

$$\lim_{r\to 0+} \frac{\mathcal{L}^n\{\mathbf{z} \in B(\mathbf{x},r) \cap E : |\mathbf{u}(\mathbf{z}) - \mathbf{u}(\mathbf{x}) - \mathbf{F}(\mathbf{x}-\mathbf{z})| < \varepsilon r\}}{\omega_n r^n} = 1,$$

and we write

$$(\mathrm{ap}\, D\mathbf{u})(\mathbf{x}) := \mathbf{F}.$$

Remark. By the Lebesgue point theorem, $D(\mathbf{x}, E) = 1$ for $\mathcal{L}^n$ a.e. $\mathbf{x} \in E$. If E is open and $\mathbf{u}$ is (a representative of an equivalence class) in $W^{1,1}(E)$ then $\mathrm{ap}\, D\mathbf{u}$ exists $\mathcal{L}^n$ a.e. and agrees with the distributional derivative (see, e.g., [Fe 69, Thms. 3.1.4 and 4.5.9] or [EG 92]).

If $\mathbf{u} : \overline{\Omega} \to \mathbf{u}(\overline{\Omega})$ is a diffeomorphism then for any $\mathbf{x} \in \partial\Omega$ the outward unit normal $\mathbf{n}$ to $\mathbf{u}(\overline{\Omega})$ at $\mathbf{u}(\mathbf{x})$ is given by

$$\mathbf{n}(\mathbf{u}(\mathbf{x})) = \mathrm{sgn}(\det D\mathbf{u}(\mathbf{x})) \frac{(\Lambda_{n-1} D\mathbf{u}(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x})}{|(\Lambda_{n-1} D\mathbf{u}(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x})|}.$$

The following proposition shows that a suitable version of this formula is also valid for Sobolev functions. See, e.g., [MS 95, Lem. 2.5] for a proof.

PROPOSITION 2.6. *Let* $\mathbf{u} : \Omega \to \mathbb{R}^n$ *be (a representative of an equivalence class) in* $W^{1,1}(\Omega; \mathbb{R}^n)$ *and assume that* $\det D\mathbf{u} \neq 0$ $\mathcal{L}^n$ *a.e. Then there is an* $\Omega_1 \subset \Omega$ *of full measure such that*

(i) $\mathbf{u}$ *is approximately differentiable on* Ω_1*;*

(ii) $\det(\mathrm{ap}\, D\mathbf{u}) \neq 0$ *on* Ω_1*;*

(iii) *if* $E \subset\subset \Omega$ *has smooth boundary,* $\mathbf{x}_0 \in \partial E \cap \Omega_1$, $\boldsymbol{\nu}$ *is the outward unit normal to* E *at* $\mathbf{x}_0$*, and*

$$\mathbf{n} = \mathrm{sgn} \det((\mathrm{ap}\, D\mathbf{u})(\mathbf{x}_0)) \frac{(\Lambda_{n-1}(\mathrm{ap}\, D\mathbf{u})(\mathbf{x}_0))\boldsymbol{\nu}}{|(\Lambda_{n-1}(\mathrm{ap}\, D\mathbf{u})(\mathbf{x}_0))\boldsymbol{\nu}|}.$$

Then for any $r > 0$ *and any* $\mathcal{L}^n$ *null set* N*, the set*

$$\mathbf{u}((E\backslash N) \cap B(\mathbf{x}_0, r)) \cap \{\mathbf{y} : (\mathbf{y} - \mathbf{u}(\mathbf{x}_0)) \cdot \mathbf{n} \leq 0\}$$

has density $1/2$ *at* $\mathbf{u}(\mathbf{x}_0)$.

Remarks. 1. Note that $\det((\mathrm{ap}\, D\mathbf{u})(\mathbf{x}_0)) \neq 0$ implies $(\Lambda_{n-1}(\mathrm{ap}\, D\mathbf{u})(\mathbf{x}_0))\boldsymbol{\nu} \neq \mathbf{0}$. Also, E need not be smooth. It suffices that E have a measure-theoretic normal $\boldsymbol{\nu}$ at $\mathbf{x}_0$.

2. If $\mathbf{u}(E\backslash N)$ has a measure-theoretic normal $\boldsymbol{\nu}$ at $\mathbf{u}(\mathbf{x}_0)$ (cf. (2.9) and (2.10)), then the above proposition yields $\tilde{\boldsymbol{\nu}} = \mathbf{n}$.

Since we are interested in the behavior of Sobolev functions on lower-dimensional sets, it is useful to consider a specific representative. The *precise representative* $\mathbf{u}^*$: $\Omega \to \mathbb{R}^n$ of an equivalence class $\{\mathbf{u}\} \in W^{1,p}(\Omega; \mathbb{R}^n)$ is defined by

$$\mathbf{u}^*(\mathbf{x}) = \begin{cases} \lim\limits_{r\to 0+} ⨍_{B(\mathbf{x},r)} \mathbf{u}(\mathbf{z})d\mathbf{z} & \text{if the limit exists,} \\ \mathbf{0} & \text{otherwise,} \end{cases} \tag{2.12}$$

where $⨍_A$ denotes the average value of the integrand over A, i.e., the integral of the function over A divided by the n-dimensional Lebesgue measure of A.

The definition of $\mathbf{u}^*$ at points where the above limit does not exist is somewhat arbitrary. For a thorough discussion of precise representatives, we refer to [EG 92] or [Zi 89]. The main properties of $\mathbf{u}^*$ are summarized below.

PROPOSITION 2.7. *Let* $\mathbf{u} \in W^{1,p}(\Omega;\mathbb{R}^n)$ *with* $1 \leq p < n$. *Let* $p^* := np/(n-p)$ *be the Sobolev conjugate exponent. Then there are sets* P *and* P' *with* $\operatorname{cap}_p(P) = \operatorname{cap}_p(P') = 0$ *such that the following properties are satisfied:*

(i) (*existence of the limit*). *The limit in the definition of* $\mathbf{u}^*$ *exists for all* $\mathbf{x} \in \Omega\backslash P$ *and, if* $\rho_r(\cdot) = r^{-n}\rho(\frac{\cdot}{r})$ *is the standard family of molifiers,*

$$(\rho_r * \mathbf{u})(\mathbf{x}) \to \mathbf{u}^*(\mathbf{x})$$

for each $\mathbf{x} \in \Omega\backslash P$.

(ii) (*Lebesgue points*). *For any* $\mathbf{x} \in \Omega\backslash P$,

$$\lim_{r\to 0+} \fint_{B(\mathbf{x},r)} |\mathbf{u}(\mathbf{z}) - \mathbf{u}^*(\mathbf{x})|^{p^*} d\mathbf{z} = 0.$$

(iii) (*invariance upon change of variables*). *If* $\mathbf{g}$ *is a bi-Lipschitz map, then*

$$(\mathbf{u}\circ\mathbf{g})^* = \mathbf{u}^*\circ\mathbf{g} \text{ on } \mathbf{g}^{-1}(\Omega)\backslash P'.$$

(iv) (*continuity on lines*). $\mathbf{u}^*$ *is absolutely continuous on* $\mathcal{L}^{n-1}$ *a.e. line segment parallel to the coordinate axes.*

(v) (*restriction to hyperplanes*). *For* $\mathcal{L}^1$ *a.e. coordinate hyperplane* H *the restriction* $\mathbf{v} := \mathbf{u}^*|_{H\cap\Omega}$ *is in* $W^{1,p}(H\cap\Omega;\mathbb{R}^n)$ *and* $\operatorname{ap} D\mathbf{v} = (\operatorname{ap} D\mathbf{u})|_{H\cap\Omega}\,\mathcal{H}^{n-1}$ *a.e. on* H. *If, moreover,* $p > n-1$, *then* $\mathbf{v}$ *is continuous.*

Remarks. 1. The property $\operatorname{cap}_p(P) = 0$ implies that $\mathcal{H}^\delta(P) = 0$ for every $\delta > n-p$. For $p > n-1$ one has in particular $\mathcal{H}^1(P) = 0$.

2. Consider $p > n-1$. If we use polar coordinates, it follows from (iii) and (v) that $\mathbf{u}^*$ is continuous on the spheres $S(\mathbf{a}, r)$ for $\mathcal{L}^1$ a.e. r.

3. The equality of the approximate derivatives in (v) is a consequence of the same equality for the corresponding distributional derivatives and the fact that each of the approximate derivatives is a representative of (the equivalence class that contains) the corresponding distributional derivative.

In the following, we will be concerned with maps $\mathbf{u} \in W^{1,p}(\Omega;\mathbb{R}^n)$ and we wish to define $\mathbf{u}^*$ on $\partial\Omega$. Let $E : W^{1,p}(\Omega;\mathbb{R}^n) \to W^{1,p}(\mathbb{R}^n;\mathbb{R}^n)$ be an extension operator (which exists if $\partial\Omega$ is Lipschitz, see, e.g., [GT 83, Thm. 7.25]) and define

$$\mathbf{u}^*(\mathbf{x}) := (E\mathbf{u})^*(\mathbf{x}) \quad \text{for } \mathbf{x} \in \partial\Omega. \tag{2.13}$$

A slightly different approach is to first consider the trace $\mathbf{v} = \mathbf{u}|_{\partial\Omega}$ (which is well defined as an equivalence class in $L^p(\partial\Omega)$). If $\mathbf{c} : U \subset \mathbb{R}^{n-1} \to \Gamma$ is a chart of Γ, one could then define

$$\mathbf{u}^{**}(\mathbf{c}(\mathbf{x})) := (\mathbf{v}\circ\mathbf{c})^*(\mathbf{x}). \tag{2.14}$$

It turns out that the two definitions agree up to a set of Hausdorff dimension $n-p$.

LEMMA 2.8. *Let* $A = \{\mathbf{x} : \mathbf{u}^*(\mathbf{x}) \neq \mathbf{u}^{**}(\mathbf{x})\}$. *Then* $H^\delta(A) = 0$ *for every* $\delta > n-p$.

Remark. This result shows that (2.13) and (2.14) are essentially independent of the choice of E and $\mathbf{c}$. The result is probably well known to experts, but we have included a proof for the convenience of the reader.

Proof. As is usual in such situations, we locally flatten $\partial\Omega$ to reduce the problem to the following case. Let $\mathbf{u} \in W^{1,p}(\mathbb{R}^n;\mathbb{R}^n)$, $H = \mathbb{R}^{n-1}\times\{0\}$, and let $\mathbf{v}$ be the trace of $\mathbf{u}$ on H. Define $A = \{\mathbf{x} \in H : \mathbf{u}^*(\mathbf{x}) \neq \mathbf{v}^*(\mathbf{x})\}$. Then we must show that $\mathcal{H}^\delta(A) = 0$.

We note that the definition of trace gives us

$$\int_V |\mathbf{u}(\mathbf{x}',x_n)-\mathbf{v}(\mathbf{x}')|\,d\mathbf{x}' \le \int_V \int_0^{x_n} \left|\frac{\partial \mathbf{u}}{\partial x_n}\right|(\mathbf{x}',t)\,dt\,d\mathbf{x}'$$

for any (measurable) $V \subset \mathbb{R}^{n-1}$. Thus, if we write $\mathbf{a} = (\mathbf{a}',0)$ and let $Q(\mathbf{a},r)$ be the cube $\mathbf{a}+r(-\frac{1}{2},\frac{1}{2})^n$ and $Q'(\mathbf{a}',r) = \mathbf{a}'+r(-\frac{1}{2},\frac{1}{2})^{n-1}$, we find that

$$\left| \fint_{Q'(\mathbf{a}',r)} \mathbf{v}(\mathbf{x}')\,d\mathbf{x}' - \mathbf{u}^*(\mathbf{a})\right|$$

$$\le \fint_{Q(\mathbf{a},r)} \left|\mathbf{v}(\mathbf{x}')-\mathbf{u}(\mathbf{x}',x_n)\right| d\mathbf{x}'\,dx_n + \fint_{Q(\mathbf{a},r)} \left|\mathbf{u}(\mathbf{x})-\mathbf{u}^*(\mathbf{a})\right| d\mathbf{x}$$

$$\le r \fint_{Q(\mathbf{a},r)} \left|\frac{\partial \mathbf{u}}{\partial x_n}\right| d\mathbf{x} + \fint_{Q(\mathbf{a},r)} \left|\mathbf{u}(\mathbf{x})-\mathbf{u}^*(\mathbf{a})\right| d\mathbf{x}.$$

Therefore, $A \subset A_1 \cup A_2$, where

$$A_1 = \left\{\mathbf{a}\in\mathbb{R}^n : \limsup_{r\to 0^+} \; r \fint_{Q(\mathbf{a},r)} |D\mathbf{u}|\,d\mathbf{x} > 0\right\}$$

$$\subset \left\{\mathbf{a}\in\mathbb{R}^n : \limsup_{r\to 0^+} \; r^{p-n} \int_{Q(\mathbf{a},r)} |D\mathbf{u}|^p d\mathbf{x} > 0\right\} =: \hat{A}_1,$$

$$A_2 = \left\{\mathbf{a}\in\mathbb{R}^n : \limsup_{r\to 0^+} \fint_{Q(\mathbf{a},r)} \left|\mathbf{u}(\mathbf{x})-\mathbf{u}^*(\mathbf{a})\right| d\mathbf{x} > 0\right\}.$$

Thus, by [EG 92, §2.4.3, Thm. 3], $\mathcal{H}^{n-p}(\hat{A}_1) = 0$, while Proposition 2.7(ii) implies that $\mathcal{H}^\delta(A_2) = 0$ for all $\delta > n-p$. □

3. Proof of Theorem TL. To simplify the notation, we let $n = 3$. First, note that since $\mathbf{d}$ is continuous and one-to-one, it is an open mapping (see, e.g., [Sc 69, Cor. 3.2]). In particular, $\mathbf{d}(\Omega)$ is open and hence (see, e.g., [Ci 88, Thm. 1.2–7]) $\partial\mathbf{d}(\Omega) \subset \mathbf{d}(\partial\Omega)$. Thus, if $\mathbf{u}$ is continuous on any line segment $L \subset \Omega$, then either

(I) $\mathbf{u}(L) \subset \mathbf{d}(\Omega)$;
(E) $\mathbf{u}(L) \subset \mathbb{R}^n \backslash \mathbf{d}(\overline{\Omega})$; or
(B) $\mathbf{u}(L)$ contains a point $\mathbf{y} = \mathbf{d}(\mathbf{x})$ with $\mathbf{x} \in \partial\Omega$.

Moreover, in the last case, the point $\mathbf{m} \in L$ for which $\mathbf{y} = \mathbf{u}(\mathbf{m})$ must be contained in the set N of hypothesis (ii) with $\mathcal{H}^2(N) = 0$.

Next, let $\mathbf{a} \in \Omega$, $\delta > 0$ be such that the open cube

$$C := (a_1, a_1+\delta) \times (a_2, a_2+\delta) \times (a_3, a_3, +\delta) \subset \Omega$$

and define

$$C_{ij} = (a_i, a_i+\delta) \times (a_j, a_j+\delta)$$

for $i,j = 1,2,3$, $i<j$, the projection of C onto the i,j-plane. Then, in particular, by hypothesis (iii), there is an $\mathcal{L}^2$-measurable set $S_{23} \subset C_{23}$ with $\mathcal{L}^2(S_{23}) = \mathcal{L}^2(C_{23}) = \delta^2$

such that the function

$$t \mapsto \mathbf{u}(t, z_2, z_3), \quad t \in (a_1, a_1, +\delta)$$

is continuous for each $z = (z_2, z_3) \in S_{23}$.

Let S_I (S_E) be the subset of S_{23} such that the image under $\mathbf{u}$ of $(a_1, a_1 + \delta) \times S_I$ ($(a_1, a_1 + \delta) \times S_E$) is contained in $\mathbf{d}(\Omega)$ ($\mathbb{R}^n \backslash \mathbf{d}(\overline{\Omega})$). Note that by the first paragraph of this proof, $S_I \cap S_E = \emptyset$ and the set $S_{23} \backslash (S_I \cup S_E)$ must be contained in the projection of the $\mathcal{H}^2$ null set N onto the 2, 3-plane. Thus the $\mathcal{L}^2$-measure of $S_{23} \backslash (S_I \cup S_E)$ is zero and hence

$$\mathcal{L}^2(S_I \cup S_E) = \delta^2.$$

We will now show that S_I and S_E are each $\mathcal{L}^2$-measurable and one of these sets has $\mathcal{L}^2$-measure zero while the other has $\mathcal{L}^2$-measure δ^2. It will then follow that

$$\mathbf{u}(\mathbf{x}) \in \mathbf{d}(\Omega) \quad \text{for a.e. } \mathbf{x} \in C, \tag{3.1}$$

or

$$\mathbf{u}(\mathbf{x}) \in \mathbb{R}^n \backslash \mathbf{d}(\overline{\Omega}) \quad \text{for a.e. } \mathbf{x} \in C. \tag{3.2}$$

Let L_I (L_E) be the projection of S_I (S_E) on the x_2-axis and define (see Figure 1)

$$A_{12} := (a_1, a_1 + \delta) \times (L_I \cap L_E) \subset C_{12}.$$

Let $\mathbf{z} = (z_1, z_2) \in A_{12}$. Then by construction there are $z_3^I, z_3^E \in (a_3, a_3 + \delta)$ such that $(z_2, z_3^I) \in S_I$ and $(z_2, z_3^E) \in S_E$. Thus the image under $\mathbf{u}$ of the line segment $\{(t, z_2, z_3^I) : a_1 < t < a_1 + \delta\}$ is contained in $\mathbf{d}(\Omega)$ while the image under $\mathbf{u}$ of the line segment $\{(t, z_2, z_3^E) : a_1 < t < a_1 + \delta\}$ is contained in $\mathbb{R}^n \backslash \mathbf{d}(\overline{\Omega})$. We note that the line segment $\{(z_1, z_2, t) : a_3 < t < a_3 + \delta\}$ intersects both of these line segments and hence that $\mathbf{u}$ cannot be continuous on this line. Therefore, by hypothesis (iii), $\mathcal{L}^2(A_{12}) = 0$ and hence $\mathcal{L}^1(L_I \cap L_E) = 0$.

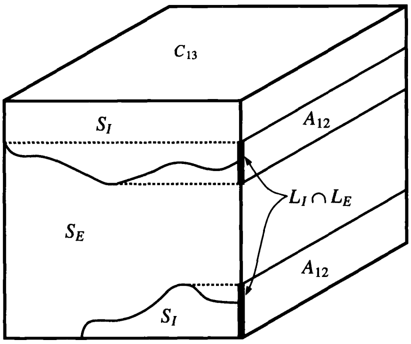


FIG. 1. *The cube C.*

Let L_I^* (L_E^*) be the projection of S_I (S_E) on the x_3-axis. Then by the argument used in the previous paragraph, $\mathcal{L}^1(L_I^* \cap L_E^*) = 0$. Now,

$$S_I \subset L_I \times L_I^* \subset C_{23}, \quad S_E \subset L_E \times L_E^* \subset C_{23},$$

and hence

$$S_I \cup S_E \subset (L_I \times L_I^*) \cup (L_E \times L_E^*) \subset C_{23}.$$

Since $\mathcal{L}^2(S_I \cup S_E) = \mathcal{L}^2(C_{23}) = \delta^2$, we conclude that the set $(L_I \times L_I^*) \cup (L_E \times L_E^*)$ is $\mathcal{L}^2$-measurable with measure δ^2 and hence, in particular, that

$$\mathcal{L}^2(L_I \times L_E^*) = 0.$$

A standard result from real analysis (e.g., Fubini's theorem) then implies that at least one of the sets L_I or L_E^* is $\mathcal{L}^1$-measurable with measure zero. Consequently, either S_I or S_E has $\mathcal{L}^2$-measure zero which gives us (3.1) or (3.2).

Next, let Ω_I (Ω_E) be the union of all open cubes $C \subset \Omega$, whose edges are parallel to the coordinate axes, such that (3.1) ((3.2)) is satisfied. Then Ω_I and Ω_E are open subsets of Ω and, by the above argument, $\Omega = \Omega_I \cup \Omega_E$. We note that $\Omega_I \cap \Omega_E$ is open. If it were nonempty, then it would contain an open cube C that would satisfy (3.1) and (3.2), which is not possible. Thus $\Omega_I \cap \Omega_E = \emptyset$. However, Ω is connected and hence either Ω_I or Ω_E must be empty.

Finally, let C_i, $i = 1, 2, 3, \ldots$ be an ordering of those open cubes in Ω_I or Ω_E that have edges of rational length and whose centers have rational coordinates. Then by the density of the rationals $\Omega = \bigcup_{i=1}^{\infty} C_i$. Now, by the conclusion of the previous paragraph, there are $M_i \subset C_i$ such that $\mathcal{L}^3(M_i) = 0$ and either $\mathbf{u}(C_i \backslash M_i) \subset \mathbf{d}(\Omega)$ for every i or $\mathbf{u}(C_i \backslash M_i) \subset \mathbb{R}^n \backslash \mathbf{d}(\overline{\Omega})$ for every i. Thus if we define $M := \bigcup_{i=1}^{\infty} M_i$, we conclude that $\mathcal{L}^3(M) = 0$ and either $\mathbf{u}(\Omega \backslash M) \subset \mathbf{d}(\Omega)$ or $\mathbf{u}(\Omega \backslash M) \subset \mathbb{R}^n \backslash \mathbf{d}(\overline{\Omega})$, which is the desired result. □

We note that Proposition 2.2 (the Jordan separation theorem) together with the proof of Theorem TL yield the following result, which we will use in the next section.

COROLLARY 3.1. *Let Ω be a domain in $\mathbb{R}^n$. Suppose that $\mathbf{u} : \overline{\Omega} \mapsto \mathbb{R}^n$ satisfies hypotheses* (ii) *and* (iii) *of Theorem TL. Assume further that the restriction $\mathbf{u}|_{\partial\Omega}$ is continuous and injective, that $\mathbb{R}^n \backslash \Omega$ consists of exactly two connected components, and hence* (*see Proposition* 2.2) *that $\mathbb{R}^n \backslash \mathbf{u}(\partial\Omega)$ consists of exactly two connected components U and V. Then*

$$\mathbf{u}(\mathbf{x}) \in U \quad \textit{for a.e. } \mathbf{x} \in \Omega \qquad \textit{or} \qquad \mathbf{u}(\mathbf{x}) \in V \quad \textit{for a.e. } \mathbf{x} \in \Omega.$$

4. Proof of Theorem AL.

We show first that for a function that satisfies the regularity and invertibility properties of Theorem AL, the Jacobian is of one sign.

LEMMA 4.1. *Let Ω be a domain in $\mathbb{R}^n$ and let $\mathbf{v} \in W^{1,p}(\Omega; \mathbb{R}^n)$ with $p > n - 1$. Suppose that $\det D\mathbf{v} \neq 0$ $\mathcal{L}^n$ a.e. and that there is an $\mathcal{H}^1$ null set $M \subset \Omega$ such that the precise representative $\mathbf{v}^*$ is injective on $\Omega \backslash M$. Then*

$$\det D\mathbf{v} > 0 \quad \mathcal{L}^n \textit{ a.e.} \qquad \textit{or} \qquad \det D\mathbf{v} < 0 \quad \mathcal{L}^n \textit{ a.e.}$$

Proof. We first show that $\operatorname{sgn}\det(\operatorname{ap} D\mathbf{v})$ is constant ($\mathcal{H}^{n-1}$ a.e.) on "most" spheres. Let $\mathbf{a} \in \Omega$ and $r_{\mathbf{a}} := \operatorname{dist}(\mathbf{a}, \partial\Omega)$. In view of Proposition 2.7, there exists an $\mathcal{L}^1$ null set $N_{\mathbf{a}}$ such that for $r \in (0, r_{\mathbf{a}}) \backslash N_{\mathbf{a}}$ one has

$$\begin{gathered}
\mathbf{v}^*|_S \in W^{1,p}(S; \mathbb{R}^n) \cap C^0(S; \mathbb{R}^n), \\
(\operatorname{ap} D)(\mathbf{v}^*|_S)(\mathbf{x}) = (\operatorname{ap} D\mathbf{v}^*)|_{T_{\mathbf{x}}S}(\mathbf{x}) \quad \text{for } \mathcal{H}^{n-1} \text{ a.e. } \mathbf{x} \in S, \\
|(\Lambda_{n-1}(\operatorname{ap} D\mathbf{v}^*))\boldsymbol{\nu}| > 0 \quad \text{for } \mathcal{H}^{n-1} \text{ a.e. on } S, \\
\mathcal{H}^{n-1}(S \backslash \Omega_1) = 0, \\
S \cap M = \emptyset,
\end{gathered}$$

where Ω_1 is the set that occurs in Proposition 2.6 and $\boldsymbol{\nu} = \frac{\mathbf{x}-\mathbf{a}}{|\mathbf{x}-\mathbf{a}|}$ is the outward unit to $S := S(\mathbf{a}, r)$. (To verify the third assertion, note that if $(\Lambda_{n-1}(\operatorname{ap} D\mathbf{v}^*)(\mathbf{x}))\mathbf{e} = \mathbf{0}$ for some $\mathbf{e} \neq \mathbf{0}$, then $\det((\operatorname{ap} D\mathbf{v}^*)(\mathbf{x})) = 0$.)

Let U be the bounded component of $\mathbb{R}^n\backslash\mathbf{v}^*(S(\mathbf{a},r))$ (see Proposition 2.2). Then by Corollary 3.1, there is an $\mathcal{L}^n$ null set N_{TL} such that

$$\mathbf{v}^*(B(\mathbf{a},r)\backslash N_{TL}) \subset U \tag{4.1}$$

or

$$\mathbf{v}^*(B(\mathbf{a},r)\backslash N_{TL}) \subset \mathbb{R}^n\backslash\overline{U}. \tag{4.2}$$

Also, by Lemma 2.4, U has finite perimeter and the generalized outer normal satisfies

$$\boldsymbol{\nu}(\mathbf{v}^*(\mathbf{x}),U) = m\frac{(\Lambda_{n-1}D\mathbf{v}^*(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x})}{|(\Lambda_{n-1}D\mathbf{v}^*(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x})|} \quad \text{for } \mathcal{H}^{n-1} \text{ a.e.} \quad \mathbf{x}\in S(\mathbf{a},r),$$

where m is a constant, which is either $+1$ or -1.

We now apply Proposition 2.6 and recall the definition (2.9) and (2.10) of the measure-theoretic normal. If (4.1) holds, one deduces

$$\operatorname{sgn}\det((\operatorname{ap}D\mathbf{v})(\mathbf{x})) = m \quad \text{for } \mathcal{H}^{n-1} \text{ a.e.} \quad \mathbf{x}\in S(\mathbf{a},r),$$

while (4.2) gives

$$\operatorname{sgn}\det((\operatorname{ap}D\mathbf{v})(\mathbf{x})) = -m \quad \text{for } \mathcal{H}^{n-1} \text{ a.e.} \quad \mathbf{x}\in S(\mathbf{a},r).$$

It follows that $\operatorname{sgn}\det((\operatorname{ap}D\mathbf{v})(\mathbf{x}))$ is constant on $S(\mathbf{a},r)$ ($\mathcal{H}^{n-1}$ a.e.) for $r\in(0,r_{\mathbf{a}})\backslash N_{\mathbf{a}}$.

Define

$$\psi := \operatorname{sgn}\det(\operatorname{ap}D\mathbf{v}), \quad \Omega^{\pm} := \{\mathbf{x}\in\Omega : \psi(\mathbf{x}) = \pm1\}.$$

We will show that the precise representative, ψ^*, of ψ is locally constant and the desired result will follow from the fact that Ω is connected.

Let $\mathbf{x}\in\Omega$ and assume for definiteness that $\psi^*(\mathbf{x}) = +1$. Suppose for the sake of contradiction that $\mathbf{y}\in B(\mathbf{x},r_{\mathbf{x}})$ satisfies $\psi^*(\mathbf{y}) = -1$. Define $\mathbf{a} = (\mathbf{x}+\mathbf{y})/2$, $\rho := |\mathbf{x}-\mathbf{y}|/2$, and note that $B(\mathbf{a},\rho)\subset\subset\Omega$. Also, define

$$A^{\pm} := \{r\in(0,r_{\mathbf{a}})\backslash N_{\mathbf{a}} : \mathcal{H}^{n-1}(\Omega^{\pm}\cap S(\mathbf{a},r)) > 0\}$$

and note that the sets $A^{\pm}$ are $\mathcal{L}^1$-measurable since the functions $t\mapsto\mathcal{H}^{n-1}(S(\mathbf{a},t)\cap\Omega^{\pm})$ are measurable (by Fubini's theorem or the coarea formula).

Now, since $\psi^*(\mathbf{x}) = +1$ and $|\mathbf{x}-\mathbf{a}| = \rho$, it follows that A^+ has density 1 at ρ. Similarly, since $\varphi^*(\mathbf{y}) = -1$, A^- must have density 1 at ρ. However, this is not possible since $A^+\cap A^- = \emptyset$. Therefore, $\varphi^*(\mathbf{y}) = +1$ for a.e. $\mathbf{y}\in B(\mathbf{x},r_{\mathbf{x}})$ which implies that $\varphi^*(\mathbf{y}) = +1$ for *every* $\mathbf{y}\in B(\mathbf{x},r_{\mathbf{x}})$. Since Ω is connected this gives the desired result. □

Proof of Theorem AL. We note that hypothesis (iii) and Lemma 4.1 yield $\det D\mathbf{u} > 0$ a.e. In order to prove that $\mathbf{u}(\mathbf{x})\in B = B(\mathbf{0},1)$, we will first show that there is an $\mathcal{H}^1$ null set N such that either

$$\mathbf{u}^*(B\backslash N)\subset B, \tag{4.3}$$

or

$$\mathbf{u}^*(B\backslash N)\subset\mathbb{R}^n\backslash\overline{B}. \tag{4.4}$$

Indeed, *define* a function $\tilde{\mathbf{u}} : \overline{B}\to\mathbb{R}^n$

$$\tilde{\mathbf{u}}(\mathbf{x}) := \begin{cases} \mathbf{u}^*(\mathbf{x}), & \mathbf{x}\in B, \\ \mathbf{x}, & \mathbf{x}\in\partial B. \end{cases}$$

Then $\tilde{\mathbf{u}}$ satisfies the hypotheses of Theorem TL since $\mathbf{u}^*(\mathbf{x}) = \mathbf{x}$ for $\mathcal{H}^1$ a.e. $\mathbf{x}\in\partial B$. (see Proposition 2.7 and Lemma 2.8). Thus we obtain (4.3) or (4.4) with $\mathcal{L}^n(N) = 0$.

Next, let P be the cap_p null set of Proposition 2.7. Then, since B is open, (4.3) and Proposition 2.7(ii) yield $\mathbf{u}^*(B\backslash P) \subset B$, while (4.4) and Proposition 2.7(ii) give us $\mathbf{u}^*(B\backslash P) \subset \mathbb{R}^n\backslash\overline{B}$.

If (4.3) is satisfied, then we are done. If instead (4.4) is satisfied, define an extension of $\mathbf{u}$ to $B(\mathbf{0},2)$ by

$$\mathbf{w}(\mathbf{x}) := \begin{cases} \mathbf{u}^*(\mathbf{x}), & \mathbf{x} \in B, \\ \mathbf{x}/|\mathbf{x}|^2, & \mathbf{x} \in B(\mathbf{0},2)\backslash B. \end{cases}$$

Then $\mathbf{w} \in W^{1,p}(B(\mathbf{0},2),\mathbb{R}^n)$ and by Lemma 2.8 $\mathbf{w}^*(\underset{\sim}{\mathrm{x}}) = \mathbf{w}(\mathbf{x})$ for $\mathcal{H}^1$ a.e. $\mathbf{x} \in B(\mathbf{0},2)$. Thus, by hypothesis (ii) and (4.4), $\mathbf{w}$ is injective off an $\mathcal{H}^1$ null set and hence $\mathbf{w}$ satisfies the hypotheses of Lemma 4.1. However, $\det D\mathbf{w} > 0$ a.e. on B and $\det D\mathbf{w} < 0$ a.e. on $B(\mathbf{0},2)\backslash\overline{B}$, which contradicts the conclusion of Lemma 4.1. Therefore, (4.3) must instead be satisfied. □

5. Counterexamples. We present two examples which show that hypothesis (ii) of Theorem TL cannot be relaxed to allow $\mathcal{H}^{n-1}(N) > 0$. For the first example, take $\Omega = B = B(\mathbf{0},1) \subset \mathbb{R}^2$, the unit disk. We construct a map $\mathbf{u} \in W^{1,p}(B;\mathbb{R}^2)$ (for all $p < 2$) that satisfies $\mathbf{u}|_{\partial B} = \mathbf{id}$ and whose image is $(B\backslash B(\mathbf{b},1/2)) \cup B(\mathbf{a},1/2)$, where $\mathbf{a} = (3/2,0)$ and $\mathbf{b} = (1/2,0)$ (see Figure 2). This will be achieved as follows. Each circle $S(\mathbf{0},r)$ is mapped to a union of two (reversely oriented) circles that touch only at the point $(1,0)$. These circles are nested (for different values or r) in such a way that as $r \longrightarrow 1$ the circles on the left increase to the unit circle while those on the right shrink to the point $(1,0)$. As $r \longrightarrow 0$, the circles approach the circles $S(\mathbf{a},1/2)$ and $S(\mathbf{b},1/2)$, respectively (see Figure 2). At the same time as the circles on the left increase, their preimage also occupies a larger position of $S(\mathbf{0},r)$. In such a way, it is possible to make $\mathbf{u}$ (Lipschitz) continuous away from the origin with $\mathbf{u} = \mathbf{id}$ on ∂B.

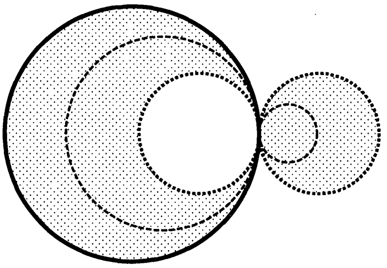


FIG. 2. *The image of a circle $S(\mathbf{0},R)$ for different values of R : $R = 1$ (solid line), $R = 1/2$ (dashed line), and asymptotically as $r \longrightarrow 0$ (dotted line). The filled region is the image under $\mathbf{u}$ of the unit disk.*

We now proceed with a detailed description of $\mathbf{u}$ and a verification of its properties. The circles on the left and the right are parametrized, respectively, by

$$\gamma_1(s,t) = (1-s,0) + s(\cos t, \sin t) \tag{5.1}$$

and

$$\gamma_2(s,t) = (1+s,0) + s(\cos t, -\sin t). \tag{5.2}$$

Let (R,θ) be polar coordinates in $\mathbb{R}^2$, and let

$$r = \frac{1+R}{2}$$

and

$$\mathbf{u}(R,\theta) = \begin{cases} \gamma_1\big(r, r^{-1}(\theta - (1-r)\pi)\big) & \text{if} \quad \pi(1-r) \le \theta \le 2\pi - \pi(1-r), \\ \gamma_2\big(1-r, (1-r)^{-1}\theta - \pi\big) & \text{if} \quad |\theta| < \pi(1-r). \end{cases}$$

One easily checks that $\mathbf{u}|_{\partial B} = \mathbf{id}$ and that $\mathbf{u}$ is Lipschitz away from the origin where $|\nabla \mathbf{u}|$ has a $1/R$ singularity. Hence $\mathbf{u} \in W^{1,p}(B; \mathbb{R}^2)$ for all $p < 2$ and, in particular, $\mathbf{u}$ is continuous on every straight line segment that does not pass through the origin. Now circles defined by (5.1) and (5.2) only have the point $(1,0)$ in common and this point corresponds to $t = 0$. Hence $\mathbf{u}|_{B\setminus N}$ is injective where

$$N = \left\{ (R\cos\theta, R\sin\theta) : \theta = \pm\pi\left(1 - \frac{R+2}{2}\right) \right\}$$

and $\mathcal{H}^1(N) > 0$. Note that one also has $\det D\mathbf{u} > 0$ a.e.

The function $\mathbf{u}$ is Lipschitz on $B\setminus\{\mathbf{0}\}$, but not C^1, since $\frac{\partial \mathbf{u}}{\partial\theta}$ is discontinuous at $\theta = \pm\pi(1-r)$. This could be fixed by reversing the orientation of the circles on the right, i.e., replacing $\gamma_2(s,t)$ by $\gamma_2(s,-t)$. Then, however, one loses the property $\det D\mathbf{u} > 0$ a.e.

For our second example, we take $\Omega = (-1,1) \times (-1,1) \subset \mathbb{R}^2$ and consider the composition of a number of maps, each of which has a simple physical interpretation. First, let

$$\mathbf{h}_0(x,y) = \frac{R+3}{4R}(x,y), \quad R = \max\{|x|, |y|\},$$

which opens a square hole at the center of the body (see Figure 3). We then compose with the map

$$\mathbf{h}_1(x,y) = \begin{cases} \big(x, 1 - (1-y)(7 - 8|x|)\big) & \text{if} \quad |x| < \frac{3}{4}, \quad \frac{3}{4} < y < 1, \\ \big(x, -1 + \frac{4}{3}|x|(y+1)\big) & \text{if} \quad |x| < \frac{3}{4}, \ -1 < y < -\frac{3}{4}, \\ (x,y) & \text{at all other points in the range of } \mathbf{h}_0, \end{cases}$$

which "pinches" part of the hole to the boundary and stretches another portion of the hole.

Next, we compose with the map

$$\mathbf{h}_2(x,y) = \begin{cases} \big(\frac{8xy}{4y+3}, y\big) & \text{if} \quad |x| < \frac{4y+3}{8}, \quad 0 \le y < \frac{3}{4}, \\ \big(\frac{-8xy}{4y+3}, y\big) & \text{if} \quad |x| < \frac{4y+3}{8}, \ -\frac{1}{4} < y \le 0, \\ (x,y) & \text{at all other points in the range of } \mathbf{h}_1 \circ \mathbf{h}_0, \end{cases}$$

which "pinches" the previously stretched portion of the hole. Finally, we compose with the map

$$\mathbf{h}_3(x,y) = \begin{cases} \big(x, 1 - \frac{(1-y)(6-7x)}{3(1-x)}\big) & \text{if} \quad |x| < \frac{3}{4}, \quad 0 \le y < 1, \\ (x, y-1) & \text{if} \quad |x| < \frac{3}{4}, \ -\frac{1}{4} < y \le 0, \\ (x,y) & \text{at all other points in the range of } \mathbf{h}_2 \circ \mathbf{h}_1 \circ \mathbf{h}_0, \end{cases}$$

which stretches the portion of the hole that was just pinched and moves a portion of it "outside" the body.

We note that each of the maps $\mathbf{h}_1, \mathbf{h}_2$, and $\mathbf{h}_3$ is Lipschitz on the range of the previous map and hence has a Lipschitz extention (see, e.g., [Fe 69, Thm. 2.10.43] or [EG 92]) to all of $\mathbb{R}^2$. In addition, Ball and Murat [BM 84] have shown that

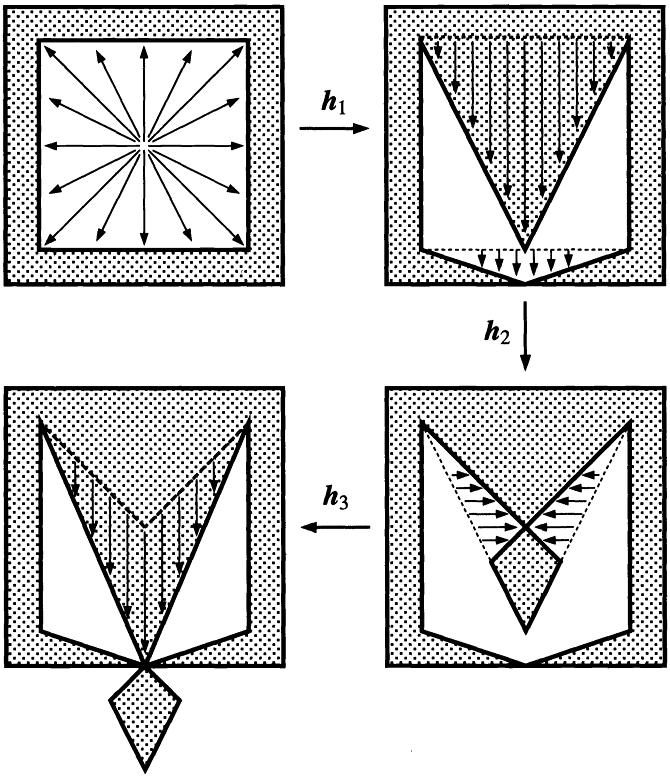


FIG. 3. *After cavitation, successive stretches and pinches cause a portion of the material to penetrate the boundary.*

$\mathbf{h}_0 \in W^{1,p}(\Omega;\mathbb{R}^2)$ for all $1 \leq p < 2$. Thus, by the chain rule (see, e.g., [Zi 89]), $\mathbf{u} = \mathbf{h}_3 \circ \mathbf{h}_2 \circ \mathbf{h}_1 \circ \mathbf{h}_0 \in W^{1,p}(\Omega;\mathbb{R}^2)$.

It is clear that one could also construct such an example that is C^1 away from the cavitation and pinching points.

REFERENCES

[Ba 81] J. M. BALL, *Global invertibility of Sobolev functions and the interpenetration of matter*, Proc. Roy. Soc. Edinburgh, 88A (1981), pp. 315–328.

[Ba 82] ———, *Discontinuous equilibrium solutions and cavitation in nonlinear elasticity*, Philos. Trans. Roy. Soc. London, 306A (1982), pp. 557–612.

[Be 50] A. S. BESICOVITCH, *Parametric surfaces*, Bull. Amer. Math. Soc., 56 (1950), pp. 228–296.

[Ci 88] P. G. CIARLET, *Mathematical Elasticity*, vol. I, North-Holland, Amsterdam, 1988.

[CN 87] P. G. CIARLET AND J. NEČAS, *Injectivity and self-contact in non-linear elasticity*, Arch. Rational Mech. Anal., 97 (1987), pp. 171–188.

[EG 92] L. C. EVANS AND R. F. GARIEPY, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[Fe 69] H. FEDERER, *Geometric Measure Theory*, Springer-Verlag, Berlin, New York, 1969.

[Gi 84] E. GIUSTI, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Basel, Switzerland, 1984.

[GL 58] A. G. GENT AND P. B. LINDLEY, *Internal rupture of bonded rubber cylinders in tension*, Proc. Roy. Soc. London Ser. A, 49 (1958), pp. 195–205.

[GT 83] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, New York, 1983.

[GV 76] V. M. GOLDSTEIN AND S. K. VODOPYANOV, *Quasiconformal mappings and spaces of func-*

tions with generalized first derivatives, Sibirsk. Mat. Zh., 17 (1976), pp. 515–531 (in Russian); Siberian Math. J., 17 (1976), pp. 399–411 (in English).

[Ma 93] J. MALÝ, *Weak lower semicontinuity of polyconvex integrals*, Proc. Roy. Soc. Edinburgh, 123A (1993), pp. 681–691.

[MM 73] M. MARCUS AND V. J. MIZEL, *Transformations by functions in Sobolev spaces and lower semicontinuity for parametric variational problems*, Bull. Amer. Math. Soc., 79 (1973), pp. 790–795.

[MM 92] J. MALÝ AND O. MARTIO, *Lusin's condition (N) and mappings of the class $W^{1,n}$*, J. Reine Angew. Math., 458 (1995), pp. 19–36.

[MZ 92] O. MARTIO AND W. ZIEMER, *Lusin's condition (N) and mappings with nonnegative Jacobian*, Michigan Math. J., 39 (1992), pp. 495–508.

[Mo 66] C. B. MORREY, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, New York, 1966.

[MS 95] S. MÜLLER AND S. J. SPECTOR, *An existence theory for nonlinear elasticity that allows for cavitation*, Arch. Rational Mech. Anal., 131 (1995), pp. 1–66.

[MTY 94] S. MÜLLER, Q. TANG, AND B. S. YAN, *On a new class of elastic deformations not allowing for cavitation*, Anal. Non Linéaire, 11 (1994), pp. 217–243.

[Po 87] S. P. PONOMAREV, *Property N of homeomorphisms of the class $W^{1,p}$*, Siberian Math. J., 28 (1987), pp. 291–298.

[Sc 69] J. T. SCHWARTZ, *Nonlinear Functional Analysis*, Gordon and Breach, New York, 1969.

[Sp 65] M. SPIVAK, *Calculus on Manifolds*, W. A. Benjamin, New York, 1965.

[Šv 88] V. ŠVERÁK, *Regularity properties of deformations with finite energy*, Arch. Rational Mech. Anal., 100 (1988), pp. 105–127.

[Ta 88] Q. TANG, *Almost-everywhere injectivity in nonlinear elasticity*, Proc. Roy. Soc. Edinburgh, 109A (1988), pp. 79–95.

[Zi 89] W. ZIEMER, *Weakly Differentiable Functions*, Springer-Verlag, Berlin, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 977–996, July 1996

QUASI-LINEAR RELAXED DIRICHLET PROBLEMS*

STEFANO FINZI VITA[†], FRANÇOIS MURAT[‡], AND NICOLETTA A. TCHOU[§]

Abstract. We study the existence and the asymptotic behavior of solutions of quasi-linear elliptic problems with homogeneous Dirichlet boundary conditions for the so-called *relaxed Dirichlet problems*. These problems, introduced in the linear case by Dal Maso and Mosco [*Arch. Rational Mech. Anal.*, 95 (1986), pp. 345–387; *Appl. Math. Optim.*, 15 (1987), pp. 15–63], involve zero-order terms with Borel measures, which can take infinite values but vanish on sets with zero capacity.

We prove two existence results when the nonlinear term has quadratic growth with respect to the gradient by extending the techniques of Boccardo, Murat, and Puel [*Res. Notes in Math.* 84, Pitman, London, 1983, pp. 19–73] to the relaxed case. We also prove in the subquadratic case a stability property for bounded solutions with respect to the γ-convergence of measures when the limit measure is sufficiently regular, making essential use of the correctors result of Finzi Vita and Tchou [*Asymptotic Anal.*, 5 (1992), pp. 269–281].

Key words. nonlinear elliptic problems, homogenization

AMS subject classifications. 35J65, 35B27

1. Introduction.

We are interested in the study of quasi-linear relaxed Dirichlet problems that can *formally* be written as

$$(1.1)\qquad \begin{cases} -\Delta u + \lambda_0 u + \mu u = f(x, u, Du) & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where Ω is a bounded domain in $\mathbb{R}^N$, λ_0 is a nonnegative constant, f is a given function that satisfies a quadratic growth hypothesis with respect to Du, and μ is a measure in the class $M_0(\Omega)$ of all nonnegative Borel measures on Ω that vanish on subsets of Ω with zero harmonic capacity. We remark that this definition allows $\mu(B) = +\infty$ on some Borel subset B of Ω with positive capacity, as shown by the example of the measure ∞_S that is defined, for a given subset S of Ω with $\text{cap}(S) > 0$, by

$$\infty_S(A) := \begin{cases} 0 & \text{if } \text{cap}(S \cap A) = 0, \\ +\infty & \text{if } \text{cap}(S \cap A) > 0, \end{cases}$$

for any Borel subset A of Ω.

To give a precise and correct meaning to solutions of problems such as (1.1), we first recall the meaning of the solution in the linear case, i.e., for problems that can be formally written as

$$(1.2)\qquad \begin{cases} -\Delta u + \mu u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

* Received by the editors April 25, 1994; accepted for publication (in revised form) October 25, 1994.

[†] Dipartimento di Matematica, Università degli Studi di Roma l, Piazzale Aldo Moro 2, 00185 Rome, Italy.

[‡] Laboratoire d'Analyse Numérique, Université de Paris 6, 4 Place Jussieu, 75252 Paris cedex 05, France.

[§] Laboratoire de Mathématiques et Informatique, Faculté des Sciences, Université de Picardie "Jules Verne", 9 Rue St. Leu, 80000 Amiens, France.

for a given $f \in H^{-1}(\Omega)$. Problems of type (1.2) have been introduced by G. Dal Maso and U. Mosco in [DM2] to study limits of Dirichlet problems in highly perturbed domains (see also [CM]).

For every $\mu \in M_0(\Omega)$, we denote by $L^2_\mu(\Omega)$ the space of square integrable functions with respect to the measure μ and by $V_\mu(\Omega)$ the Hilbert space $H^1_0(\Omega) \cap L^2_\mu(\Omega)$, equipped with the scalar product

$$((v,w))_\mu := \int_\Omega DvDw\,dx + \int_\Omega vw\,d\mu.$$

The solution u of problem (1.2) is then defined as the unique function of $V_\mu(\Omega)$, which solves

$$((u,v))_\mu = \langle f,v\rangle_{H^{-1},H^1_0} \quad \text{for all } v \in V_\mu(\Omega), \tag{1.3}$$

or, equivalently, as the unique minimizing point in $V_\mu(\Omega)$ of the functional $((v,v))_\mu - 2\langle f,v\rangle_{H^{-1},H^1_0}$.

We refer to [DM1] and [DM2] for detailed studies of the properties of these solutions. Let us only mention that (1.2) is only a formal writing of (1.3) and that in general one does not have $-\Delta u + \mu u = f$ in the distributional sense in Ω, since in general the $C^\infty_0(\Omega)$ functions do not belong to $L^2_\mu(\Omega)$ (consider, for example, the case $\mu = \infty_S$ described above).

We now recall the notion of γ-convergence of measures, which is defined by means of the Γ-convergence of the corresponding functionals defined in $L^2(\Omega)$: for every $\mu \in M_0(\Omega)$, let J_μ be the functional defined by

$$J_\mu(v) = \begin{cases} ((v,v))_\mu = \displaystyle\int_\Omega |Dv|^2\,dx + \int_\Omega v^2\,d\mu & \text{if } v \in V_\mu(\Omega), \\ +\infty & \text{if } v \in L^2(\Omega) \setminus V_\mu(\Omega). \end{cases}$$

DEFINITION 1.1. *The sequence μ_ϵ in $M_0(\Omega)$ is said to* γ-converge *to a measure $\mu_0 \in M_0(\Omega)$ ($\mu_\epsilon \xrightarrow{\gamma} \mu_0$) if the corresponding functionals J_{μ_ϵ} Γ-converge to the functional J_{μ_0} in the space $L^2(\Omega)$; that is, if*

(i) *for every $v \in L^2(\Omega)$ and for every sequence v_ϵ converging to v in $L^2(\Omega)$ one has*

$$J_{\mu_0}(v) \le \liminf_\epsilon J_{\mu_\epsilon}(v_\epsilon),$$

(ii) *for every $v \in L^2(\Omega)$ there exists a sequence v_ϵ converging to v in $L^2(\Omega)$ such that*

$$J_{\mu_0}(v) = \lim_\epsilon J_{\mu_\epsilon}(v_\epsilon).$$

The previous convergence can be proved to be equivalent to the strong convergence in $L^2(\Omega)$ of the resolvent operators of problem (1.3) (see [DM2]): in other words, μ_ϵ γ-converges to μ_0 if and only if for every $f \in L^2(\Omega)$ the solutions u_ϵ of (1.3) for μ_ϵ converge strongly in $L^2(\Omega)$ (and weakly in $H^1_0(\Omega)$) to the solution u_0 of (1.3) for μ_0. In [DM2] it is shown that the set $M_0(\Omega)$ is closed and sequentially compact with respect to the γ-convergence and that the measures ∞_S defined above as well as the regular measures with bounded density with respect to the Lebesgue measure are dense in $M_0(\Omega)$. These results give motivation to the use of the class $M_0(\Omega)$. Relaxed Dirichlet problems such as (1.2) form in fact the smallest family of equations, stable under the $L^2(\Omega)$ convergence of the solutions, which contains Dirichlet problems for the Laplace operator on any subdomain A of Ω (which are formally equivalent, by definition, to relaxed problems of type (1.2) with measures $\mu = \infty_S$, $S = \Omega - A$). Moreover, as already remarked, this class of measures gives the structure of a Hilbert space to the space $V_\mu(\Omega)$.

Let us consider the space

$$V^1_\mu(\Omega) := \{v \in H^1(\Omega) \cap L^2_\mu(\Omega) : v - 1 \in H^1_0(\Omega)\}.$$

We are now in position to state the definition of correctors (see [T], [CM]).

DEFINITION 1.2. *If the sequence μ_ϵ in $M_0(\Omega)$ γ-converges to μ_0, a sequence w_ϵ in $V^1_{\mu_\epsilon}(\Omega)$ is said to be a sequence of* correctors *for problem* (1.2) *if, for any $f \in L^2(\Omega)$, defining u_ϵ and u_0 as the solutions of* (1.3) *for μ_ϵ and μ_0, one has, as ϵ tends to zero,*

$$u_\epsilon - w_\epsilon u_0 \to 0 \textit{ strongly in } W^{1,1}_0(\Omega). \tag{1.4}$$

In other words,

$$Du_\epsilon = D(w_\epsilon u_0) + r_\epsilon, \textit{ with } r_\epsilon \to 0 \textit{ strongly in } (L^1(\Omega))^N.$$

We recall the following definition (see [DM2]).

DEFINITION 1.3. *Assume that $V^1_\mu(\Omega)$ is not empty. The* μ-capacitary potential *z associated with a measure $\mu \in M_0(\Omega)$ is defined as the unique minimum point in $V^1_\mu(\Omega)$ of the functional $J_\mu(.)$.*

Note that Definition 1.3 has a meaning only if the space $V^1_\mu(\Omega)$ is not empty—a hypothesis that we will always suppose to be satisfied. This is, for example, the case if μ is assumed to be zero in a neighborhood of $\partial\Omega$.

By definition, then z solves the variational problem

$$z \in V^1_\mu(\Omega), \quad ((z, v))_\mu = 0 \text{ for all } v \in V_\mu(\Omega); \tag{1.5}$$

i.e., z is a solution of the relaxed problem that can formally be written as

$$\begin{cases} -\Delta z + \mu z = 0 & \text{in } \Omega, \\ z = 1 & \text{on } \partial\Omega. \end{cases}$$

Moreover, $z \in L^\infty(\Omega)$ since it can be easily proved that $0 \le z \le 1$ (in the sense of $H^1(\Omega)$).

Making use of Definition 1.3, [FT] suggests a method for constructing a sequence of correctors. Let us denote by $z_\epsilon \in V^1_{\mu_\epsilon}(\Omega)$ and $z_0 \in V^1_{\mu_0}(\Omega)$ the capacitary potentials associated with μ_ϵ and μ_0, respectively. Under the assumption (1.8) on the limit measure μ_0, the functions

$$w_\epsilon = \frac{z_\epsilon}{z_0} \tag{1.6}$$

define a sequence of correctors for problem (1.3), in accordance with Definition 1.2. More precisely one has (see [FT, Thm. 3.6]) the following theorem.

THEOREM 1.4. *Let Ω be sufficiently smooth, and assume that a sequence μ_ϵ is given in $M_0(\Omega)$ such that for every ϵ the spaces $V^1_{\mu_\epsilon}(\Omega)$ are not empty and*

$$\mu_\epsilon \xrightarrow{\gamma} \mu_0. \tag{1.7}$$

Assume also that

$$\mu_0 = m(x)dx, \qquad m \in L^\infty(\Omega),\ m \geq 0, \tag{1.8}$$

and that

$$f \in L^\infty(\Omega). \tag{1.9}$$

Then the functions w_ϵ *defined by* (1.6) *are in* $V^1_{\mu_\epsilon}(\Omega) \cap L^\infty(\Omega)$, *and for the solutions* u_ϵ *and* u_0 *of* (1.3) *corresponding to* μ_ϵ *and* μ_0 *one has the corrector result*

$$\|u_\epsilon - w_\epsilon u_0\|_{H^1_0(\Omega)} \to 0. \tag{1.10}$$

Remark 1.1. It is easily shown that $0 \leq z_\epsilon \leq 1$, while hypothesis (1.8) implies that $\alpha \leq z_0 \leq 1$ for some $\alpha > 0$ (see [FT, Prop. 3.2]). Thus w_ϵ is bounded in $L^\infty(\Omega)$. The sequence w_ϵ is also bounded in $H^1_0(\Omega)$ because $1/z_0$ can be proved to belong to $W^{1,\infty}(\Omega)$ under the hypothesis (1.8). Thus

$$w_\epsilon \rightharpoonup 1 \text{ weakly in } H^1(\Omega).$$

This property actually holds (see [DM2] or [FT, Rem. 3.1]) under the assumption that Ω belongs, with respect to μ_0, to a *rich* class of sets. This is the case here, since for a Radon measure it has been proved in [DM2] that this class contains all bounded Borel sets E with $\mu_0(\partial E) = 0$.

Theorem 1.1 extends, with a different definition of correctors, the results obtained by D. Cioranescu and F. Murat [CM] in an "abstract" setting that applies to the case of periodically perforated domains with holes of critical size.

Hypotheses (1.8) and (1.9) of Theorem 1.1 can be weakened (see [FT]). Moreover, we mention that another sequence of correctors has been recently proposed by A. Garroni and G. Dal Maso [GDM] (see also [Ca2], [CG], [DMM1], and [DMM2] for the nonlinear case) whose main advantage is that no hypothesis (on, for example, $V^1_{\mu_\epsilon}(\Omega)$, $\partial\Omega$) is required for it to work.

Let us come back to the quasi-linear problem (1.1). We shall consider in §2 nonlinear terms f with quadratic growth with respect to the gradient under two different sets of hypotheses.

(a) *Bounded case.* We assume $\lambda_0 > 0$ and that $|f(x,s,p)| \leq c_0 + b(|s|)|p|^2$ for some constant c_0 and function b; we are interested in bounded weak solutions and offer the following more precise definition.

DEFINITION 1.5. *A bounded weak solution of* (1.1) *is a function* u *such that*

$$\begin{cases} u \in V_\mu(\Omega) \cap L^\infty(\Omega), \\ ((u,v))_\mu + \lambda_0 \displaystyle\int_\Omega uv\,dx = \int_\Omega f(x,u,Du)v\,dx \quad \textit{for all } v \in V_\mu(\Omega) \cap L^\infty(\Omega). \end{cases} \tag{1.11}$$

(b) *Unbounded case.* We discuss the interesting case where $f(x,s,p) = h(x) - g(x,s,p)$, h is in $H^{-1}(\Omega)$, and g satisfies $|g(x,s,p)| \leq b(|s|)(1+|p|^2)$ as well as the sign condition $g(x,s,p)s \geq 0$. The term $\lambda_0 u$ can be included in g in this case, so we will only consider $\lambda_0 = 0$. Then (1.1) becomes

$$\begin{cases} -\Delta u + \mu u + g(x,u,Du) = h & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.12}$$

and we cannot expect a possible solution to be in $L^\infty(\Omega)$ (consider the case where $\mu = 0$ and $g = 0$; the solution is only in $H_0^1(\Omega)$ since h belongs to $H^{-1}(\Omega)$).

DEFINITION 1.6. *A weak solution of* (1.12) *is a function u such that*

$$(1.13)\quad \begin{cases} u \in V_\mu(\Omega), \quad g(x,u,Du) \in L^1(\Omega), \quad g(x,u,Du)u \in L^1(\Omega), \\ ((u,v))_\mu + \displaystyle\int_\Omega g(x,u,Du)v\,dx = \langle h,v\rangle_{H^{-1},H_0^1} \quad \textit{for all } v \in V_\mu(\Omega) \cap L^\infty(\Omega). \end{cases}$$

In the bounded case (a), by extending the techniques of [BMP1] to the relaxed case, we will prove in Theorem 2.1 the existence of a bounded weak solution of problem (1.11).

The proof will also imply that the bounded weak solutions of (1.11), in the sense of Definition 1.4, are uniformly bounded in $L^\infty(\Omega) \cap H_0^1(\Omega)$, independently of the measure μ. This property of solutions will be essential for us to prove, in §3, the homogenization result of Theorem 3.1. In this theorem we prove a stability property of solutions of (1.11) with respect to the γ-convergence of measures when the growth of f with respect to Du is assumed to be strictly subquadratic. An important role is played in the proof of this result by the use of the correctors (1.6) of the linear equation.

Note that this homogenization result is no more true in the case of an "exactly quadratic" (not strictly subquadratic) growth for f. Indeed, it has recently been shown by J. Casado-Diaz [Ca1] that in the exactly quadratic case the linear corrector result is no more valid. A new corrector has been constructed, and an extra nonlinear term appears in front of the measure μ_0 (see Remark 3.2 at the end of §3).

In the unbounded case (b), we shall prove in Theorem 2.2 of §2 the existence of weak solutions (without any L^∞ bound) for the relaxed quasi-linear problem (1.13). This result extends the results of [BMP2] and [BBM] to relaxed problems.

2. Existence results.

In this section we prove two existence results for the solutions of the relaxed quasi-linear elliptic problems (1.11) and (1.13).

2.1. The bounded case.

Let us assume that

$$(E_a1)\qquad \lambda_0 > 0, \qquad \mu \in M_0(\Omega),$$

$$(E_a2)\qquad \begin{cases} f : \Omega \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N \text{ is a Carathéodory function satisfying} \\ |f(x,s,p)| \le c_0 + b(|s|)|p|^2, \end{cases}$$

where $b(.)$ is an increasing function from $\mathbb{R}^+$ to $\mathbb{R}^+$ and $c_0 \in \mathbb{R}^+$.

THEOREM 2.1. *Under assumptions* (E_a1) *and* (E_a2), *there exists at least a bounded weak solution u of problem* (1.11).

Remark 2.1. Theorem 2.1 extends the result of [BMP1, Thm. 2.1] to the case of relaxed Dirichlet problems. The proof is henceforth very similar; nevertheless, for the sake of completeness we prefer to repeat its main steps, underlying the changes we have to make in the present case (particularly in Step 3).

Proof. We divide the proof into four steps:

Step 1: *Existence of approximate solutions.* We construct a sequence of problems that approximate (1.11) by introducing for $\epsilon > 0$ the bounded Carathéodory function

$$f_\epsilon(x,s,p) = \frac{f(x,s,p)}{1 + \epsilon|f(x,s,p)|};$$

note that $|f_\epsilon(x,s,p)| \leq \frac{1}{\epsilon}$ and $|f_\epsilon(x,s,p)| \leq |f(x,s,p)|$.

We shall first prove, for $\epsilon > 0$ fixed, the existence of a solution u_ϵ of the quasilinear problem

$$(1.11_\epsilon)\quad \begin{cases} \displaystyle\int_\Omega Du_\epsilon Dv\,dx + \lambda_0 \int_\Omega u_\epsilon v\,dx + \int_\Omega u_\epsilon v\,d\mu = \int_\Omega f_\epsilon(x,u_\epsilon,Du_\epsilon)v\,dx \quad \text{for all } v \in V_\mu(\Omega), \\ u_\epsilon \in V_\mu(\Omega) \cap L^\infty(\Omega), \end{cases}$$

which can be formally written as

$$\begin{cases} -\Delta u_\epsilon + \lambda_0 u_\epsilon + \mu u_\epsilon = f_\epsilon(x,u_\epsilon,Du_\epsilon) & \text{in } \Omega, \\ u_\epsilon = 0 & \text{on } \partial\Omega. \end{cases}$$

Since for ϵ fixed the application $w \to f_\epsilon(x,w,Dw)$ is bounded from $H_0^1(\Omega)$ into $L^\infty(\Omega)$, the mapping $S : V_\mu(\Omega) \to V_\mu(\Omega)$ that associates with each function $w \in V_\mu(\Omega)$ the unique solution $\overline{w} = Sw$ of the linear problem

$$\begin{cases} \displaystyle\int_\Omega D\overline{w}Dv\,dx + \lambda_0 \int_\Omega \overline{w}v\,dx + \int_\Omega \overline{w}v\,d\mu = \int_\Omega f_\epsilon(x,w,Dw)v\,dx \quad \text{for all } v \in V_\mu(\Omega), \\ \overline{w} \in V_\mu(\Omega), \end{cases}$$

which formally reads as

$$\begin{cases} -\Delta\overline{w} + \lambda_0\overline{w} + \mu\overline{w} = f_\epsilon(x,w,Dw) & \text{in } \Omega, \\ \overline{w} = 0 & \text{on } \partial\Omega, \end{cases}$$

satisfies the hypotheses of the Schauder fixed-point theorem. Thus (1.11_ϵ) has at least a solution that belongs to $V_\mu(\Omega)$. Moreover, this solution is in $L^\infty(\Omega)$, with

$$(2.1)\qquad -\frac{1}{\lambda_0\epsilon} \leq u_\epsilon \leq \frac{1}{\lambda_0\epsilon}.$$

To prove this, we use the function $z_\epsilon = (u_\epsilon - 1/(\lambda_0\epsilon))^+$ as the test function in problem (1.11_ϵ). This is possible since $z_\epsilon \in H_0^1(\Omega)$; moreover, $0 \leq z_\epsilon \leq u_\epsilon^+$, so z_ϵ belongs to $L^2_\mu(\Omega)$ and thus to $V_\mu(\Omega)$. Then, since $\mu \geq 0$ and $|f_\epsilon| \leq \frac{1}{\epsilon}$,

$$\begin{aligned} \int_\Omega |Dz_\epsilon|^2\,dx &\leq \int_\Omega (-\lambda_0 u_\epsilon + f_\epsilon(x,u_\epsilon,Du_\epsilon))z_\epsilon\,dx \\ &\leq \int_\Omega \left(-\lambda_0 u_\epsilon + \frac{1}{\epsilon}\right) z_\epsilon\,dx = \int_\Omega \left(-\lambda_0 u_\epsilon + \frac{1}{\epsilon}\right)\left(u_\epsilon - \frac{1}{\lambda_0\epsilon}\right)^+ dx \leq 0, \end{aligned}$$

so that $z_\epsilon = 0$. This proves one of the inequalities of (2.1). The other inequality is proved in the same way by considering the test function $z_\epsilon = -(u_\epsilon + 1/(\lambda_0\epsilon))^-$. Thus $u_\epsilon \in L^\infty(\Omega)$.

By this method we have also proved that $u_\epsilon \in L^\infty_\mu(\Omega)$. Indeed, from $z_\epsilon = 0$ in the sense of $H_0^1(\Omega)$, we deduce that $z_\epsilon = 0$ except on a set E of zero capacity; thanks to the definition of the space $M_0(\Omega)$, we have $\mu(E) = 0$ and we deduce $z_\epsilon = 0$ in $L^\infty_\mu(\Omega)$ too. In fact, this proof asserts that $H_0^1(\Omega) \cap L^\infty(\Omega) \subseteq H_0^1(\Omega) \cap L^\infty_\mu(\Omega)$.

Step 2: The solutions u_ϵ of problem (1.11_ϵ) are uniformly bounded in $V_\mu(\Omega) \cap L^\infty(\Omega)$. It is enough here to repeat the proofs of Steps 2 and 3 of Theorem 2.1 of

[BMP1], which is concerned with the case where $\mu = 0$. To prove the uniform bound in $L^\infty(\Omega)$, we remark that the functions $T_\epsilon(z_\epsilon^+)$, where

$$T_\epsilon(v) = v\ \exp(t_\epsilon v^2), \quad z_\epsilon = u_\epsilon - \frac{c_0}{\lambda_0}, \quad t_\epsilon = \frac{b(\|u_\epsilon\|_{L^\infty})^2}{2},$$

can be used as test functions in (1.11_ϵ) since $z_\epsilon^+ \in V_\mu(\Omega) \cap L^\infty(\Omega) \cap L^\infty_\mu(\Omega)$. Then the same computations as in [BMP1] yield, using the hypothesis (E_a1),

$$\|u_\epsilon\|_{L^\infty} \le \frac{c_0}{\lambda_0},$$

and the same estimate holds in $L^\infty_\mu(\Omega)$. Let

$$c_1 = b\left(\frac{c_0}{\lambda_0}\right).$$

To show the uniform estimate in $H_0^1(\Omega)$, we use as test function in (1.11_ϵ) the function $T(u_\epsilon) \in V_\mu(\Omega)$, where $T(v) = v\ \exp(tv^2)$ and $t = c_1^2/2$. This leads us to the inequality

$$\int_\Omega |Du_\epsilon|^2\,dx + \int_\Omega u_\epsilon^2\,d\mu \le K, \tag{2.2}$$

which means that u_ϵ is uniformly bounded in $V_\mu(\Omega)$.

Extracting a subsequence (still denoted by u_ϵ), we have proved the existence of a function $u \in V_\mu(\Omega) \cap L^\infty(\Omega)$ such that

$$u_\epsilon \rightharpoonup u \text{ weakly in } H_0^1(\Omega);$$

$$u_\epsilon \rightharpoonup u \text{ weakly in } L^2_\mu(\Omega);$$

$$u_\epsilon \to u \text{ strongly in } L^p(\Omega) \text{ for any } p < +\infty \text{ and weakly * in } L^\infty(\Omega);$$

$$u_\epsilon \to u \text{ a.e. in } \Omega.$$

Hence, we conclude in particular that

$$\|u\|_{L^\infty} \le \frac{c_0}{\lambda_0}.$$

Note that this $L^\infty(\Omega)$ bound as well as the $H_0^1(\Omega)$ bound

$$\|u\|_{H_0^1(\Omega)} \le K,$$

which is easily derived from (2.2), do not depend on the measure μ but only on c_0, λ_0, c_1, and Ω.

Step 3: *The sequence u_ϵ converges strongly in $H_0^1(\Omega)$ and $L^2_\mu(\Omega)$ to the function u.* This step is the more original one with respect to the proof of [BMP1] and uses an argument that was used in [Mo] in the context of nonlinear parabolic equations. Let ϵ and η be two positive parameters and u_ϵ and u_η be the corresponding solutions of (1.11_ϵ) and (1.11_η). Let $T(v) = v \exp(tv^2)$ and $t = 16c_1^2$. Subtracting (1.11_η) from (1.11_ϵ) and using the test function $T(u_\epsilon - u_\eta)$, which belongs to $V_\mu(\Omega)$, we obtain

$$\begin{aligned} \int_\Omega |Du_\epsilon - Du_\eta|^2 T'(u_\epsilon - u_\eta)\,dx &+ \lambda_0 \int_\Omega (u_\epsilon - u_\eta)T(u_\epsilon - u_\eta)\,dx \\ &+ \int_\Omega (u_\epsilon - u_\eta)T(u_\epsilon - u_\eta)\,d\mu \\ &= \int_\Omega [f_\epsilon(x, u_\epsilon, Du_\epsilon) - f_\eta(x, u_\eta, Du_\eta)]T(u_\epsilon - u_\eta)\,dx \\ &\le \int_\Omega [2c_0 + c_1|Du_\epsilon|^2 + c_1|Du_\eta|^2]|T(u_\epsilon - u_\eta)|\,dx \\ &\le \int_\Omega [2c_0 + 3c_1|Du_\epsilon|^2]|T(u_\epsilon - u_\eta)|\,dx \\ &+ \int_\Omega 2c_1|Du_\epsilon - Du_\eta|^2|T(u_\epsilon - u_\eta)|\,dx, \end{aligned} \tag{2.3}$$

where we have used the hypothesis (E_a2) on f and $|f_\epsilon| \le |f|$. Since the second and third integrals of the left-hand side are nonnegative, we get

$$\begin{aligned} \int_\Omega |Du_\epsilon - Du_\eta|^2 [T'(u_\epsilon - u_\eta) - 2c_1|T(u_\epsilon - u_\eta)|]\,dx \\ \le \int_\Omega [2c_0 + 3c_1|Du_\epsilon|^2]|T(u_\epsilon - u_\eta)|\,dx. \end{aligned}$$

The choice of t implies $T'(v) - 8c_1|T(v)| \ge \frac{1}{2}$. Now let η go to zero; by the results of Step 2 on the sequence u_η, the continuity of the functions T and T', and the weak lower semicontinuity of the left-hand side, we easily pass to the limit (as η tends to zero and ϵ is fixed) in the previous inequality, so that

$$\begin{aligned} \int_\Omega |Du_\epsilon - Du|^2 [T'(u_\epsilon - u) - 2c_1|T(u_\epsilon - u)|]\,dx \\ \le \int_\Omega [2c_0 + 3c_1|Du_\epsilon|^2]|T(u_\epsilon - u)|\,dx \\ \le \int_\Omega [2c_0 + 6c_1|Du|^2 + 6c_1|Du_\epsilon - Du|^2]|T(u_\epsilon - u)|\,dx. \end{aligned}$$

Then for t fixed as before,

$$\begin{aligned} \frac{1}{2}\int_\Omega |Du_\epsilon - Du|^2\,dx &\le \int_\Omega |Du_\epsilon - Du|^2[T'(u_\epsilon - u) - 8c_1|T(u_\epsilon - u)|]\,dx \\ &\le \int_\Omega [2c_0 + 6c_1|Du|^2]|T(u_\epsilon - u)|\,dx. \end{aligned}$$

Since the last integral tends to zero as ϵ tends to zero, we have proved that u_ϵ tends to u strongly in $H_0^1(\Omega)$. This convergence is strong in $L^2_\mu(\Omega)$ too, since, coming back to the first inequality in (2.3), we get

$$\begin{aligned} \int_\Omega |u_\epsilon - u_\eta|^2\,d\mu &\le \int_\Omega (u_\epsilon - u_\eta)T(u_\epsilon - u_\eta)\,d\mu \\ &\le \int_\Omega [2c_0 + c_1|Du_\epsilon|^2 + c_1|Du_\eta|^2]|T(u_\epsilon - u_\eta)\,dx, \end{aligned}$$

which is easily shown to converge to zero.

Step 4: *Passing to the limit in* (1.11_ϵ) *and proving that u is a solution of problem* (1.11). By Step 3 we know that (up to the extraction of a subsequence) $Du_\epsilon \to Du$ a.e. in Ω. Then $f_\epsilon(u_\epsilon, Du_\epsilon) \to f(u, Du)$ a.e. in Ω. Since $|f_\epsilon(u_\epsilon, Du_\epsilon)| \le c_0 + c_1|Du_\epsilon|^2$, which converges strongly in $L^1(\Omega)$, Vitali's theorem ensures that $f_\epsilon(u_\epsilon, Du_\epsilon) \to f(u, Du)$ strongly in $L^1(\Omega)$. This shows that u solves problem (1.11), as we wanted to prove. □

Remark 2.2. As already observed in [BMP1, Rem. 3.3], the strict positivity of λ_0 (first part of assumption (E_a1)) is essential in the proof of Theorem 1.1 because it allows us to obtain the $L^\infty(\Omega)$ bound on u_ϵ. Since the term containing the measure could degenerate somewhere in Ω (either with $\mu = 0$ or with $\mu = +\infty$), the existence of a solution is no more guaranteed in the absence of the zero-order term in the operator. As a counterexample, one could consider the counterexample introduced by J. L. Kazdan and R. J. Kramer in [KK] (see also [BMP1, Contre-ex. 3.1]) with an extra term μu, with either $\mu \equiv 0$ or $\mu = \infty_E$, E being a closed subset of Ω.

Of course, hypothesis (E_a1) can be replaced by

$$\lambda_0 \ge 0, \quad \mu \in M_0(\Omega), \quad \mu + \lambda_0\, dx \ge a_0 dx$$

for a strictly positive constant a_0.

2.2. The unbounded case. Let us now consider the problem (1.13) and assume that

$$(E_b1) \qquad \mu \in M_0(\Omega), \quad h = -\operatorname{div} r \in H^{-1}(\Omega), \quad r \in (L^2(\Omega))^N,$$

$$(E_b2) \qquad \begin{cases} g : \Omega \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N \text{ is a Carathéodory function satisfying} \\ |g(x,s,p)| \le b(|s|)(1+|p|^2) \\ \text{and} \\ g(x,s,p)s \ge 0, \end{cases}$$

where $b(.)$ is a continuous and increasing function from $\mathbb{R}^+$ to $\mathbb{R}^+$.

As already remarked, the main feature of this case is that, due to the lack of regularity of h, a solution of (1.13) is no more in $L^\infty(\Omega)$. This property, which was crucial in the proof of Theorem 2.1, now must be replaced by the use of truncations (see also [BBM], [BGM], [LM], and [M], where the same idea is used).

THEOREM 2.2. *Under assumptions* (E_b1) *and* (E_b2), *there exists at least a weak solution u of problem* (1.13).

Proof. We divided the proof into four steps.

Step 1: *Existence of approximate solutions.* We construct a sequence of problems that approximates (1.13) by introducing for $\epsilon > 0$ the Carathéodory function:

$$g_\epsilon(x,s,p) = \frac{g(x,s,p)}{1+\epsilon|g(x,s,p)|}.$$

By this definition g_ϵ has the same properties as g (growth and sign condition) but is bounded by $\frac{1}{\epsilon}$. We can then repeat the argument of Step 1 of the proof of Theorem 2.1, which shows the existence of a weak solution u_ϵ (which here does not belong to

$L^\infty(\Omega)$ since h only belongs to $H^{-1}(\Omega)$) of the quasi-linear problem

$$(1.13_\epsilon) \qquad \begin{cases} \displaystyle\int_\Omega Du_\epsilon Dv\,dx + \int_\Omega u_\epsilon v\,d\mu + \int_\Omega g_\epsilon(x,u_\epsilon,Du_\epsilon)v\,dx = \langle h,v\rangle_{H^{-1},H_0^1}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for all } v \in V_\mu(\Omega), \\ u_\epsilon \in V_\mu(\Omega). \end{cases}$$

Step 2: *The solutions u_ϵ of problem* (1.13_ϵ) *are uniformly bounded in* $V_\mu(\Omega)$. It is enough to use u_ϵ as the test function in problem (1.13_ϵ) and to use the sign condition on g to get a uniform bound in $V_\mu(\Omega)$:

$$\int_\Omega |Du_\epsilon|^2\,dx + \int_\Omega u_\epsilon^2\,d\mu \le K.$$

Extracting a subsequence (still denoted by u_ϵ), we have

$$\begin{aligned} u_\epsilon &\rightharpoonup u \quad \text{weakly in } H_0^1(\Omega), \\ u_\epsilon &\rightharpoonup u \quad \text{weakly in } L^2_\mu(\Omega), \\ u_\epsilon &\to u \quad \text{a.e. in } \Omega. \end{aligned}$$

The above estimate also proves that

$$(2.4) \qquad \int_\Omega g_\epsilon(x,u_\epsilon,Du_\epsilon)u_\epsilon\,dx \le K.$$

Step 3: *Strong convergence of* $T_k(u_\epsilon)$. For a positive constant k, let $T_k(w)$ denote the truncation of the function w at level k, that is,

$$T_k(w) = \begin{cases} k & \text{if } w > k, \\ w & \text{if } |w| \le k, \\ -k & \text{if } w < -k, \end{cases}$$

and let

$$G_k(w) = w - T_k(w).$$

We remark that if $w \in V_\mu(\Omega)$, then both $T_k(w)$ and $G_k(w)$ are in $V_\mu(\Omega)$. We will prove the following convergence results:

$$(2.5) \qquad \lim_{k\to+\infty} \limsup_{\epsilon\to 0} \|G_k(u_\epsilon)\|_{H_0^1(\Omega)} = 0,$$

$$(2.6) \qquad T_k^\pm(u_\epsilon) \to T_k^\pm(u) \text{ strongly in } H_0^1(\Omega) \text{ and in } L^2_\mu(\Omega),$$

where w^+ (w^-) denotes the positive (negative) part of a function w.

Take $v = G_k(u_\epsilon)$ as a test function in (1.13_ϵ); then

$$\int_\Omega Du_\epsilon DG_k(u_\epsilon)\,dx + \int_\Omega g_\epsilon(x,u_\epsilon,Du_\epsilon)G_k(u_\epsilon)\,dx \\ + \int_\Omega u_\epsilon G_k(u_\epsilon)\,d\mu = \langle h, G_k(u_\epsilon)\rangle_{H^{-1},H_0^1}.$$

Since $G_k(u_\epsilon)$ has the sign of u_ϵ, the second and third terms of the left-hand side are nonnegative, so we obtain

$$\begin{aligned}\int_\Omega |DG_k(u_\epsilon)|^2\,dx \le \langle h, G_k(u_\epsilon)\rangle_{H^{-1},H_0^1} &= \int_{|u_\epsilon|\ge k} rDG_k(u_\epsilon)\,dx \\ &\le \frac{1}{2}\int_{|u_\epsilon|\ge k} |r|^2\,dx + \frac{1}{2}\int_\Omega |DG_k(u_\epsilon)|^2\,dx.\end{aligned}$$

Thus

$$\limsup_{\epsilon\to 0} \|DG_k(u_\epsilon)\|^2_{L^2(\Omega)} \le \int_{|u|\ge k} |r|^2\,dx,$$

and (2.5) follows when $k \to +\infty$.

The proof of (2.6) is more delicate. We will write it only for the positive parts (the other part of the proof is analogous). We adopt the approach already used in Step 3 of the proof of Theorem 2.1. For two positive parameters ϵ and η, let u_ϵ and u_η be the corresponding solutions of (1.13_ϵ) and (1.13_η). By subtracting these two equations we get, for any $v \in V_\mu(\Omega)$,

$$\int_\Omega D(u_\epsilon - u_\eta)Dv\,dx + \int_\Omega (u_\epsilon - u_\eta)v\,d\mu + \int_\Omega (g_\epsilon(x,u_\epsilon,Du_\epsilon) - g_\eta(x,u_\eta,Du_\eta))v\,dx = 0.$$

Let us take in this weak formulation the test function $T(T_k^+(u_\epsilon) - T_k^+(u_\eta)) \in V_\mu(\Omega)$, where as usual $T(v) = v\exp(tv^2)$, with a positive constant t to be chosen later. Since

$$Dv = DT_k^+(v) + DG_k^+(v) - Dv^-,$$

we obtain

$$\begin{aligned}&\int_\Omega |D(T_k^+(u_\epsilon) - T_k^+(u_\eta))|^2 T'(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,dx \\ &\quad + \int_\Omega (u_\epsilon - u_\eta)T(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,d\mu \\ &\quad = -\int_\Omega [g_\epsilon(x,u_\epsilon,Du_\epsilon) - g_\eta(x,u_\eta,Du_\eta)]T(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,dx \\ &\qquad - \int_\Omega D(G_k^+(u_\epsilon) - G_k^+(u_\eta))D(T_k^+(u_\epsilon) - T_k^+(u_\eta))T'(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,dx \\ &\qquad + \int_\Omega D(u_\epsilon^- - u_\eta^-)D(T_k^+(u_\epsilon) - T_k^+(u_\eta))T'(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,dx.\end{aligned}\tag{2.7}$$

Let us consider separately each term of (2.7). First, it is easy to check that the second term of the left-hand side of (2.7) is nonnegative; more precisely, an analysis of all possible cases (namely $u_\epsilon \le 0$, $u_\epsilon > 0$, $u_\eta \le 0$, $u_\eta > 0$) shows that

$$\int_\Omega (u_\epsilon - u_\eta)T(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,d\mu \ge \int_\Omega [T_k^+(u_\epsilon) - T_k^+(u_\eta)]^2\,d\mu.\tag{2.8}$$

Let us now define

$$\begin{aligned}L_{k,\epsilon} &= \{x \in \Omega : 0 < u_\epsilon(x) < k\}, \\ U_{k,\epsilon} &= \{x \in \Omega : u_\epsilon(x) \ge k\}.\end{aligned}$$

Concerning the first term in the right-hand side of (2.7), using a detailed analysis of all possible cases (namely $u_\epsilon \leq 0$, $0 < u_\epsilon \leq k$, $u_\epsilon \geq k$, $u_\eta \leq 0$, $0 < u_\eta \leq k$, $u_\eta \geq k$), then the sign condition on g_ϵ and g_η, and finally the growth condition, one shows that

$$\begin{aligned}
&-\int_\Omega [g_\epsilon(x,u_\epsilon,Du_\epsilon) - g_\eta(x,u_\eta,Du_\eta)]T(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,dx\\
&\quad\leq -\int_{L_{k,\epsilon}\cap L_{k,\eta}} [g_\epsilon(x,u_\epsilon,Du_\epsilon) - g_\eta(x,u_\eta,Du_\eta)]T(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,dx\\
&\qquad -\int_{L_{k,\epsilon}\cap U_{k,\eta}} g_\epsilon(x,u_\epsilon,Du_\epsilon)T(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,dx\\
&\qquad +\int_{U_{k,\epsilon}\cap L_{k,\eta}} g_\eta(x,u_\eta,Du_\eta)T(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,dx \qquad (2.9)\\
&\quad\leq b(k)\int_\Omega [4 + 2|DT_k^+(u_\epsilon)|^2 + 2|DT_k^+(u_\eta)|^2]|T(T_k^+(u_\epsilon) - T_k^+(u_\eta))|\,dx\\
&\quad\leq b(k)\int_\Omega [4 + 6|DT_k^+(u_\epsilon)|^2]|T(T_k^+(u_\epsilon) - T_k^+(u_\eta))|\,dx\\
&\qquad + b(k)\int_\Omega 4|D(T_k^+(u_\epsilon) - T_k^+(u_\eta))|^2|T(T_k^+(u_\epsilon) - T_k^+(u_\eta))|\,dx.
\end{aligned}$$

Finally, the last two terms of (2.7) can be written as

$$\begin{aligned}
&-\int_\Omega D(G_k^+(u_\epsilon) - G_k^+(u_\eta))D(T_k^+(u_\epsilon) - T_k^+(u_\eta))T'(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,dx\\
&\quad = \int_\Omega [DG_k^+(u_\epsilon)DT_k^+(u_\eta) + DG_k^+(u_\eta)DT_k^+(u_\epsilon)]T'(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,dx, \qquad (2.10)
\end{aligned}$$

$$\begin{aligned}
&\int_\Omega D(u_\epsilon^- - u_\eta^-)D(T_k^+(u_\epsilon) - T_k^+(u_\eta))T'(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,dx\\
&\quad = -\int_\Omega [Du_\epsilon^- DT_k^+(u_\eta) + Du_\eta^- DT_k^+(u_\eta)]T'(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,dx. \qquad (2.11)
\end{aligned}$$

Substituting the results (2.8)–(2.11) into (2.7), we have

$$\begin{aligned}
&\int_\Omega |D(T_k^+(u_\epsilon) - T_k^+(u_\eta))|^2[T'(T_k^+(u_\epsilon) - T_k^+(u_\eta)) - 4b(k)|T(T_k^+(u_\epsilon) - T_k^+(u_\eta))|]\,dx\\
&\quad + \int_\Omega [T_k^+(u_\epsilon) - T_k^+(u_\eta)]^2 d\mu\\
&\qquad \leq b(k)\int_\Omega [4 + 6|DT_k^+(u_\epsilon)|^2]|T(T_k^+(u_\epsilon) - T_k^+(u_\eta))|\,dx\\
&\qquad\quad + \int_\Omega [DG_k^+(u_\epsilon)DT_k^+(u_\eta) + DG_k^+(u_\eta)DT_k^+(u_\epsilon)]\,T'(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,dx\\
&\qquad\quad - \int_\Omega [Du_\epsilon^- DT_k^+(u_\eta) + Du_\eta^- DT_k^+(u_\epsilon)]T'(T_k^+(u_\epsilon) - T_k^+(u_\eta))\,dx.
\end{aligned}$$

Now fix $t = 64(b(k))^2$, which implies $T'(v) - 16b(k)|T(v)| \geq \frac{1}{2}$. For a fixed ϵ, the fact that the sequence u_η weakly converges in $H_0^1(\Omega)$ to u allows us to pass to the limit in the previous inequality as η goes to zero (we use in particular the weak lower semicontinuity in $H_0^1(\Omega)$ of the left-hand side), so

$$\begin{aligned}\int_\Omega &|D(T_k^+(u_\epsilon) - T_k^+(u))|^2[T'(T_k^+(u_\epsilon) - T_k^+(u)) - 16b(k)|T(T_k^+(u_\epsilon) - T_k^+(u))|]\,dx\\ &+ \int_\Omega [T_k^+(u_\epsilon) - T_k^+(u)]^2 d\mu\\ &\quad\le b(k)\int_\Omega [4 + 12|DT_k^+(u)|^2]|T(T_k^+(u_\epsilon) - T_k^+(u))|\,dx\\ &\qquad + \int_\Omega [DG_k^+(u_\epsilon)DT_k^+(u) + DG_k^+(u)DT_k^+(u_\epsilon)]T'(T_k^+(u_\epsilon) - T_k^+(u))\,dx\\ &\qquad - \int_\Omega [Du_\epsilon^- DT_k^+(u) + Du^- DT_k^+(u_\epsilon)]T'(T_k^+(u_\epsilon) - T_k^+(u))\,dx.\end{aligned}$$

Now let ϵ tend to zero. The strong convergence of $T_k^+(u_\epsilon)$ to $T_k^+(u)$ in $H_0^1(\Omega) \cap L_\mu^2(\Omega)$ (i.e., (2.6)) then follows since the three integrals in the right-hand side converge to zero.

Step 4: *The sequence u_ϵ converges strongly in $H_0^1(\Omega)$ and in $L_\mu^2(\Omega)$ and the end of the proof.* Since, by definition, $u_\epsilon^+ - u^+ = (T_k^+(u_\epsilon) - T_k^+(u)) + (G_k^+(u_\epsilon) - G_k^+(u))$, we have for the $H_0^1(\Omega)$ or $L_\mu^2(\Omega)$ norm

$$\|u_\epsilon^+ - u^+\| \le \|T_k^+(u_\epsilon) - T_k^+(u)\| + \|G_k^+(u_\epsilon)\| + \|G_k^+(u)\|.$$

Using (2.5) and (2.6), this proves the strong convergence of u_ϵ^+ to u^+. The analogous result for u_ϵ^- proves that

$$u_\epsilon \to u \text{ strongly in } H_0^1(\Omega) \text{ and in } L_\mu^2(\Omega). \tag{2.12}$$

Extracting a subsequence such that $Du_\epsilon \to Du$ a.e. in Ω, we deduce from (2.4) and Fatou's lemma that

$$g(x, u, Du)u \in L^1(\Omega).$$

Writing

$$\int_\Omega |g(x, u, Du)|\,dx \le \int_{|u|\le 1} b(1)(1 + |Du|^2)\,dx + \int_{|u|\ge 1} g(x, u, Du)u\,dx$$

implies that $g(x, u, Du) \in L^1(\Omega)$.

Finally, in view of (2.4) we have

$$\int_{|u_\epsilon|>m} |g_\epsilon(x, u_\epsilon, Du_\epsilon)|\,dx \le \frac{1}{m}\int_{|u_\epsilon|>m} g_\epsilon(x, u_\epsilon, Du_\epsilon)u_\epsilon\,dx \le \frac{K}{m}. \tag{2.13}$$

On the other hand, for m fixed, Vitali's theorem, the estimate

$$|\chi_{|u_\epsilon|\le m} g_\epsilon(x, u_\epsilon, Du_\epsilon)| \le b(m)(1 + |Du_\epsilon|^2),$$

and (2.12) imply that

$$\chi_{|u_\epsilon|\le m} g_\epsilon(x, u_\epsilon, Du_\epsilon) \to \chi_{|u|\le m} g(x, u, Du) \text{ strongly in } L^1(\Omega). \tag{2.14}$$

Combining (2.13), (2.14), and

$$\int_{|u|\geq m} |g(x,u,Du)|\,dx \leq \frac{1}{m}\int_\Omega g(x,u,Du)u\,dx,$$

we obtain

$$g_\epsilon(x,u_\epsilon,Du_\epsilon) \to g(x,u,Du) \text{ strongly in } L^1(\Omega),$$

which allows us to pass to the limit in (1.13_ϵ) and to obtain (1.13).

The proof of the Theorem 2.2 is complete. □

3. Homogenization. In this section we study the convergence of the solution of (1.1) when the measure μ varies. We shall consider only nonlinear terms with subquadratic growth with respect to the gradient and prove the stability (Theorem 3.1). The general quadratic case is, however, very different, as it has recently been shown in Casado-Diaz [Ca1] (see Remark 3.2).

Let us consider the sequence of problems

$$(3.1_\epsilon)\qquad \begin{cases} \displaystyle\int_\Omega Du_\epsilon Dv\,dx + \lambda_0\int_\Omega u_\epsilon v\,dx + \int_\Omega u_\epsilon v\,d\mu_\epsilon = \int_\Omega f(x,Du_\epsilon)v\,dx \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{for all } v \in V_{\mu_\epsilon}(\Omega)\cap L^\infty(\Omega), \\ u_\epsilon \in V_{\mu_\epsilon}(\Omega)\cap L^\infty(\Omega), \end{cases}$$

where Ω is a bounded domain in $\mathbb{R}^N$ with smooth boundary, $\mu_\epsilon \in M_0(\Omega)$, and for the sake of simplicity, we consider only nonlinear terms that do not depend explicitly on u_ϵ. The general case is evoked in Remark 3.3. We assume the following hypotheses:

(H1)

$$\begin{cases} |f(x,p_1) - f(x,p_2)| \leq K(1+|p_1|^{s-\gamma}+|p_2|^{s-\gamma})|p_1-p_2|^\gamma, \\ \qquad\qquad \text{for any } p_1,p_2 \in \mathbb{R}^N, \text{ with } 0<\gamma\leq 1,\ \gamma\leq s<2; \\ |f(x,0)| \leq c_0, \end{cases}$$

$$(\text{H2})\qquad \begin{cases} \lambda_0 > 0, \qquad \mu_\epsilon \xrightarrow{\gamma} \mu_0, \qquad V^1_{\mu_\epsilon}(\Omega) \neq \emptyset; \\ \mu_0 = m(x)dx, \quad \text{where } m \in L^\infty(\Omega),\ m \geq 0. \end{cases}$$

Note that (H1) implies in particular that f has a strictly subquadratic growth in the gradient since

$$(3.2)\qquad |f(x,p)| \leq c_1(1+|p|^s), \quad \text{for any } p \in \mathbb{R}^N\ (s<2),$$

so hypothesis (E_a2) is satisfied. Therefore, the existence result of Theorem 2.1 holds in this case.

On the other hand, hypothesis (H2) allows us to use the corrector result of Theorem 1.1.

We also make an additional hypothesis on the correctors w_ϵ defined by (1.6). We assume that

$$(\text{H3})\qquad Dw_\epsilon \to 0 \text{ a.e. in } \Omega.$$

Hypothesis (H3) was recently proved to hold true by Casado-Diaz [Ca1] and Dal Maso and Murat [DMM] for the correctors considered by these authors. We will prove in Remark 3.1 that (H3) holds in this case of periodically perforated domains with holes of critical size. This property was also observed by Boccardo and Donato [BD] and Labani and Picard [LP].

We shall prove the following theorem.

THEOREM 3.1. *Assume* (H1), (H2), *and* (H3), *and let* $u_\epsilon \in V_{\mu_\epsilon}(\Omega) \cap L^\infty(\Omega)$ *be any sequence of solutions of* (3.1_ϵ). *Up to the extraction of a subsequence we have*

$$u_\epsilon \rightharpoonup u_0 \textit{ weakly in } H^1(\Omega),$$

where u_0 *is a solution of*

$$(3.1_0) \qquad \begin{cases} \displaystyle\int_\Omega Du_0 Dv\,dx + \lambda_0 \int_\Omega u_0 v\,dx + \int_\Omega u_0 v\,d\mu_0 = \int_\Omega f(x, Du_0) v\,dx \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad \textit{for all } v \in V_{\mu_0}(\Omega) \cap L^\infty(\Omega), \\ u_0 \in V_{\mu_0}(\Omega) \cap L^\infty(\Omega), \end{cases}$$

which reads formally as

$$\begin{cases} -\Delta u_0 + \lambda_0 u_0 + \mu_0 u_0 = f(x, Du_0) & \textit{in } \Omega, \\ u_0 = 0 & \textit{on } \partial\Omega. \end{cases}$$

In other words, Theorem 3.1 asserts that by passing to the limit on μ_ϵ both the operator and the right-hand side remain unchanged. This is a stability result.

Proof. We distinguish four steps in the proof.

Step 1: *Bounds for the solutions* u_ϵ *and* $f(x, Du_\epsilon) = g_\epsilon$. In view of (3.2), we know from Theorem 2.1 the existence of at least one solution of (3.1_ϵ). It also follows from the proof of Theorem 2.1 (see the end of Step 2 of that proof) that any solution of this problem is bounded in $H_0^1(\Omega)$ and $L^\infty(\Omega)$ by a constant that depends only on c_0, λ_0, and Ω, i.e.,

$$(3.3) \qquad \|u_\epsilon\|_{H_0^1} + \|u_\epsilon\|_{L^2_{\mu_\epsilon}} + \|u_\epsilon\|_{L^\infty} \le K.$$

We can thus extract a subsequence (still denoted by u_ϵ) such that

$$u_\epsilon \rightharpoonup u \text{ weakly in } H_0^1(\Omega),$$
$$u_\epsilon \to u \text{ strongly in } L^p(\Omega), \text{ for any } p < \infty, \text{ weakly * in } L^\infty(\Omega) \text{ and a.e. in } \Omega.$$

Let $g_\epsilon = f(x, Du_\epsilon)$. Take $q = \frac{2}{s} > 1$. Using Hölder's inequality, we get

$$\int_\Omega |g_\epsilon|^q\,dx = \int_\Omega |f(x, Du_\epsilon)|^q\,dx \le c_1^q \int_\Omega (1 + |Du_\epsilon|^s)^q\,dx \le \text{Cst}.$$

Extracting a new subsequence, there exists a function g_0 such that

$$f(x, Du_\epsilon) = g_\epsilon \rightharpoonup g_0 \text{ weakly in } L^q(\Omega).$$

Step 2: *A first passage to the limit in* (3.1_ϵ). In this step and in Step 3, we use the corrector result. Note that hypothesis (H2) is used only here and in Step 3. Note also that we could have $\lambda_0 = 0$ now and in what follows. The hypothesis $\lambda_0 > 0$ is indeed essential only in the proofs of the existence of u_ϵ and of the uniform estimates on u_ϵ (Step 1).

Now consider the sequence of functions w_ϵ defined by (1.6) in §1. Since, as observed in Remark 1.1, w_ϵ is bounded in L^∞, we have

$$w_\epsilon \to 1 \text{ strongly in } L^p(\Omega) \text{ for any } p < +\infty, \text{ and weakly * in } L^\infty(\Omega).$$

We claim

$$\text{(3.4)} \qquad \begin{cases} v_\epsilon \in V_{\mu_\epsilon}(\Omega), \qquad v_0 \in V_{\mu_0}(\Omega), \qquad v_\epsilon \rightharpoonup v_0 \text{ weakly in } H_0^1(\Omega) \\ \Rightarrow \displaystyle\int_\Omega Dw_\epsilon Dv_\epsilon\, dx + \int_\Omega w_\epsilon v_\epsilon\, d\mu_\epsilon \to \int_\Omega v_0\, d\mu_0. \end{cases}$$

Indeed, recalling the definitions of $w_\epsilon = z_\epsilon/z_0$ and z_ϵ, which solves (1.5) for the measure μ_ϵ, and the fact that z_0 and $1/z_0$ belong to $W^{1,\infty}(\Omega)$ (this is a consequence of hypothesis (H2), see [FT, Prop. 3.2]), and finally the fact that z_0 solves (1.5) for the measure μ_0, we have

$$\begin{aligned} &\int_\Omega Dw_\epsilon Dv_\epsilon\, dx + \int_\Omega w_\epsilon v_\epsilon\, d\mu_\epsilon \\ &= \left(\left(z_\epsilon, \frac{v_\epsilon}{z_0}\right)\right)_{\mu_\epsilon} + \int_\Omega D\left(\frac{1}{z_0}\right) Dv_\epsilon z_\epsilon\, dx - \int_\Omega Dz_\epsilon D\left(\frac{1}{z_0}\right) v_\epsilon\, dx \\ &\to \int_\Omega D\left(\frac{1}{z_0}\right) Dv_0 z_0\, dx - \int_\Omega Dz_0 D\left(\frac{1}{z_0}\right) v_0\, dx \\ &= -\int_\Omega Dz_0 D\left(\frac{v_0}{z_0}\right) dx = \int_\Omega v_0\, d\mu_0. \end{aligned}$$

For $\phi \in V_{\mu_0}(\Omega) \cap L^\infty(\Omega)$, take ϕw_ϵ, which belongs to $V_{\mu_\epsilon}(\Omega)$, as a test function in (3.1_ϵ). This yields

$$\text{(3.5)} \qquad \begin{aligned} &\int_\Omega Du_\epsilon D(\phi w_\epsilon)\, dx + \lambda_0 \int_\Omega u_\epsilon \phi w_\epsilon\, dx + \int_\Omega u_\epsilon \phi w_\epsilon\, d\mu_\epsilon \\ &\quad = \int_\Omega f(x, Du_\epsilon)\phi w_\epsilon\, dx = \int_\Omega g_\epsilon \phi w_\epsilon\, dx. \end{aligned}$$

Now w_ϵ tends to 1 strongly in $L^p(\Omega)$ for any $p < \infty$, while g_ϵ tends to g_0 weakly in $L^q(\Omega)$ with $q = \frac{2}{s} > 1$; the right-hand side of (3.5) then converges to $\int_\Omega g_0 \phi\, dx$.

We rewrite the left-hand side of (3.5) and pass to the limit using in particular (3.4) with $v_\epsilon = \phi u_\epsilon$. This yields

$$\begin{aligned} &\int_\Omega Du_\epsilon D(\phi w_\epsilon)\, dx + \lambda_0 \int_\Omega u_\epsilon \phi w_\epsilon\, dx + \int_\Omega u_\epsilon \phi w_\epsilon\, d\mu_\epsilon \\ &= \int_\Omega Dw_\epsilon D(\phi u_\epsilon)\, dx + \int_\Omega u_\epsilon \phi w_\epsilon\, d\mu_\epsilon + \lambda_0 \int_\Omega u_\epsilon \phi w_\epsilon\, dx \\ &\quad - \int_\Omega Dw_\epsilon D\phi u_\epsilon\, dx + \int_\Omega Du_\epsilon D\phi w_\epsilon\, dx \\ &\to \int_\Omega u_0 \phi\, d\mu_0 + \lambda_0 \int_\Omega u_0 \phi\, dx + \int_\Omega Du_0 D\phi\, dx. \end{aligned}$$

From Definition 1.1 of the γ-convergence and the bound (3.3), we finally deduce that u_0 belongs to V_{μ_0}. Hence u_0 satisfies

$$\begin{cases} u_0 \in V_{\mu_0}(\Omega) \cap L^\infty(\Omega), \\ \displaystyle\int_\Omega Du_0 D\phi\, dx + \lambda_0 \int_\Omega u_0 \phi\, dx + \int_\Omega u_0 \phi\, d\mu_0 = \int_\Omega g_0 \phi\, dx \\ \qquad\qquad\qquad\qquad \text{for all } \phi \in V_{\mu_0}(\Omega) \cap L^\infty(\Omega). \end{cases}$$

Step 3: *Corrector result for the nonlinear problem.* Let us first prove the convergence of the energies. Taking $v = u_\epsilon$ as a test function in (3.1_ϵ), we pass easily to the limit since u_ϵ converges strongly to u in $L^p(\Omega)$ for any $p < \infty$. We obtain

$$(3.6)\qquad \begin{aligned} &\int_\Omega |Du_\epsilon|^2\, dx + \int_\Omega u_\epsilon^2\, d\mu_\epsilon = \int_\Omega u_\epsilon g_\epsilon\, dx - \lambda_0 \int_\Omega u_\epsilon^2\, dx \\ &\qquad \to \int_\Omega u_0 g_0\, dx - \lambda_0 \int_\Omega u_0^2\, dx = \int_\Omega |Du_0|^2\, dx + \int_\Omega u_0^2\, d\mu_0. \end{aligned}$$

We now claim that for any $\phi \in V_{\mu_0}(\Omega) \cap L^\infty(\Omega)$

$$(3.7)\qquad \begin{aligned} &\int_\Omega |D(u_\epsilon - w_\epsilon \phi)|^2\, dx + \int_\Omega (u_\epsilon - w_\epsilon \phi)^2\, d\mu_\epsilon \\ &\qquad \to \int_\Omega |D(u_0 - \phi)|^2\, dx + \int_\Omega (u_0 - \phi)^2\, d\mu_0. \end{aligned}$$

To prove this claim, we write the left-hand side of (3.7) as

$$\begin{aligned} &\left[\int_\Omega |Du_\epsilon|^2\, dx + \int_\Omega u_\epsilon^2 d\mu_\epsilon\right] + \left[\int_\Omega Dw_\epsilon D(\phi^2 w_\epsilon)\, dx + \int_\Omega (\phi^2 w_\epsilon) w_\epsilon\, d\mu_\epsilon\right] \\ &\quad - 2\left[\int_\Omega Dw_\epsilon D(\phi u_\epsilon)\, dx + \int_\Omega (\phi u_\epsilon) w_\epsilon\, d\mu_\epsilon\right] \\ &\quad + \int_\Omega w_\epsilon^2 |D\phi|^2\, dx - 2\int_\Omega w_\epsilon Du_\epsilon D\phi\, dx + 2\int_\Omega u_\epsilon Dw_\epsilon D\phi\, dx, \end{aligned}$$

where it is easy to pass to the limit by using (3.4) with $v_\epsilon = \phi^2 w_\epsilon$ and (3.6) with $v_\epsilon = \phi u_\epsilon$. This proves (3.7).

Taking now $\phi = u_0$ in (3.7) we obtain

$$(3.8)\qquad u_\epsilon - w_\epsilon u_0 \to 0 \text{ strongly in } H_0^1(\Omega).$$

The previous proof is similar to the proof of Theorem 3.4 of [CM] and Theorem 3.6 of [FT]. Note, however, that we assumed here ϕ (and u_0) belongs to $V_{\mu_0}(\Omega) \cap L^\infty(\Omega)$, which allows us to obtain (3.8), a result that is stronger than the result of [CM].

Step 4: *Identifying* g_0 *as* $f(x, Du_0)$. In the proof of this step, we use hypothesis (H1), which implies, with $p_1 = D(u_\epsilon)$, $p_2 = D(w_\epsilon u_0)$, that

$$|f(x, Du_\epsilon) - f(x, D(w_\epsilon u_0))| \le K(1 + |Du_\epsilon|^{s-\gamma} + |D(w_\epsilon u_0)|^{s-\gamma})|D(u_\epsilon - w_\epsilon u_0)|^\gamma.$$

It can easily be proved that any term in the right-hand side converges to zero strongly in $L^1(\Omega)$. Let us consider, for example, the last term in the case $s > \gamma$. It is enough to apply Hölder's inequality with $p = \frac{2}{s-\gamma}$ and $p' = \frac{2}{2-s+\gamma}$ to get

$$\int_\Omega |D(w_\epsilon u_0)|^{s-\gamma} |D(u_\epsilon - w_\epsilon u_0)|^\gamma \, dx$$
$$\leq \left(\int_\Omega |D(w_\epsilon u_0)|^2 \, dx \right)^{\frac{s-\gamma}{2}} \left(\int_\Omega |D(u_\epsilon - w_\epsilon u_0)|^{\frac{2\gamma}{2-s+\gamma}} \, dx \right)^{\frac{2-s+\gamma}{2}}$$

The condition $s < 2$ implies that $\frac{2\gamma}{2-s+\gamma} < 2$, and the result follows from the corrector result (3.8). We have proved that

$$f(x, Du_\epsilon) - f(x, D(w_\epsilon u_0)) \to 0 \text{ strongly in } L^1(\Omega). \tag{3.9}$$

Let us now prove that

$$f(x, D(w_\epsilon u_0)) \to f(x, Du_0) \text{ strongly in } L^1(\Omega). \tag{3.10}$$

By hypothesis (H3), $Dw_\epsilon \to 0$ a.e. Because of the continuity of f with respect to p (see hypothesis (H1)), we have

$$f(x, D(w_\epsilon u_0)) = f(x, u_0 Dw_\epsilon + w_\epsilon Du_0) \to f(x, Du_0) \text{ a.e.}$$

On the other hand, (3.2) implies

$$f(x, D(w_\epsilon u_0))| \leq C_1(1 + |Dw_\epsilon u_0 + w_\epsilon Du_0|^s) \leq 2C_1(1 + |Dw_\epsilon|^s |u_0|^s + |Du_0|^s).$$

The proof of (3.10) is then achieved using Vitali's convergence theorem since $Dw_\epsilon \to 0$ a.e. and Dw_ϵ is bounded in $(L^2(\Omega))^N$.

From (3.9) and (3.10) we obtain that

$$g_\epsilon = f(x, Du_\epsilon) \to f(x, Du_0) \text{ strongly in } L^1(\Omega),$$

which proves that $g_0 = f(x, Du_0)$. Using this result in the limit problem for u_0 at the end of Step 2, we complete the proof of the theorem. □

Remark 3.1. Hypothesis (H3) is satisfied, for instance, in the case of domains periodically perforated by holes that are balls of critical size. More precisely, we consider in this remark the case where the measure μ_ϵ is given by $\mu_\epsilon = \infty_{T_\epsilon}$ (see the definition in the introduction), where T_ϵ is the union of the balls of radius $a_\epsilon = C_0 \epsilon^{N/(N-2)}$ (when $N \geq 3$) or $a_\epsilon = C_0 \exp(-c_0/\epsilon^2)$ (when $N = 2$), whose centers are distributed at the vertices of a lattice in $\mathbb{R}^N$, with cell size 2ϵ (see [CM, Ex. 2.1]).

An explicit computation, based on the proof of Theorem 2.2 of [CM] (see in particular formula (2.2)), in the case $N \geq 3$ (the case $N = 2$ is similar) gives

$$\int_\Omega |Dw_\epsilon|^\theta \, dx \sim \frac{\text{meas}\, \Omega}{(2\epsilon)^N} S_N \int_{a_\epsilon}^{\epsilon} \left| \frac{dw_\epsilon}{dr} \right|^\theta r^{N-1} \, dr$$
$$\sim \frac{\text{meas}\, \Omega}{(2\epsilon)^N} S_N \int_{a_\epsilon}^{\epsilon} |(N-2)(a_\epsilon)^{N-2} r^{-(N-1)}|^\theta r^{N-1} \, dr$$
$$\sim \text{ Cst } \frac{1}{\epsilon^N} (a_\epsilon)^{(N-2)\theta} (a_\epsilon)^{-(N-1)(\theta-1)+1}$$
$$= \text{Cst} \frac{(a_\epsilon)^{N-\theta}}{\epsilon^N} = \text{ Cst } C_0^{N-\theta} \epsilon^{N \frac{2-\theta}{N-2}},$$

which gives, if $\theta < 2$,

$$\int_\Omega |Dw_\epsilon|^\theta \, dx \to 0.$$

This result proves that, at least for a subsequence, hypothesis (H3) is satisfied for this example.

Remark 3.2. The restriction $s < 2$ is crucial in the above proof but also in the statement of Theorem 3.1, as proved by Casado-Diaz [Ca1].

Indeed the result is drastically different in the "exactly quadratic" case ($s = 2$) since when, for example,

$$f(x, Du) = h(x) + \gamma |Du|^2, \qquad \text{with } h \in L^\infty(\Omega), \tag{3.11}$$

the result of the passage to the limit reads as

$$\begin{cases} -\Delta u_0 + \lambda_0 u_0 + \dfrac{1}{\gamma}(1 - e^{-\gamma u_0})\mu_0 = f(x, Du_0) & \text{in } \Omega, \\ u_0 = 0 & \text{on } \partial\Omega. \end{cases} \tag{3.12}$$

If one compares this result with (3.1_0), one sees that the nonlinearity with respect to the gradient is the same (namely $f(x, Du_0)$), in both (3.1_0) and (3.12), but that the term $u_0\mu_0$ in (3.1_0) has been replaced by $\frac{1}{\gamma}(1 - e^{-\gamma u_0})\mu_0$ in (3.12).

Also, an important fact is that the corrector result (3.8) does not hold true when $s = 2$. In the previous example (3.11) we have in place of (3.8) the result

$$\left\| u_\epsilon - \frac{1}{\gamma} \log(1 + (e^{\gamma u_0} - 1)w_\epsilon) \right\|_{H_0^1(\Omega)} \to 0,$$

while the expression of the corrector in the general case is more complicated.

For more details on the quadratic case, we then refer the reader to the work of Casado-Diaz [Ca1].

Remark 3.3. It is possible to obtain a result analogous to Theorem 3.1 if the nonlinear term also depends on the function u_ϵ. We only need to replace hypothesis (H2) by

$$\text{(H2)}' \quad \begin{cases} |f(x,t,p)| \le C_1(1 + |p|^s), \\ |f(x,t_1,p_1) - f(x,t_2,p_2)| \\ \le n(t_1,t_2)\{(1 + |p_1|^{s-\gamma} + |p_2|^{s-\gamma})|p_1 - p_2|^\gamma + |t_1 - t_2|^r(|p_1|^s + |p_2|^s)\}, \\ \hfill \text{for any } t_1, t_2 \in \mathbb{R}, \text{for any } p_1, p_2 \in \mathbb{R}^N, \end{cases}$$

where $0 < \gamma \le 1$, $\gamma \le s < 2$, $r > 0$, and $n : \mathbb{R}^2 \to \mathbb{R}$ is a function that is bounded on the bounded subsets of $\mathbb{R}^2$. Indeed the first part of (H2)$'$ allows us to obtain the existence of a solution u_ϵ and its boundedness in $H_0^1(\Omega) \cap L^\infty(\Omega)$ (as in Step 1 above) while the second part of (H2)$'$ allows to perform the proof we made in Step 4.

Acknowledgments. We thank L. Boccardo and P. Donato for interesting discussions concerning this paper.

REFERENCES

[BBM] A. Bensoussan, L. Boccardo, and F. Murat, *On a nonlinear partial differential equation having natural growth terms and unbounded solution*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 5 (1988), pp. 347–364.

[BD] L. Boccardo and P. Donato, personal communication, 1993.

[BGM] L. Boccardo, T. Gallouet, and F. Murat, *A unified presentation of two existence results for problems with natural growth*, in Progress in Partial Differential Equations: The Metz Surveys 2, M. Chipot, ed., Pitman Res. Notes in Math. 296, Longman Scientific and Technical, Harlow, UK, 1993, pp. 127–137.

[BMP1] L. Boccardo, F. Murat, and J.-P. Puel, *Existence de solutions faibles pour des équations elliptiques quasi-linéaires à croissance quadratique*, in Nonlinear Partial Differential Equations and their Applications, Collège de France Seminar, vol. IV, H. Brezis and J. L. Lions, eds., Res. Notes in Math. 84, Pitman, London, 1983, pp. 19–73.

[BMP2] ———, *Existence de solutions non bornées pour certaines équations quasi-linéaires*, Portugal. Math., 41 (1982), pp. 507–534.

[Ca1] J. Casado-Diaz, *Sobre la homogeneización de problemas no coercivos y problemas en dominios con agujeros*, Ph.D. thesis, University of Seville, Seville, Spain, 1993.

[Ca2] ———, *The homogenization problem in perforated domains for monotone operators*, to appear.

[CG] J. Casado-Diaz and A. Garroni, *Asymptotic behaviour of nonlinear Dirichlet systems on varying domains*, to appear.

[CM] D. Cioranescu and F. Murat, *Un terme étrange venu d'ailleurs*, in Nonlinear Partial Differential Equations and their Applications, Collège de France Seminar, vols. II and III, H. Brezis and J. L. Lions, eds., Res. Notes in Math. 60 and 70, Pitman, London, 1982, pp. 98–138 and 154–178. English translation: *A strange term coming from nowhere*, in Topics in the Mathematical Modelling of Composite Materials, R. V. Kohn, ed., Progr. Nonlinear Differential Equations Appl., Birkhaüser, Boston, 1995.

[DM1] G. Dal Maso and U. Mosco, *Wiener criteria and energy decay for relaxed Dirichlet problems*, Arch. Rational Mech. Anal., 95 (1986), pp. 345–387.

[DM2] ———, *Wiener criterion and* Γ*-convergence*, Appl. Math. Optim., 15, (1987), pp. 15–63.

[DMM1] G. Dal Maso and F. Murat, *Dirichlet problems in perforated domains for homogeneous monotone operators on* H_0^1, in Calculus of Variations, Homogenization and Continuum Mechanics, World Scientific, Singapore, 1994.

[DMM2] ———, *Asymptotic behaviour and correctors for Dirichlet problems in perforated domains with homogeneous monotone operators*, to appear.

[GDM] A. Garroni and G. Dal Maso, *New results on the asymptotic behaviour of Dirichlet problems in perforated domains*, Math. Models Methods Appl. Sci., 3 (1996), pp. 373–407.

[FT] S. Finzi Vita and N. A. Tchou, *Correctors results for relaxed Dirichlet problems*, Asymptotic Anal., 5 (1992), pp. 269–281.

[KK] J. L. Kazdan and R. J. Kramer, *Invariant criteria for existence of solutions of second order quasilinear elliptic equations*, Comm. Pure Appl. Math., 31 (1978), pp. 619–645.

[LM] P. L. Lions and F. Murat, *Solutions renormalisées d'équations elliptiques*, to appear.

[LP] N. Labani and C. Picard, *Homogenization of a nonlinear Dirichlet problem in a periodically perforated domain*, in Recent Advances in Nonlinear Elliptic Problems, Res. Notes in Math. 208, Pitman, London, 1989, pp. 296–305.

[Mo] A. Mokrane, *Existence of bounded solutions for some nonlinear parabolic equations*, Proc. Roy. Soc. Edinburgh Sect. A, 107, (1987), pp. 313–326.

[M] F. Murat, *Soluciones renormalizades de EDP elipticas no lineales*, in Lectures at the University of Seville, 1992, Publications du Laboratoire d'Analyse Numérique, Université Paris 6, Paris, 1993.

[T] L. Tartar, Cours Peccot, Collège de France, Paris, 1977; partially written in F. Murat, *H-convergence*, Séminaire d'Analyse Numérique et Fonctionnelle, Université d'Alger, Alger, 1977/78. English translation: *H-convergence*, in Topics in the Mathematical Modelling of Composite Materials, R. V. Kohn, ed., Progr. Nonlinear Differential Equations Appl., Birkhaüser, Boston, 1995.

SIAM J. MATH. ANAL.
Vol. 27, No. 3, pp. 997–1014, July 1996

006

STRUCTURE OF RADIAL SOLUTIONS TO $\Delta u + K(|x|)|u|^{p-1}u = 0$ IN $\mathbf{R}^n$*

EIJI YANAGIDA[†]

Abstract. The elliptic equation equation $\Delta u + K(r)|u|^{p-1}u = 0$ in $\mathbf{R}^n$ is studied, where $r = |x|, p > 1, n > 2$, and $K(r) > 0$ for $r \in (0, \infty)$. If $rK_r(r)/K(r)$ is nonincreasing in r, then the structure of radial solutions is determined completely by analyzing linearized equations around the solutions.

Key words. semilinear elliptic equation, radial solutions, structure

AMS subject classifications. 34C10, 35J25

1. Introduction. The purpose of this paper is to investigate the structure of radial solutions to the semilinear elliptic equation

$$\Delta u + K(|x|)|u|^{p-1}u = 0, \qquad x \in \mathbf{R}^n,$$

where $p > 1, n > 2, \Delta = \sum_{i=1}^n \partial^2/\partial x_i^2$, and $|x| = \{\sum_{i=1}^n x_i^2\}^{1/2}$. This equation appears in various fields such as astrophysics, combustion, and differential geometry.

Since radial solutions (i.e., solutions with $u = u(|x|)$ for all $x \in \mathbf{R}^n$) are of particular interest, we will study the initial value problem

$$\text{(1.1)} \qquad \begin{cases} \{r^{n-1}u_r(r)\}_r + r^{n-1}K(r)|u(r)|^{p-1}u(r) = 0, & r > 0, \\ u(0) = \alpha > 0, \end{cases}$$

where $r = |x|$. Throughout this paper, we assume that

$$\text{(K.0)} \qquad \begin{cases} K(r) & \in C^1((0,\infty)), \\ K(r) & > 0 \quad \text{on } (0,\infty), \\ rK(r) & \in L^1(0,1). \end{cases}$$

Then, for any $\alpha > 0$, problem (1.1) has the unique global solution $u(r) \in C([0,\infty)) \cap C^2((0,\infty))$ (see Ni and Yotsutani [12] for the local existence and regularity and Coffman and Ullrich [3] for the global existence). We denote the solution by $u(r;\alpha)$.

We classify the solutions as follows.

Type $R(i)$. $u(r;\alpha)$ has exactly i zeros on $(0,\infty)$, and $r^{n-2}|u(r;\alpha)| \to \beta$ as $r \to \infty$ for some constant $\beta > 0$.

Type $S(i)$. $u(r;\alpha)$ has exactly i zeros on $(0,\infty)$, and $r^{n-2}|u(r;\alpha)| \to \infty$ as $r \to \infty$.

Type O. $u(r;\alpha)$ has infinitely many zeros on $(0,\infty)$.

By virtue of (d) of Lemma 2.1, any solution of (1.1) is classified into one of the above types.

Let λ be a number given by

$$\lambda := \frac{(n-2)p-(n+2)}{2},$$

* Received by the editors January 10, 1994; accepted for publication (in revised form) October 21, 1994.

[†] Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Meguro-ku, Tokyo 152, Japan (yanagida@is.titech.ac.jp).

which is related to the Pohozaev identity (see Lemma 2.2) and plays an important role for the existence of solutions of each type. The existence of a solution of Type $R(i)$ was studied by Yanagida and Yotsutani [17] and later generalized by Naito [11] and Kabeya, Yanagida, and Yotsutani [5]. According to their results, the following theorem holds.

THEOREM A. *Suppose that* $\lim_{r\to 0}\inf rK_r(r)/K(r) > \lambda$ *and* $\lim_{r\to\infty}\sup rK_r(r)/K(r) < \lambda$. *Then there exist* $0 < \alpha_1 < \alpha_2 < \cdots < \infty$ *such that* $u(r;\alpha_{i+1})$ *is of Type* $R(i)$.

Once this theorem is established, it is natural to ask whether the solution of Type $R(i)$ is unique or not, or, more generally, what the entire structure of solutions is. Under various assumptions on $K(r)$ and p, uniqueness of a solution of Type $R(0)$ was proved by Kwong and Li [10] and Yanagida [14], and the entire structure of positive solutions was studied by Kawano, Yanagida, and Yotsutani independently and in conjunction with one another [6, 13, 18]. However, little is known concerning the uniqueness or structure of radial solutions that may change sign. In this paper, we study the structure of radial solutions of (1.1) under the following quite simple condition on $K(r)$:

$$\text{(K.1)} \qquad rK_r(r)/K(r) \text{ is nonincreasing in } r \in (0,\infty).$$

We note that, under this condition, we can define σ and ℓ by

$$\sigma := \lim_{r\to 0} rK_r(r)/K(r) \in (-\infty, +\infty],$$
$$\ell := \lim_{r\to\infty} rK_r(r)/K(r) \in [-\infty, +\infty).$$

The following theorem is a main result of this paper.

THEOREM 1. *Suppose that* (K.1) *and the inequalities* $-\infty \le \ell < \lambda < \sigma \le +\infty$ *are satisfied. Then the following hold.*

(a) *There exist* $0 = \alpha_0 < \alpha_1 < \alpha_2 < \cdots < \infty$ *with* $\lim_{i\to\infty}\alpha_i = \infty$ *such that* $u(r;\alpha)$ *is of Type* $R(i)$ *if and only if* $\alpha = \alpha_{i+1}$, *and* $u(r;\alpha)$ *is of Type* $S(i)$ *for every* $\alpha \in (\alpha_i, \alpha_{i+1})$.

(b) *The ith zero of* $u(r;\alpha)$ *is a strictly decreasing function of* $\alpha \in (\alpha_i,\infty)$ *for every* i.

(c) *Let* $\beta_i := \lim_{r\to\infty} r^{n-2}|u(r;\alpha_i)|$. *Then* $\{\beta_i\}$ *is a monotone increasing positive sequence with* $\lim_{i\to\infty}\beta_i = \infty$.

(d) *The zeros of* $u(r;\alpha_i)$ *separate and are separated by those of* $u(r;\alpha_{i+1})$.

When the inequalities $\ell < \lambda < \sigma$ are not satisfied, we have the next theorem.

THEOREM 2. (a) *If* $rK_r(r)/K(r) \equiv \lambda$ *on* $(0,\infty)$, *then* $u(r;\alpha)$ *is of Type* $R(0)$ *for every* $\alpha \in (0,\infty)$.

(b) *If* $rK_r(r)/K(r) \le \lambda$ *and* $rK_r(r)/K(r) \not\equiv \lambda$ *on* $(0,\infty)$, *then* $u(r;\alpha)$ *is of Type* $S(0)$ *for every* $\alpha \in (0,\infty)$.

(c) *If* $rK_r(r)/K(r) \ge \lambda$ *and* $rK_r(r)/K(r) \not\equiv \lambda$ *on* $(0,\infty)$, *then* $u(r;\alpha)$ *is of Type* O *for every* $\alpha \in (0,\infty)$. *In addition, if* (K.1) *holds, then the ith zero of* $u(r;\alpha)$ *is a strictly decreasing function of* $\alpha \in (0,\infty)$ *for every* i.

This theorem, except the latter part of (c), was proved in [1, 4, 8]. We will give a simplified proof for self-containedness. We note that, by virtue of Theorems 1 and 2, the structure of solutions to (1.1) under the condition (K.1) can be completely determined.

The outline of this paper is as follows. In §2, we give several fundamental properties of solutions to (1.1). In §3, we give two important propositions concerning the properties of zeros of $u(r;\alpha)$. Then we give a proof of Theorem 1 by using these propositions. A proof of Theorem 2 is given in §4. Sections 5 and 6 are devoted to proofs of the propositions.

2. Fundamental properties of solutions.

In this section, we collect fundamental facts that will be used in subsequent sections.

LEMMA 2.1. *Any solution of* (1.1) *has the following properties:*

(a) $u_r(r;\alpha) = -\int_0^r (s/r)^{n-1}K(s)|u(s;\alpha)|^{p-1}u(s;\alpha)\,ds$.

(b) $u(r;\alpha) = \alpha - \frac{1}{n-2}\int_0^r \{1-(s/r)^{n-2}\}sK(s)|u(s;\alpha)|^{p-1}u(s;\alpha)\,ds$.

(c) $\lim_{r\to 0} ru_r(r;\alpha) = 0$.

(d) *If* $u(r;\alpha)>0$ (*resp.*, $u(r;\alpha)<0$) *on* (R,∞) *for some* $R\geq 0$, *then* $\{r^{n-2}u(r;\alpha)\}_r > 0$ (*resp.*, $\{r^{n-2}u(r;\alpha)\}_r < 0$) *on* (R,∞).

(e) *If* $r^{n-2}u(r;\alpha)\to\gamma$ *as* $r\to\infty$ *for some constant* $\gamma\neq 0$, *then* $r^{n-1}u_r(r;\alpha)\to -(n-2)\gamma$ *as* $r\to\infty$.

(f) *If* $u(r;\alpha)$ *is not of Type O and satisfies* $u(r;\alpha)\to 0$ *as* $r\to\infty$, *then* $ru_r(r;\alpha)\to 0$ *as* $r\to\infty$.

Proof. For (a) and (b), see Propositions 4.1 and 4.2 of [12]. For (c), see (4.5) of [12]. Assertions (d) and (e) were proved in Lemmas 7.1 and 7.2 of [12] for positive solutions. The proof applies to solutions of (1.1) with an obvious change. By (a) and (d), if $u(r;\alpha)>0$ (resp., $u(r;\alpha)<0$) in a neighborhood of $r=\infty$ and $u(r;\alpha)\to 0$ as $r\to\infty$, then there exists $R\geq 0$ such that

$$0 < -ru_r < (n-2)u \quad (\text{resp.}, 0 < ru_r < -(n-2)u) \tag{2.1}$$

on (R,∞). Since $u\to 0$ as $r\to\infty$, (f) holds. □

The following identity is a variant of the well-known Pohozaev identity.

LEMMA 2.2. *Any solution* $u = u(r;\alpha)$ *of* (1.1) *satisfies the identity*

$$\frac{d}{dr}P(r;u) \equiv \frac{1}{p+1}r^{n-1}K(r)\{rK_r(r)/K(r)-\lambda\}|u|^{p+1},$$

where

$$P(r;u) := \frac{1}{2}r^{n-1}u_r\{ru_r+(n-2)u\} + \frac{1}{p+1}r^nK(r)|u|^{p+1}.$$

Proof. The equation in (1.1) can be written as

$$\{ru_r+(n-2)u\}_r = -rK(r)|u|^{p-1}u. \tag{2.2}$$

Carrying out the differentiation and using (1.1) and (2.2), we obtain

$$\begin{aligned}\frac{d}{dr}P(r;u) &= \frac{1}{2}(r^{n-1}u_r)_r\{ru_r+(n-2)u\} + \frac{1}{2}r^{n-1}u_r\{ru_r+(n-2)u\}_r \\ &\quad + \frac{1}{p+1}\{r^nK(r)|u|^{p+1}\}_r \\ &= -\frac{1}{2}\{r^{n-1}K(r)|u|^{p-1}u\}\{ru_r+(n-2)u\} - \frac{1}{2}r^{n-1}u_r\{rK(r)|u|^{p-1}u\} \\ &\quad + \frac{1}{p+1}\{r^nK(r)|u|^{p+1}\}_r \\ &= \frac{1}{p+1}r^{n-1}K(r)\{rK_r(r)/K(r)-\lambda\}|u|^{p+1}.\end{aligned}$$

Thus we get the desired identity. □

The following characterizations of solutions to (1.1) in terms of $P(r;u)$ are useful.

LEMMA 2.3. (a) *For any solution* $u = u(r;\alpha)$ *of* (1.1), *there exists a sequence* $\{\varepsilon_j\}$ *such that* $\varepsilon_j \to 0, \varepsilon_j^n K(\varepsilon_j) \to 0$, *and* $P(\varepsilon_j;u) \to 0$ *as* $j \to \infty$.

(b) *If* $u = u(r;\alpha)$ *is of Type* $R(i)$, *then there exists a sequence* $\{\bar{r}_j\}$ *such that* $\bar{r}_j \to \infty$ *and* $P(\bar{r}_j;u) \to 0$ *as* $j \to \infty$.

(c) *If* $u = u(r;\alpha)$ *is of Type* $S(i)$ *and satisfies* $uu_{\mathrm{r}} < 0$ *in a neighborhood of* $r = \infty$, *then there exists a sequence* $\{\hat{r}_j\}$ *such that* $\hat{r}_j \to \infty$ *as* $j \to \infty$ *and* $P(\hat{r}_j;u) < 0$ *for every* j.

(d) *If* $u = u(r;\alpha)$ *is of Type* O, *then there exists a sequence* $\{\tilde{r}_j\}$ *such that* $\tilde{r}_j \to \infty$ *as* $j \to \infty$ *and* $P(\tilde{r}_j;u) > 0$ *for every* j.

Proof. Since $rK(r) \in L^1(0,1)$, there exists a sequence $\{\varepsilon_j\}$ such that $\varepsilon_j \to 0$ and $\varepsilon_j^2 K(\varepsilon_j) \to 0$ as $j \to \infty$. On the other hand, by Lemma 2.1(c), we have

$$\lim_{r\to 0} r^{n-1}u_r\{ru_r + (n-2)u\} = 0.$$

Hence it follows that $\varepsilon_j^n K(\varepsilon_j) \to 0$ and $P(\varepsilon_j;u) \to 0$ as $j \to \infty$ in view of $n > 2$. Thus (a) holds.

If $u = u(r;\alpha)$ is of Type $R(i)$ and $r^{n-2}u(r;\alpha) \to \gamma$ as $r \to \infty$, it follows from Lemma 2.1(a),(e) that

$$r^{n-1}u_r(r;\alpha) = -\int_0^r s^{n-1}K(s)|u(s;\alpha)|^{p-1}u(s;\alpha)\,ds \to -(n-2)\gamma$$

as $r \to \infty$. Since the above integral is convergent as $r \to \infty$, there exists a sequence $\{\bar{r}_j\}$ such that $\bar{r}_j \to \infty$ and $\bar{r}_j^n K(\bar{r}_j)|u(\bar{r}_j;\alpha)|^p \to 0$ as $j \to \infty$. On the other hand, we have

$$\lim_{r\to\infty} r^{n-1}u_r\{ru_r + (n-2)u\} = 0$$

in view of Lemma 2.1(e),(f). These imply that $P(\bar{r}_j;u) \to 0$ as $j \to \infty$. Thus (b) holds.

Next, we prove (c) by assuming that $u > 0$ and $u_r < 0$ in a neighborhood of $r = \infty$. If r is sufficiently large, we have

$$\begin{aligned} P(r;u) &= \frac{1}{2}r^2u_r\{r^{n-2}u_r + (n-2)r^{n-3}u\} + \frac{1}{p+1}ru\{r^{n-1}K(r)|u|^{p-1}u\} \\ &= \frac{1}{2}r^2u_r(r^{n-2}u)_r - \frac{1}{p+1}ru(r^{n-1}u_r)_r \\ &= -r^nuu_r\left\{-\frac{1}{2}\frac{(r^{n-2}u)_r}{r^{n-2}u} + \frac{1}{p+1}\frac{(r^{n-1}u_r)_r}{r^{n-1}u_r}\right\} \\ &= -r^nuu_r\frac{d}{dr}\left\{-\frac{1}{2}\log(r^{n-2}u) + \frac{1}{p+1}\log(-r^{n-1}u_r)\right\} \\ &= -r^nuu_r\frac{d}{dr}\left\{\left(\frac{1}{p+1} - \frac{1}{2}\right)\log(r^{n-2}u) + \frac{1}{p+1}\log\left(\frac{-r^{n-1}u_r}{r^{n-2}u}\right)\right\}. \end{aligned}$$

Here $r^{n-2}u \to \infty$ as $r \to \infty$ if $u = u(r;\alpha)$ is of Type $S(i)$. Moreover, it follows from (2.1) that $-ru_r(r)/u(r) < n-2$ if r is sufficiently large. Hence we have

$$\left(\frac{1}{p+1} - \frac{1}{2}\right)\log(r^{n-2}u) + \frac{1}{p+1}\log\left(\frac{-r^{n-1}u_r}{r^{n-2}u}\right) \to -\infty$$

as $r \to \infty$. This implies (c). The proof in the case $u < 0$ and $u_r > 0$ in a neighborhood of $r = \infty$ can be obtained in the same manner.

Finally let $u = u(r;\alpha)$ be of Type O and let $z_j(\alpha)$ denote the jth zero of $u = u(r;\alpha)$. If $u_r(z_j(\alpha);\alpha) = 0$, then $u(r;\alpha) \equiv 0$ for all r, contradicting $u(0;\alpha) = \alpha > 0$. Hence we obtain $u_r(z_j(\alpha);\alpha) \neq 0$ for every j. Then

$$P(z_j(\alpha);u) = \frac{1}{2} z_j(\alpha)^n u_r(z_j(\alpha);\alpha)^2 > 0. \tag{2.3}$$

Thus (d) holds by taking $\tilde{r}_j = z_j(\alpha)$. □

LEMMA 2.4. *Suppose that $rK(r) \notin L^1(1,\infty)$. If $u = u(r;\alpha)$ is not of Type O, then $uu_r < 0$ in a neighborhood of $r = \infty$.*

Proof. We consider the case where $u(r;\alpha) > 0$ in a neighborhood of $r = \infty$. The proof in the case where $u(r;\alpha) < 0$ in a neighborhood of $r = \infty$ is obtained in the same manner.

Since $(r^{n-1}u_r)_r = -r^{n-1}K(r)|u|^{p-1}u < 0$ if $u > 0, r^{n-1}u_r$ is a strictly decreasing function of r in a neighborhood of $r = \infty$. This means that $u_r(r;\alpha) > 0$ near $r = \infty$ or $u_r(r;\alpha) < 0$ near $r = \infty$. Suppose that the former holds. Then there exist $\delta > 0$ and $R > 0$ such that $u(r;\alpha) > \delta$ on $[R,\infty)$. Integrating (2.2) over $[R,r]$, we get

$$\begin{aligned} &ru_r(r;\alpha) + (n-2)u(r;\alpha) \\ &\quad = Ru_r(R;\alpha) + (n-2)u(R;\alpha) \\ &\qquad - \int_R^r sK(s)|u(s;\alpha)|^{p-1}u(s;\alpha)\,ds \\ &\quad \le Ru_r(R;\alpha) + (n-2)u(R;\alpha) \\ &\qquad - \delta^p \int_R^r sK(s)\,ds. \end{aligned}$$

Here, by assumption, the right-hand side diverges to $-\infty$ as $r \to \infty$, while the left-hand side is positive in view of Lemma 2.1(d). This is a contradiction. Thus it is shown that $u_r(r;\alpha) < 0$ in a neighborhood of $r = \infty$. □

LEMMA 2.5. *If $\lim_{r\to\infty} \sup rK_r(r)/K(r) < \lambda$, then $u(r;\alpha)$ is not of Type O for any $\alpha > 0$.*

Proof. Set

$$w(t;\alpha) := r^{n-2}u(r;\alpha), \qquad t := r^{n-2}.$$

Then the equation in (1.1) is rewritten as

$$w_{tt} + M(t)|w|^{p-1}w = 0, \tag{2.4}$$

where

$$M(t) := (n-2)^{-2}t^{-p-(n-4)/(n-2)}K(t^{1/(n-2)}).$$

It was shown in Theorem 2 of Kiguradze [7] that if $M(t)$ satisfies

$$\frac{d}{dt}\{t^{(p+3)/2+\varepsilon}M(t)\} \le 0 \quad \text{on } (T,\infty) \tag{2.5}$$

for some $\varepsilon > 0$ and $T > 0$, then any solution of (2.4) has at most a finite number of zeros on (T,∞).

Changing the variable from t to r, we get

$$\begin{aligned}\frac{d}{dt}\{t^{(p+3)/2+\varepsilon}M(t)\} &= (n-2)^2\frac{d}{dt}\{t^{(3-p)/2+\varepsilon-(n-4)/(n-2)}K(t^{1/(n-2)})\}\\ &= (n-2)^2\frac{dr}{dt}\frac{d}{dr}\{r^{(n-2)((3-p)/2+\varepsilon)-(n-4)}K(r)\}\\ &= (n-2)r^{n-3}\frac{d}{dr}\{r^{-\lambda+(n-2)\varepsilon}K(r)\}\\ &= (n-2)r^{n-3}K(r)\{rK_r/K(r)-\lambda+(n-2)\varepsilon\}.\end{aligned}$$

By assumption on $K(r)$, if we take $\varepsilon > 0$ sufficiently small and $T > 0$ sufficiently large, then the condition (2.5) is satisfied. This completes the proof. □

LEMMA 2.6. *If* (K.1) *holds, then* $\sigma = \lim_{r\to 0} rK_r(r)/K(r) \in (-2, +\infty]$.

Proof. Suppose that $\sigma \le -2$. Then, by (K.1), we have $rK_r(r)/K(r) \le -2$ on $(0,\infty)$, which is equivalent to $\{r^2K(r)\}_r \le 0$ on $(0,\infty)$. Hence there exists a constant $C > 0$ such that $r^2K(r) \ge C$ on (0, 1). However this contradicts the assumption $rK(r) \in L^1(0,1)$. □

LEMMA 2.7. *If* $\ell = \lim_{r\to\infty} rK_r(r)/K(r) \in (-\infty, \lambda)$, *then* $r^{n-1-(n-2)p}K(r) \in L^1(1,\infty)$.

Proof. We have

$$\lambda + n - (n-2)p = -\frac{(n-2)(p-1)}{2} < 0.$$

Hence, if we take $\varepsilon > 0$ small enough, then

$$\begin{aligned}&\{r^{n-(n-2)p+\varepsilon}K(r)\}_r\\ &\quad = r^{n-1-(n-2)p+\varepsilon}K(r)\{rK_r/K(r)+n-(n-2)p+\varepsilon\}\\ &\quad < r^{n-1-(n-2)p}K(r)\{rK_r/K(r)-\lambda\} < 0\end{aligned}$$

for every sufficiently large $r > 0$. Therefore there exists a constant $C > 0$ such that

$$r^{n-1-(n-2)p}K(r) < Cr^{-1-\varepsilon} \quad \text{for } r > 1.$$

This implies that $r^{n-1-(n-2)p}K(r) \in L^1(1,\infty)$. □

3. Proof of Theorem 1.

In this section, we give a proof of Theorem 1.

Let $U = U(r)$ be the unique solution of the following linearized equation of (1.1) at $u(r;\alpha)$:

$$(3.1) \qquad \begin{cases} (r^{n-1}U_r)_r + pr^{n-1}K(r)|u(r;\alpha)|^{p-1}U = 0, & r > 0,\\ U(0) = 1. \end{cases}$$

Differentiating (1.1) with respect to α, we see that the unique solution of this equation is given by

$$U(r) = \frac{\partial}{\partial\alpha}u(r;\alpha).$$

Let $z_j(\alpha)$ denote the jth zero of $u(r;\alpha)$, and let τ_j denote the jth zero of $U(r)$.

The following two propositions are the most technical part in this paper. We will give their proofs in §§5 and 6, respectively.

PROPOSITION 3.1. *Suppose that* (K.1) *holds. If $u(r;\alpha)$ has at least $i\,(>0)$ zeros on $(0,\infty)$, then $U(r)$ has exactly i zeros on $(0,z_i(\alpha))$ satisfying $\tau_1\in(0,z_1(\alpha))$ and $\tau_j\in(z_{j-1}(\alpha),z_j(\alpha)), j=2,3,\dots,i$.*

PROPOSITION 3.2. *Suppose that* (K.1) *holds and $rK_r(r)/K(r)\not\equiv\lambda$. If $u(r;\alpha)$ is of Type $R(i)$, then there exists $\delta>0$ such that $u(r;\alpha')$ has at least $i+1$ zeros for every $\alpha'\in(\alpha,\alpha+\delta)$.*

Now let us complete the proof of Theorem 1 by using these propositions.

Proof of Theorem 1(b). Differentiating $u(z_j(\alpha);\alpha)=0$ with respect to α, we obtain

$$u_r(z_j(\alpha);\alpha)\frac{d}{d\alpha}z_j(\alpha)+U(z_j(\alpha))=0. \tag{3.2}$$

Here, by Proposition 3.1, U satisfies

$$\begin{cases} U(z_j(\alpha))<0 & \text{if } u_r(z_j(\alpha);\alpha)<0,\\ U(z_j(\alpha))>0 & \text{if } u_r(z_j(\alpha);\alpha)>0.\end{cases}$$

Hence we obtain $\frac{d}{d\alpha}z_j(\alpha)<0$. □

Proof of Theorem 1(a). Since any zero of $u(r;\alpha)$ is a strictly decreasing continuous function of α, the zero does not disappear as α increases.

Let $\{\alpha_i\}$ be a nondecreasing sequence defined by $\alpha_0=0$ and

$$\alpha_i:=\sup\{\alpha|u(r;\alpha)\text{ has at most } i-1 \text{ zeros }\},\qquad i=1,2,3,\dots. \tag{3.3}$$

Since the number of zeros of $u(r;\alpha)$ never decreases as α increases, we see from Lemma 2.5 that $\alpha_i<\infty$ for every i and that $\alpha_i\to\infty$ as $i\to\infty$. Moreover, in view of Proposition 3.2, $u(r;\alpha)$ must be of Type $S(i)$ for every $\alpha\in(\alpha_i,\alpha_{i+1})$. This implies that the sequence $\{\alpha_i\}$ given in Theorem A coincides with the sequence $\{\alpha_i\}$ given by (3.3). Thus the proof is complete. □

Proof of Theorem 1(c). We introduce the Kelvin transformation

$$v(s):=r^{n-2}u(r),\qquad s:=r^{-1}. \tag{3.4}$$

By this transformation, the equation in (1.1) is rewritten as

$$(s^{n-1}v_s)_s+s^{n-1}L(s)|v|^{p-1}v=0,$$

where

$$L(s):=s^{2\lambda}K(s^{-1}).$$

It is clear that $L(s)\in C^1((0,\infty))$ and $L(r)>0$ on $(0,\infty)$. Also, if $\ell<\lambda$, then it follows from Lemma 2.7 that

$$\lim_{\varepsilon\to0}\int_\varepsilon^1 sL(s)\,ds=\int_1^\infty r^{n-1-(n-2)p}K(r)\,dr<\infty.$$

Thus $L(s)$ satisfies the condition (K.0). Moreover, we have

$$\begin{aligned} sL_s(s)/L(s)&=s\{2\lambda s^{2\lambda-1}K(s^{-1})-s^{2\lambda-2}K_r(s^{-1})\}/s^{2\lambda}K(s^{-1})\\ &=2\lambda-rK_r(s^{-1})/K(s^{-1}).\end{aligned}$$

Hence $sL_s(s)/L(s)$ is a nonincreasing function of s and satisfies

$$\lim_{s\to 0} sL_s(s)/L(s) = 2\lambda - \ell > \lambda,$$
$$\lim_{s\to 0} sL_s(s)/L(s) = 2\lambda - \sigma < \lambda.$$

Thus $L(s)$ satisfies the same condition as $K(r)$.

Now let us consider the initial value problem

$$(3.5) \qquad \begin{cases} (s^{n-1}v_s)_s + s^{n-1}L(s)|v|^{p-1}v = 0, \\ v(0) = \beta > 0. \end{cases}$$

We denote by $v(s;\beta)$ the unique solution of this problem. By applying Theorem 1(a) to (3.5), there exist $0 = \beta_0 < \beta_1 < \beta_2 < \cdots < \infty$ with $\lim_{i\to\infty}\beta_i = \infty$ such that $v(s;\beta)$ is of Type $R(i)$ if and only if $\beta = \beta_i$. In view of (3.4), we see that $u(r;\alpha_i)$ and $v(s;\beta_i)$ are related by

$$(3.6) \qquad u(r;\alpha_i) \equiv (-1)^{i+1}r^{-(n-2)}v(r^{-1};\beta_i) \quad \text{on } (0,\infty)$$

and that $u(r;\alpha_i)$ satisfies

$$\lim_{r\to\infty} r^{n-2}|u(r;\alpha_i)| = \beta_i.$$

This proves (c). □

Proof of Theorem 1(d). Let $z_j(\alpha)$ and $y_j(\beta)$ be the jth zero of $u(r;\alpha)$ and $v(s;\beta)$, respectively. By (b) and $\alpha_i < \alpha_{i+1}$, we have

$$(3.7) \qquad z_j(\alpha_i) > z_j(\alpha_{i+1}).$$

Similarly, by applying (a) and (b) to (3.5), we have

$$y_{i-j}(\beta_i) > y_{i-j}(\beta_{i+1}).$$

Here, in view of (3.4) and (3.6), we see that $y_{i-j}(\beta_i) = 1/z_j(\alpha_i)$ and $y_{i-j}(\beta_{i+1}) = 1/z_{j+1}(\alpha_{i+1})$. Hence we get

$$(3.8) \qquad z_j(\alpha_i) < z_{j+1}(\alpha_{i+1}).$$

Thus, by (3.7) and (3.8), we obtain

$$z_j(\alpha_{i+1}) < z_j(\alpha_i) < z_{j+1}(\alpha_{i+1})$$

for every i and j. This completes the proof of (d). □

4. Proof of Theorem 2. In this section we give a proof of Theorem 2.

Proof of Theorem 2(a). Assume that $rK_r(r)/K(r) \equiv \lambda$ on $(0,\infty)$. Then, by Lemma 2.2 and Lemma 2.3(a), we have $P(r;u) \equiv P(\varepsilon_j;u)$ on $[\varepsilon_j,\infty)$. Letting $j\to\infty$, we obtain $P(r;u) \equiv 0$ on $(0,\infty)$. Hence, in view of (2.3), $u(r;\alpha)$ has no zero on $(0,\infty)$. Then it follows from Lemma 2.1 that $u > 0$ and $u_r < 0$ on $(0,\infty)$. Thus, by Lemma 2.3, $u(r;\alpha)$ must be of Type $R(0)$ for every $\alpha \in (0,\infty)$. □

Proof of Theorem 2(b). Assume that $rK_r(r)/K(r) \le \lambda$ and $rK_r(r)/K(r) \not\equiv \lambda$ on $(0,\infty)$. Then, by Lemma 2.2 and Lemma 2.3(a), we have $P(r;u) \le 0$ on $(0,\infty)$ and there exist $R > 0$ and $\delta > 0$ such that $P(r;u) < -\delta$ on (R,∞). Hence, in view of (2.3), $u(r;\alpha)$ has no zero on $(0,\infty)$. Then it follows from Lemma 2.1 that $u > 0$ and $u_r < 0$ on $(0,\infty)$. Thus, by Lemma 2.3, $u(r;\alpha)$ must be of Type $S(0)$ for every $\alpha \in (0,\infty)$. Thus (b) is proved. □

Proof of Theorem 2(c). Assume that $rK_r(r)/K(r) \geq \lambda$ and $rK_r(r)/K(r) \not\equiv \lambda$ on $(0,\infty)$. The inequality $rK_r(r)/K(r) \geq \lambda$ is equivalent to $\{r^{-\lambda}K(r)\}_r \geq 0$. Hence there exists a constant $C > 0$ such that $K(r) \geq Cr^{\lambda}$ on $(1,\infty)$. Here

$$\lambda = \frac{(n-2)p-(n+2)}{2} = \frac{(n-2)(p-1)}{2} - 2 > -2.$$

This implies that $rK(r) \notin L^1(1,\infty)$. On the other hand, by Lemma 2.2, there exist $R > 0$ and $\delta > 0$ such that $P(r;u) > \delta$ on (R,∞). Hence, by Lemmas 2.3 and 2.4, $u(r;\alpha)$ must be of Type O for every $\alpha \in (0,\infty)$. Moreover, by Proposition 3.1 and (3.2), we obtain $\frac{d}{d\alpha}z_j(\alpha) < 0$ for every j. Thus the proof is complete. □

5. Proof of Proposition 3.1. In this section, we give a proof of Proposition 3.1. Throughout this section, we assume that $u(r;\alpha)$ has at least $i\,(> 0)$ zeros on $(0,\infty)$. By developing the ideas in [9, 15, 16], we will construct a comparison function that oscillates faster than $U(r)$.

We prepare a few lemmas. For convenience, we set $z_0(\alpha) = 0$ in the following.

LEMMA 5.1. *Suppose that* (K.1) *holds. Then the inequality* $(ru_r/u)_r < 0$ *holds for* $r \in (z_{j-1}(\alpha), z_j(\alpha)), j = 1,2,3,\dots,i$.

Proof. By (2.2), we have

$$\begin{aligned}(ru_r/u)_r &= \frac{(ru_r)_r u - ru_r^2}{u^2}\\ &= -\frac{u_r\{ru_r + (n-2)u\} + rK(r)|u|^{p+1}}{u^2}.\end{aligned}$$

Since $p > 1$ and $K(r) > 0$, we obtain

$$(ru_r/u)_r < -\frac{2P(r;u)}{r^{n-1}u^2}.$$

Thus it is sufficient to show that $P(r;u) \geq 0$ on $(0, z_i(\alpha))$.

By (K.1) and Lemma 2.2, there exists $R \in [0,\infty]$ such that

$$\begin{aligned}&\frac{d}{dr}P(r;u) \geq 0 \quad \text{for } r \in (0,R),\\ &\frac{d}{dr}P(r;u) \leq 0 \quad \text{for } r \in (R,\infty).\end{aligned}$$

Hence we have $P(r;u) \geq P(\varepsilon_j;u)$ for $r \in [\varepsilon_j, R)$, where $\{\varepsilon_j\}$ is the sequence as in Lemma 2.3(a). Letting $j \to \infty$, we obtain $P(r;u) \geq 0$ for $r \in (0,R)$. Thus, if $z_i(\alpha) \leq R$, then $P(r;u) \geq 0$ for $r \in (0, z_i(\alpha))$. Conversely, if $z_i(\alpha) > R$, then it follows from (2.3) that $P(r;u) \geq P(z_i(\alpha);u) > 0$ for $r \in (R, z_i(\alpha)]$. Thus the proof is complete. □

Let L be a linear operator defined by

$$L[U] = (r^{n-1}U_r)_r + pr^{n-1}K(r)|u(r;\alpha)|^{p-1}U.$$

Then (3.1) is equivalent to $L[U(r)] = 0$ and $U(0) = 1$.

LEMMA 5.2. *Suppose that* (K.1) *holds. Then there exist sequences* $\{\mu_j\}$ *and* $\{\rho_j\}, j = 1, 2, \ldots, i$, *such that*

(a) $0 < \mu_1 < +\infty$;

(b) $-\infty < \mu_i \leq \mu_{i-1} \leq \cdots \leq \mu_1 < \infty$;

(c) $\rho_j \in (z_{j-1}(\alpha), z_j(\alpha))$;

(d) *if* j *is odd, then*

$$\begin{cases} \mu_j u + r u_r > 0 \textit{ and } L[\mu_j u + r u_r] \leq 0 & \textit{for } r \in (z_{j-1}(\alpha), \rho_j), \\ \mu_j u + r u_r < 0 \textit{ and } L[\mu_j u + r u_r] \geq 0 & \textit{for } r \in (\rho_j, z_j(\alpha)), \end{cases}$$

and if j *is even, then*

$$\begin{cases} \mu_j u + r u_r < 0 \textit{ and } L[\mu_j u + r u_r] \geq 0 & \textit{for } r \in (z_{j-1}(\alpha), \rho_j), \\ \mu_j u + r u_r > 0 \textit{ and } L[\mu_j u + r u_r] \leq 0 & \textit{for } r \in (\rho_j, z_j(\alpha)). \end{cases}$$

Proof. We have

$$\begin{aligned} L[u] &= (r^{n-1}u_r)_r + pr^{n-1}K(r)|u|^{p-1}u, \\ L[ru_r] &= \{r^{n-1}(ru_r)_r\}_r + pr^{n-1}K(r)|u|^{p-1}(ru_r). \end{aligned}$$

Using

$$(r^{n-1}u_r)_r = -r^{n-1}K(r)|u|^{p-1}u, \tag{5.1}$$

we get

$$L[u] = (p-1)r^{n-1}K(r)|u|^{p-1}u. \tag{5.2}$$

Using (5.1) and

$$\begin{aligned} (r^{n-1}u_r)_{rr} &= -\{r^{n-1}K(r)|u|^{p-1}u\}_r \\ &= -(n-1)r^{n-2}K(r)|u|^{p-1}u - r^{n-1}K_r(r)|u|^{p-1}u - pr^{n-1}K(r)|u|^{p-1}u_r, \end{aligned}$$

we get

$$L[ru_r] = -r^{n-1}K(r)\{rK_r(r)/K(r) + 2\}|u|^{p-1}u. \tag{5.3}$$

Thus, by (5.2) and (5.3), we obtain

$$L[\mu u + ru_r] = r^{n-1}K(r)|u|^{p-1}u\{(p-1)\mu - rK_r(r)/K(r) - 2\}, \tag{5.4}$$

where μ is a parameter.

Let Γ be a curve on the two-dimensional plane defined by

$$\Gamma := \{(r, rK_r(r)/K(r) + 2); r \in (0, \infty)\}.$$

According to Lemma 5.1, for any $\rho \in (0, z_1(\alpha))$, there exists a unique $\mu = \bar{\mu}_1(\rho) \in (0, +\infty)$ such that

$$\begin{cases} \bar{\mu}_1(\rho)u + ru_r > 0 & \text{for } r \in (0, \rho), \\ \bar{\mu}_1(\rho)u + ru_r < 0 & \text{for } r \in (\rho, z_1(\alpha)), \end{cases}$$

and $\bar{\mu}_1(\rho)$ is a strictly increasing continuous function of ρ satisfying

$$\begin{cases} \bar{\mu}_1(\rho) \downarrow 0 & \text{as } \rho \downarrow 0, \\ \bar{\mu}_1(\rho) \uparrow +\infty & \text{as } \rho \uparrow z_1(\alpha). \end{cases}$$

By Lemma 2.6, this implies that the point $(\rho, (p-1)\bar{\mu}_1(\rho))$ is below Γ if $\rho = +0$ and above Γ if $\rho = z_1(\alpha) - 0$. Hence, by the continuity of Γ, there exist $\rho_1 \in (0, z_1(\alpha))$ and $\mu_1 := \bar{\mu}_1(\rho_1) > 0$ such that $(\rho_1, (p-1)\mu_1)$ is just on the curve Γ, i.e.,

$$(p-1)\mu_1 - \rho_1 K_r(\rho_1)/K(\rho_1) - 2 = 0.$$

Then, since $rK_r(r)/K(r)$ is nonincreasing, we have

$$\begin{cases} (p-1)\mu_1 - rK_r(r)/K(r) - 2 \le 0 & \text{for } r \in (0, \rho_1), \\ (p-1)\mu_1 - rK_r(r)/K(r) - 2 \ge 0 & \text{for } r \in (\rho_1, z_1(\alpha)). \end{cases}$$

By (5.4), this implies

$$\begin{cases} L[\mu_1 u + r u_r] \le 0 & \text{for } r \in (0, \rho_1), \\ L[\mu_1 u + r u_r] \ge 0 & \text{for } r \in (\rho_1, z_1(\alpha)). \end{cases}$$

Similarly, for any $\rho \in (z_1(\alpha), z_2(\alpha))$, there exists $\mu = \bar{\mu}_2(\rho) \in (-\infty, +\infty)$ such that

$$\begin{cases} \bar{\mu}_2(\rho) u + r u_r < 0 & \text{for } r \in (z_1(\alpha), \rho), \\ \bar{\mu}_2(\rho) u + r u_r > 0 & \text{for } r \in (\rho, z_2(\alpha)), \end{cases}$$

and $\bar{\mu}_2(\rho)$ is a strictly increasing continuous function of ρ satisfying

$$\begin{cases} \bar{\mu}_2(\rho) \downarrow -\infty & \text{as } \rho \downarrow z_1(\alpha), \\ \bar{\mu}_2(\rho) \uparrow +\infty & \text{as } \rho \uparrow z_2(\alpha). \end{cases}$$

This implies that the point $(\rho, (p-1)\bar{\mu}_2(\rho))$ is below Γ if $\rho = z_1(\alpha) + 0$ and above Γ if $\rho = z_2(\alpha) - 0$. Hence, by the continuity of Γ, there exists $\rho_2 \in (z_1(\alpha), z_2(\alpha))$ and $\mu_2 := \bar{\mu}_2(\rho_2)$ such that $(\rho_2, (p-1)\mu_2)$ is just on the curve Γ, i.e.,

$$(p-1)\mu_2 - \rho_2 K_r(\rho_2)/K(\rho_2) - 2 = 0.$$

Then, by (K.1), we have

$$\begin{cases} (p-1)\mu_2 - rK_r(r)/K(r) - 2 \le 0 & \text{for } r \in (z_1(\alpha), \rho_2), \\ (p-1)\mu_2 - rK_r(r)/K(r) - 2 \ge 0 & \text{for } r \in (\rho_2, z_2(\alpha)). \end{cases}$$

By (5.4), this implies

$$\begin{cases} L[\mu_2 u + r u_r] \ge 0 & \text{for } r \in (z_1(\alpha), \rho_2), \\ L[\mu_2 u + r u_r] \le 0 & \text{for } r \in (\rho_2, z_2(\alpha)). \end{cases}$$

Repeating this process, we obtain (a), (c), and (d).

Finally, since $\{\rho_j\}$ is a strictly increasing sequence and

$$(p-1)\mu_j - \rho_j K_r(\rho_j)/K(\rho_j) - 2 = 0,$$

it follows from (K.1) that $\{\mu_j\}$ is a nonincreasing sequence. Thus (b) holds. □

Let $V(r)$ be a function on $(0, z_i(\alpha)]$ defined by

$$V(r) = \mu_j u(r;\alpha) + r u_r(r;\alpha)) \quad \text{for } r \in (z_{j-1}(\alpha), z_j(\alpha)],\ j = 1,2,3,\dots,i, \tag{5.5}$$

where $\{\mu_j\}$ is the sequence given in Lemma 5.2. Since $u(r;\alpha) = 0$ at $r = z_j(\alpha), V(r)$ is a continuous function. However $V(r)$ is not smooth at $r = z_j(\alpha)$ if $\mu_j \neq \mu_{j+1}$. Note that, by Lemma 5.2(d), $V(r) = 0$ if and only if $r = \rho_j$.

Let us introduce the Prüfer transformation

$$\begin{cases} u(r;\alpha) = Q(r;u)\cos(\theta(r;u)), \\ -r^{n-1}u_r(r;\alpha) = Q(r;u)\sin(\theta(r;u)), \end{cases}$$

where $Q(r;u) := \{u(r;\alpha)^2 + (r^{n-1}u_r(r;\alpha))^2\}^{1/2} > 0$ and $\theta(r;u)$ is a smooth function of r satisfying $\theta(0;u) = 0$. Since $u_r = r^{1-n}(r^{n-1}u_r)$ and $-(r^{n-1}u_r)_r = r^{n-1}K(r)|u|^{p-1}u$, we have

$$Q_r\cos(\theta) - Q\sin(\theta)\theta_r = -r^{1-n}Q\sin(\theta), \tag{5.6}$$

$$Q_r\sin(\theta) + Q\cos(\theta)\theta_r = r^{n-1}K(r)|u|^{p-1}Q\cos(\theta), \tag{5.7}$$

where $Q = Q(r;u)$ and $\theta = \theta(r;u)$. Multiplying (5.6) and (5.7) by $-\sin(\theta)$ and $\cos(\theta)$, respectively, and adding them, we obtain

$$\theta_{\mathrm{r}}(r;u) = r^{1-n}\sin^2(\theta(r;u)) + r^{n-1}K(r)|u|^{p-1}\cos^2(\theta(r;u)). \tag{5.8}$$

Similarly, put

$$\begin{cases} U(r) = Q(r;U)\cos(\theta(r;U)), \\ -r^{n-1}U_r(r) = Q(r;U)\sin(\theta(r;U)). \end{cases}$$

Then we may assume that $\theta(0;U) = 0$ and

$$\theta_r(r;U) = r^{1-n}\sin^2(\theta(r;U)) + pr^{n-1}K(r)|u|^{p-1}\cos^2(\theta(r;U)). \tag{5.9}$$

As for the function $V(r)$, the situation is a little delicate. Put

$$\begin{cases} V(r) = Q(r;V)\cos(\theta(r;V)), \\ -r^{n-1}V_r(r) = Q(r;V)\sin(\theta(r;V)). \end{cases}$$

Since $0 < \mu_1 < +\infty$, we have $V(0) > 0$. Hence we may assume that $\theta(r;V) \in (-\pi/2, \pi/2)$ if $r > 0$ is sufficiently small. We may assume that $\theta(r;V)$ is continuous on $(z_{j-1}(\alpha), z_j(\alpha))$. However, since $V(r)$ may not be smooth at $r = z_j(\alpha), \theta(r;V)$ may be discontinuous at $r = z_j(\alpha)$. In fact, we have

$$V_r(z_j(\alpha)+0) - V_r(z_j(\alpha)-0) = (\mu_{j+1} - \mu_j)u_r(z_j(\alpha);\alpha).$$

Since $\mu_j \geq \mu_{j+1}$ and

$$\begin{cases} u_r(z_j(\alpha);\alpha) < 0 & \text{if } j > 0 \text{ is odd}, \\ u_r(z_j(\alpha);\alpha) > 0 & \text{if } j > 0 \text{ is even}, \end{cases}$$

we obtain

$$\begin{cases} V_r(z_j(\alpha)-0) \le V_r(z_j(\alpha)+0) & \text{if } j \text{ is odd,} \\ V_r(z_j(\alpha)-0) \ge V_r(z_j(\alpha)+0) & \text{if } j \text{ is even.} \end{cases}$$

Since

$$\begin{cases} V(z_j(\alpha)) < 0 & \text{if } j \text{ is odd,} \\ V(z_j(\alpha)) > 0 & \text{if } j \text{ is even,} \end{cases}$$

we may assume that $\theta(r;V)$ satisfies

$$\theta(z_j(\alpha)+0;V) - \theta(z_j(\alpha)-0;V) \in [0,\pi) \tag{5.10}$$

for $j = 1, 2, \ldots, i-1$.

Now, by Lemma 5.2(d), $V(r)$ satisfies $-V(r)L[V(r)] \ge 0$. This is equivalent to

$$-Q\cos(\theta)\{-Q_r\sin(\theta) - Q\cos(\theta)\theta_r + pr^{n-1}K(r)|u|^{p-1}Q\cos(\theta)\} \ge 0,$$

where $Q = Q(r;V)$ and $\theta = \theta(r;V)$. On the other hand, $r^{n-1}V_r\{V_r - r^{1-n}(r^{n-1}V_r)\} = 0$ is equivalent to

$$-Q\sin(\theta)\{Q_r\cos(\theta) - Q\sin(\theta)\theta_r + r^{1-n}Q\sin(\theta)\} = 0.$$

Adding these two equalities, we obtain the differential inequality

$$\theta_r(r;V) \ge r^{1-n}\sin^2(\theta(r;V)) + pr^{n-1}K(r)|u|^{p-1}\cos^2(\theta(r;V)) \tag{5.11}$$

for $r \ne z_j(\alpha)$.

Now we have the next lemma.

LEMMA 5.3. *Suppose that* (K.1) *holds. Then* $\theta(r;u) < \theta(r;U) \le \theta(r;V)$ *for* $r \in (0, z_i(\alpha)]$.

Proof. We want to apply the comparison theorem (see [2, Chap. 8, Thm. 1.2]). However, since the right-hand sides of (5.8), (5.9), and (5.11) have singularities at $r = 0$, we need careful consideration in a neighborhood of $r = 0$.

First we prove the inequality $\theta(r;u) < \theta(r;U)$. Integrating $uL[U] = 0$ by parts on $[\varepsilon_j, r]$, we get

$$[r^{n-1}(U_r u - U u_r)]_{\varepsilon_j}^r = -(p-1)\int_{\varepsilon_j}^r s^{n-1}K(s)|u|^{p-1}uU\,ds, \tag{5.12}$$

where $\{\varepsilon_j\}$ is the sequence as in Lemma 2.3(a). Here, as $j \to \infty$, we have

$$\begin{aligned} u(\varepsilon_j;\alpha) &\to \alpha, \\ \varepsilon_j^{n-1}u_r(\varepsilon_j;\alpha) &\to 0, \\ U(\varepsilon_j) &\to 1, \\ \varepsilon_j^{n-1}U_r(\varepsilon_j) &= -\textstyle\int_0^{\varepsilon_j} pr^{n-1}K(r)|u|^{p-1}U\,dr \to 0. \end{aligned}$$

Thus we obtain

$$r^{n-1}(U_r u - U u_r) = -(p-1)\lim_{j\to\infty}\int_{\varepsilon_j}^r s^{n-1}K(s)|u|^{p-1}uU\,ds.$$

If $0 < r < \min\{z_1(\alpha), \tau_1\}$, then the right-hand side of this equality is negative. Hence we obtain $-r^{n-1}u_r/u < -r^{n-1}U_r/U$, i.e., $\theta(r; u) < \theta(r; U)$ if $r > 0$ is small. For $r > 0$, the right-hand sides of (5.8) and (5.9) have no singularity. Hence we can apply the comparison theorem to show that $\theta(r; u) < \theta(r; U)$ for all $r > 0$.

Next we prove $\theta(r; U) \leq \theta(r; V)$. Let $r > 0$ be sufficiently small. Integrating $VL[U] = 0$ by parts on $[\varepsilon_j, r]$, we get

$$[r^{n-1}(U_rV - UV_r)]_{\varepsilon_j}^r = -\int_{\varepsilon_j}^r L[V]U\,ds.$$

Here, as $j \to \infty$, we have

$$\begin{aligned} U(\varepsilon_j) &\to 1, \\ \varepsilon_j^{n-1}U_r(\varepsilon_j) &\to 0, \\ V(\varepsilon_j) &\to \mu_1\alpha, \\ \varepsilon_j^{n-1}V_r(\varepsilon_j) &= \varepsilon_j^{n-1}\{(\mu_1+1)u_r(\varepsilon_j;\alpha) + \varepsilon_j u_{rr}(\varepsilon_j;\alpha)\} \\ &= \varepsilon_j^{n-1}\{(\mu_1 - n + 2)u_r(\varepsilon_j;\alpha) - \varepsilon_j K(\varepsilon_j)|u(\varepsilon_j;\alpha)|^{p-1}u(\varepsilon_j;\alpha)\} \to 0. \end{aligned}$$

Thus we obtain

$$r^{n-1}(U_rV - UV_r) = -\lim_{j\to\infty}\int_{\varepsilon_j}^r L[V]U\,ds.$$

Here the right-hand side of this equality is nonnegative in view of Lemma 5.2(d). Thus we obtain $-r^{n-1}U_r/U \leq -r^{n-1}V_r/V$, i.e., $\theta(r; U) \leq \theta(r; V)$ if $r > 0$ is small. For $r > 0$, the right-hand sides of (5.9) and (5.11) have no singularity. Hence we can apply the comparison theorem to show that $\theta(r; U) \leq \theta(r; V)$ for $r \in (0, z_1(\alpha))$. Then, by (5.10), we obtain $\theta(z_1(\alpha) - 0; U) \leq \theta(z_1(\alpha) + 0; V)$. Again we can apply the comparison theorem to show that $\theta(r; U) \leq \theta(r; V)$ for $r \in (z_1(\alpha), z_2(\alpha))$. Repeating this process, we obtain $\theta(r; U) \leq \theta(r; V)$ for $r \in (0, z_i(\alpha))$. □

Let us complete the proof of Proposition 3.1.

Proof of Proposition 3.1. By Lemma 5.3, the inequalities $\theta(r; u) < \theta(r; U) \leq \theta(r; V)$ hold. These imply that $U(r)$ oscillates faster than $u(r; \alpha)$ and more slowly than $V(r)$. Since ρ_j is the unique zero of $V(r)$ in $[z_{j-1}(\alpha), z_j(\alpha)]$, we obtain $z_{j-1}(\alpha) < \rho_j \leq \tau_j < z_j(\alpha)$ for $j = 1, 2, \ldots, i$. This completes the proof. □

6. Proof of Proposition 3.2. In this section we give a proof of Proposition 3.2. For convenience, we put $z_0(\alpha) = 0$ and $z_{i+1}(\alpha) = \infty$ if $u(r; \alpha)$ is of Type $R(i)$.

The following lemma can be proved in the same way as Lemma 5.2 by noting Lemma 2.1(d) and Lemma 2.3(b). We omit the proof.

LEMMA 6.1. *Suppose that* (K.1) *holds. If* $u = u(r; \alpha)$ *is of Type* $R(i)$, *then there exist sequences* $\{\mu_j\}$ *and* $\{\rho_j\}$, $j = 1, 2, \ldots, i+1$, *such that*

(a) $0 < \mu_1 < +\infty$ *and* $-\infty < \mu_{i+1} < n - 2$,

(b) $-\infty < \mu_{i+1} \leq \mu_i \leq \cdots \leq \mu_1 < +\infty$,

(c) $\rho_j \in (z_{j-1}(\alpha), z_j(\alpha))$,

(d) *if* j *is odd, then*

$$\begin{cases} \mu_j u + r u_r > 0 \text{ and } L[\mu_j u + r u_r] \leq 0 & \text{for } r \in (z_{j-1}(\alpha), \rho_j), \\ \mu_j u + r u_r < 0 \text{ and } L[\mu_j u + r u_r] \geq 0 & \text{for } r \in (\rho_j, z_j(\alpha)), \end{cases}$$

and if j is even, then

$$\begin{cases} \mu_j u + r u_r < 0 \text{ and } L[\mu_j u + r u_r] \geq 0 & \text{for } r \in (z_{j-1}(\alpha), \rho_j), \\ \mu_j u + r u_r > 0 \text{ and } L[\mu_j u + r u_r] \leq 0 & \text{for } r \in (\rho_j, z_j(\alpha)). \end{cases}$$

Let $u(r;\alpha)$ be of Type $R(i)$, and let $V(r)$ be defined by

$$V(r) = \mu_j u(r;\alpha) + r u_r(r;\alpha) \quad \text{for } r \in [z_{j-1}(\alpha), z_j(\alpha)), j = 1, 2, \ldots, i+1.$$

Further let $\theta(r;u), \theta(r;U)$, and $\theta(r;V)$ be as in the previous section. Then, similar to Lemma 5.3, we have

$$\theta(r;u) < \theta(r;U) \leq \theta(r;V) \quad \text{on } (0,\infty). \tag{6.1}$$

Moreover, if $rK_r(r)/K(r) \not\equiv \text{Constant}$, then $L[V] \not\equiv 0$ in view of the proof of Lemma 5.2. By [2, Chap. 8, Thm. 1.2], this implies that there exists $R > 0$ such that

$$\theta(r;u) < \theta(r;U) < \theta(r;V) \quad \text{on } [R,\infty). \tag{6.2}$$

LEMMA 6.2. *Suppose that* (K.1) *holds. If $u(r;\alpha)$ is of Type $R(i)$, then $U(r)$ has exactly $i+1$ zeros on $(0,\infty)$ satisfying $\tau_j \in (z_{j-1}(\alpha), z_j(\alpha)), j = 1, 2, \ldots, i+1$.*

Proof. By Proposition 3.1, $U(r)$ has one and only one zero in $(z_{j-1}(\alpha), z_j(\alpha))$ for $j = 1, 2, \ldots, i$. By (6.1), $U(r)$ has at most one zero in $(z_i(\alpha), \infty)$. Thus it is sufficient to show that $U(r)$ has a zero in $(z_i(\alpha), \infty)$.

First we consider the case where $i\,(> 0)$ is even. Then $u(r;\alpha) > 0$ in a neighborhood of $r = \infty$. We will derive a contradiction by assuming that $U(r) > 0$ on $[z_i(\alpha), \infty)$. Similar to (5.12), we have

$$[r^{n-1}(U_r u - U u_r)]^r_{z_j(\alpha)} = -(p-1) \int_{z_i(\alpha)}^r s^{n-1} K(s) u^p U \, ds < 0. \tag{6.3}$$

Since $u(z_i(\alpha);\alpha) = 0, u_r(z_i(\alpha);\alpha) > 0$, and $U(z_i(\alpha)) > 0$, we obtain $r^{n-1}(U_r u - U u_r) < 0$ or $(U/u)_r < 0$ on $[z_i(\alpha), \infty)$. Hence there exists a constant $C > 0$ such that $0 < U < Cu$ on $[z_i(\alpha), \infty)$. Then we have

$$|r^{n-1} u_r U| < C |r^{n-1} u_r u| \to 0 \quad \text{as } r \to \infty.$$

Moreover, integrating (3.1) over $[z_i(\alpha), r]$, we obtain

$$\begin{aligned} |r^{n-1} U_r(r)| &\leq z_i(\alpha)^{n-1} |U_r(z_i(\alpha))| + p \int_{z_i(\alpha)}^r s^{n-1} K(s) u^{p-1} U \, ds \\ &\leq z_i(\alpha)^{n-1} |U_r(z_i(\alpha))| + Cp \int_{z_i(\alpha)}^r s^{n-1} K(s) u^p \, ds. \end{aligned}$$

Since $u(r;\alpha)$ is of Type $R(i)$, the integral in the right-hand side is convergent as $r \to \infty$. Hence $r^{n-1} U_r u \to 0$ as $r \to \infty$. Thus it is shown that the left-hand side of (6.3) is positive if r is sufficiently large. This contradicts (6.3). Hence $U(r)$ must have a zero in $(z_i(\alpha), \infty)$.

The proof in the case i is odd is obtained in the same manner. Finally the proof in the case $i = 0$ can be obtained by replacing $z_i(\alpha)$ by ε_j and letting $j \to \infty$, where $\{\varepsilon_j\}$ is the sequence as in Lemma 2.3(a). □

Now let us complete the proof of Proposition 3.2.

Proof of Proposition 3.2. We consider the case where i is even. The proof in the case i is odd can be obtained similarly. Note that if i is even, then $u(r;\alpha) > 0, U(r) < 0$, and $V(r) < 0$ in a neighborhood of $r = \infty$.

Assume that $u(r;\alpha)$ is of Type $R(i)$. Then, by the continuity of solutions with respect to initial data, there exists $\delta > 0$ such that $u(r;\alpha')$ has at least i zeros for $a' \in (\alpha, \alpha + \delta)$. Put

$$W(r) := \frac{u(r;\alpha') - u(r;\alpha)}{\alpha' - \alpha}.$$

Then $W(r)$ satisfies

$$\begin{cases} (r^{n-1}W_r)_r + r^{n-1}K(r)h(r)W = 0, & r > 0, \\ W(0) = 1, \end{cases}$$

where $h(r)$ is a continuous function given by

$$h(r) := \begin{cases} \dfrac{|u(r;\alpha')|^{p-1}u(r;\alpha') - |u(r;\alpha)|^{p-1}u(r;\alpha)}{u(r;\alpha') - u(r;\alpha)} \\ \qquad\qquad\qquad \text{if } u(r;\alpha') \neq u(r;\alpha), \\ p|u(r;\alpha)|^{p-1} \quad \text{if } u(r;\alpha') = u(r;\alpha). \end{cases}$$

We fix R so large that (6.2) and $R > \tau_{i+1}$ hold, where τ_{i+1} is the $(i+1)$th zero of $U(r)$. By definition, $W(r) \to U(r)$ and $W_r(r) \to U_r(r)$ as $\alpha' \to \alpha$ uniformly in $[0, R]$. Hence, if $\alpha' - \alpha > 0$ is sufficiently small, $W(r)$ has exactly $i + 1$ zeros on $[0, R]$ and satisfies $W(R) < 0$. Moreover, since the inequality $\theta(R; U) < \theta(R; V)$ is equivalent to

$$U_r(R)V(R) - U(R)V_r(R) > 0,$$

we have

$$W_r(R)V(R) - W(R)V_r(R) > 0 \tag{6.4}$$

if $\alpha' - \alpha > 0$ is sufficiently small.

For the behavior of $u(r;\alpha)$ in (R, ∞), one of the following three cases occurs.

Case 1. There exists $R_0 \in (R, \infty)$ such that $0 < u(r;\alpha') < u(r;\alpha)$ for $r \in (R, R_0)$ and $u(R_0;\alpha') - u(R_0;\alpha) = 0$.

Case 2. $0 < u(r;\alpha') < u(r;\alpha)$ for all $r \in (R, \infty)$.

Case 3. There exists $z_{i+1}(\alpha')$ such that $0 < u(r;\alpha') < u(r;\alpha)$ for $r \in (R, z_{i+1}(\alpha'))$ and $u(z_{i+1}(\alpha');\alpha') = 0$.

First suppose that Case 1 holds. Then $W(r)L[V(r)] \leq 0$ on $[R, R_0]$. Integrating this inequality by parts on $[R, R_0]$, we get

$$[r^{n-1}(W_rV - WV_r)]_R^{R_0} \geq \int_R^{R_0} r^{n-1}K(r)\{p|u(r;\alpha)|^{p-1} - h(r)\}V(r)W(r)\,dr. \tag{6.5}$$

Here, by $W(R_0) = 0, W_r(R_0) > 0, V(R_0) < 0$, and (6.4), the left-hand side of (6.5) is negative. On the other hand, since $p|u(r;\alpha)|^{p-1} > h(r)$ in view of $0 < u(r;\alpha') < u(r;\alpha)$ for $r \in (R, R_0)$, the right-hand side of (6.5) is positive. This is a contradiction.

Next suppose that Case 2 holds. Then $u(r;\alpha')$ also must be of Type $R(i)$. Similar to (6.5), integrating $W(r)L[V(r)] \leq 0$ by parts on $[R,r]$ and using (6.4), we obtain

$$r^{n-1}(W_rV - WV_r) > \int_R^r s^{n-1}K(s)\{p|u(s;\alpha)|^{p-1} - h(s)\}V(s)W(s)\,ds.$$

Here the integrand is positive. Hence we can take a constant $C_1 > 0$ such that

$$r^{n-1}(W_rV - WV_r) > C_1 \quad \text{for } r \in (R,\infty),$$

which is equivalent to

$$(W/V)_r > C_1 r^{1-n}V^{-2} \quad \text{for } r \in (R,\infty). \tag{6.6}$$

On the other hand, by the definition of V, we have

$$V(r)/u(r;\alpha) = \mu_{i+1} + ru_r(r;\alpha)/u(r;\alpha) \quad \text{for } r \in (z_i(\alpha),\infty).$$

Hence we see from Lemma 2.1(e) and Lemma 6.1(a) that

$$\lim_{r\to\infty} V(r)/u(r;\alpha) = \mu_{i+1} - (n-2) < 0.$$

Therefore there exist constants $C_2 > C_3 > 0$ such that

$$-C_2r^{2-n} < V(r) < -C_3r^{2-n} \quad \text{for } r \in (R,\infty). \tag{6.7}$$

Hence, by (6.6), we obtain

$$\begin{aligned}(W/V)_r &> C_1r^{1-n}V^{-2}\\ &\geq (C_1/C_2^2)r^{n-3} \quad \text{for } r \in (R,\infty).\end{aligned}$$

This implies that $W/V \to \infty$ as $r \to \infty$. Then it follows from (6.7) that $r^{n-2}W \to -\infty$ as $r \to \infty$. However, this contradicts the fact that both $u(r;\alpha')$ and $u(r;\alpha)$ are of Type $R(i)$.

Consequently, we conclude that Case 3 must hold. □

REFERENCES

[1] K.-S. Cheng, *Positive solutions of semilinear elliptic equations*, Chinese J. Math., 15 (1987), pp. 437–477.

[2] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, 1955.

[3] C. V. Coffman and D. F. Ullrich, *On the continuation of solutions of a certain non-linear differential equation*, Monatsch. Math., 71 (1967), pp. 385–392.

[4] W.-Y. Ding and W.-M. Ni, *On the elliptic equation $\Delta u + Ku^{(n+2)/(n-2)} = 0$ and related topics*, Duke Math. J., 52 (1985), pp 485–506.

[5] Y. Kabeya, E. Yanagida, and S. Yotsutani, *Existence of nodal fast-decay solutions to* $\mathrm{div}(|\nabla u|^{m-2}\nabla u) + K(|x|)|u|^{q-1}u = 0$ *in* $\mathbf{R}^n$, Differential Integral Equations, to appear.

[6] N. Kawano, E. Yanagida, and S. Yotsutani, *Structure theorems for positive radial solutions to $\Delta u + K(|x|)u^p = 0$ in $\mathbf{R}^n$*, Funkcial. Ekvac., 36 (1993), pp. 557–579.

[7] I. T. Kiguradze, *On the conditions of oscillation of solutions of the equation $u'' + a(t)|u|^n \cdot$ sgn $u = 0$*, Časopis Pěst. Mat., 87 (1962), pp. 492–495 (in Russian).

[8] T. Kusano and M. Naito, *Oscillation theory of entire solutions of second order superlinear elliptic equations*, Funkcial. Ekvac., 30 (1987), pp. 269–282.

[9] M.-K. Kwong, *Uniqueness of positive solutions of $\Delta u - u + u^p = 0$* in $\mathbf{R}^n$, Arch. Rational Mech. Anal., 105 (1989), pp. 243–266.

[10] M.-K. Kwong and Y. Li, *Uniqueness of radial solutions of semilinear elliptic equations*, Trans. Amer. Math. Soc., 333 (1992), pp. 339–363.

[11] Y. Naito, *Bounded solutions with prescribed numbers of zeros for the Emden–Fowler differential equation*, Hiroshima Math. J., 24 (1994), pp. 177–220.

[12] W.-M. Ni and S. Yotsutani, *Semilinear elliptic equations of Matukuma-type and related topics*, Japan J. Appl. Math., 5 (1988), pp. 1–32.

[13] E. Yanagida, *Structure of positive radial solutions of Matukuma's equation*, Japan J. Indust. Appl. Math., 8 (1991), pp. 165–173.

[14] ———, *Uniqueness of positive radial solutions of* $\Delta u + g(r)u + h(r)u^p = 0$ in $\mathbf{R}^n$, Arch. Rational Mech. Anal., 115 (1991), pp. 257–274.

[15] ———, *Uniqueness of positive radial solutions of* $\Delta u + f(u, |x|) = 0$, Nonlinear Anal., 19 (1992), pp. 1143–1154.

[16] ———, *Sturmian theory for a class of nonlinear second-order differential equations*, J. Math. Anal. Appl., 187 (1994), pp. 650–662.

[17] E. Yanagida and S. Yotsutani, *Existence of nodal fast-decay solutions to* $\Delta u + K(|x|)|u|^{p-1}u = 0$ *in* $\mathbf{R}^n$, Nonlinear Anal., 22 (1994), pp. 1005–1015.

[18] ———, *Classification of the structure of positive radial solutions to* $\Delta u + K(|x|)u^p = 0$ in $\mathbf{R}^n$, Arch. Rational Mech. Anal., 124 (1993), pp. 239–259.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 1015–1023, July 1996

INITIAL AND INITIAL-BOUNDARY VALUE PROBLEMS FOR A VORTEX FILAMENT WITH OR WITHOUT AXIAL FLOW*

TAKAHIRO NISHIYAMA† AND ATUSI TANI†

Abstract. The equations which describe the motion of a vortex filament with or without an axial flow inside its core are considered. The initial and the initial-boundary value problems are proved to have unique and smooth solutions globally in time. These results are obtained by adding vanishing parabolic terms which conserve the length of the filament.

Key words. vortex filament, perfect fluid, localized induction equation, initial and initial-boundary value problems, unique and smooth solvability

AMS subject classifications. 35Q, 76C

1. Introduction.

The system of equations

$$x_t = x_s \times x_{ss} + a\left\{x_{sss} + \frac{3}{2}x_{ss} \times (x_s \times x_{ss})\right\} \tag{1.1}$$

approximately describes the deformation of a vortex filament with or without axial velocity in its thin core, in a perfect fluid. Here $x = x(s,t)$ denotes the position of a point on the filament in $\boldsymbol{R}^3$ as a vector-valued function of arclength s ($\in \boldsymbol{R}$) and time t (> 0), and a real constant a represents the magnitude of the effect of the axial flow. In particular, (1.1) with $a = 0$, from which the axial-flow effect is absent, is called the localized induction equation (LIE).

Since Da Rios [1] formulated the LIE in 1906, many authors have studied it from various points of view (see [9], [10], and the references therein). In [8], we proved the weak solvability of some initial and initial-boundary value problems for the LIE, although the expected uniqueness and smoothness of the solution were not found. On the other hand, (1.1) with $a \neq 0$ was originally derived by Fukumoto and Miyazaki [2] as a generalization of the LIE from the Moore–Saffman equation in [7].

Differentiating (1.1) with respect to s and setting $v = x_s$, we have

$$v_t = v \times v_{ss} + a\left\{v_{sss} + \frac{3}{2}v_{ss} \times (v \times v_s) + \frac{3}{2}v_s \times (v \times v_{ss})\right\}. \tag{1.2}$$

Impose the initial condition

$$v(s,0) = v_0(s), \quad |v_0| = 1, \tag{1.3}$$

on (1.2) for $s \in \boldsymbol{R}$. One of our aims in this paper is to establish the unique and smooth solvability of the initial value problem (1.2) with (1.3) in the space where the curvature of the vortex filament $|v_s|$ tends to zero as $s \to \pm\infty$, on the time interval $[0,T]$ with any $T > 0$. In order to achieve it, we first investigate the parabolic regularization

$$\begin{aligned} v_t = v \times v_{ss} + a\left\{v_{sss} + \frac{3}{2}v_{ss} \times (v \times v_s) + \frac{3}{2}v_s \times (v \times v_{ss})\right\} \\ -\epsilon\left\{v_{ssss} + 4(v_s \cdot v_{sss})v + 3|v_{ss}|^2 v\right\} \end{aligned} \tag{1.4}$$

for $\epsilon > 0$. After that, we let $\epsilon \to 0$.

* Received by the editors September 2, 1994; accepted for publication November 21, 1994.

† Department of Mathematics, Keio University, Yokohama 223, Japan.

The other aim is to obtain the unique and smooth solvability of an initial-boundary value problem for (1.2) by the above method. At this time, we treat the case $a = 0$ only.

By the way, (1.1) or (1.2) can be transformed into the Hirota equation (or the nonlinear Schrödinger equation if $a = 0$),

$$i\Psi_t + \Psi_{ss} + \frac{1}{2}|\Psi|^2\Psi - ia\left(\Psi_{sss} + \frac{3}{2}|\Psi|^2\Psi_s\right) = 0 \tag{1.5}$$

for $\Psi = \kappa(s,t)\exp\left(i\int_0^s \tau(s,t)ds + i\eta(t)\right)$, where $\kappa(s,t)$ and $\tau(s,t)$ are the curvature and the torsion of the filament, respectively, and $\eta(t)$ is a real function of t (see [2], [3], [5]). But, as in [8], we should remark that (1.5) is always equivalent to neither (1.1) nor (1.2). In fact, if the filament has a segment where $|x_{ss}|$ vanishes and τ is indefinite, then $\operatorname{Arg}\Psi$ is not well defined even outside there.

We introduce the notation and a result for a linear parabolic system in §2. Then a solution of (1.4) with (1.3) is obtained uniquely on $[0,T]$ with ϵ small enough in §3. In §4, we establish the theorem for (1.2) and (1.3) and obtain a corollary on the vanishing axial flow. In §5, an initial-boundary value problem is discussed.

2. Preliminaries. Let us introduce our notation. m denotes an arbitrary non-negative integer unless we particularly note otherwise. The norms of vector-valued functions in $L^2(\Omega)$ and in the Sobolev space $W_2^m(\Omega)$ are denoted by $\|\cdot\|_\Omega$ and $\|\cdot\|_\Omega^{(m)}$, respectively. Then $\|\cdot\|_\Omega^{(0)} = \|\cdot\|_\Omega$. When $\Omega = \boldsymbol{R}$, we write the norms as simply $\|\cdot\|$ and $\|\cdot\|^{(m)}$. The set of all continuous (resp. once continuously differentiable) functions in a Hilbert space X on a finite time interval $[0,T]$ is denoted by $C(0,T;X)$ (resp. $C^1(0,T;X)$). The class of Hölder-continuous X-valued functions on $[0,T]$ is written as $C^\beta(0,T;X)$, $0 < \beta < 1$. The norm $\langle\cdot\rangle_T$ (resp. $\langle\cdot\rangle_T^{(\beta)}$) represents the supremum (resp. the Hölder norm) over $[0,T]$. Positive constants, denoted by c, c_*, and c_a, change from line to line, but the second is independent of both ϵ and a and the third is monotonically increasing in $|a|$ and independent of ϵ. The operator D is equal to $\partial/\partial s$.

Next, consider a linear equation

$$u_t = -\epsilon u_{ssss} + f(s,t) \tag{2.1}$$

with

$$u(s,0) = u_0(s) \tag{2.2}$$

for $s \in \boldsymbol{R}$. Then we get the following result.

LEMMA 2.1. *If $\epsilon > 0$, $u_0 \in W_2^{4+m}(\boldsymbol{R})$, and $f \in C^\beta(0,T;W_2^m(\boldsymbol{R}))$ for $T > 0$, $0 < \beta < 1$, then there exists a unique solution of (2.1), (2.2) in $C(0,T;W_2^{4+m}(\boldsymbol{R}))\cap C^1(0,T;W_2^m(\boldsymbol{R}))$. Moreover, the following estimate is valid:*

$$\langle\|u\|^{(4+m)}\rangle_T + \langle\|u_t\|^{(m)}\rangle_T \le c\left(\|u_0\|^{(4+m)} + \langle\|f\|^{(m)}\rangle_T^{(\beta)}\right), \tag{2.3}$$

where c is independent of u_0 and f.

This lemma was proved in more general form by the theory of analytic semigroups in [6, Thm. 5.8], [11, Thm. IV.6.E].

3. Solvability of (1.4) with (1.3). Noting that v is a tangential vector and is not square integrable over $\boldsymbol{R}$, we obtain the following.

PROPOSITION 3.1. *Let $\epsilon > 0$, $a \in \boldsymbol{R}$, and $v_{0s} \in W_2^{3+m}(\boldsymbol{R})$. Then on some time interval $[0, T_0]$, $T_0 > 0$, there exists a unique solution v of* (1.4) *with* (1.3) *such that* $(v - v_0) \in C(0, T_0; W_2^{4+m}(\boldsymbol{R})) \cap C^1(0, T_0; W_2^m(\boldsymbol{R}))$.

Proof. Let $u^{(0)} = 0$ and $u^{(n)}$ $(n = 1, 2, \ldots)$ be a solution in Lemma 2.1 on a time interval $[0, t]$ with $0 < \beta < 1/4$, $u_0 = 0$, and

$$\begin{aligned} f = f^{(n-1)} \equiv v^{(n-1)} \times v_{ss}^{(n-1)} &+ a\Big\{ v_{sss}^{(n-1)} + \frac{3}{2} v_{ss}^{(n-1)} \times \left(v^{(n-1)} \times v_s^{(n-1)} \right) \\ &+ \frac{3}{2} v_s^{(n-1)} \times \left(v^{(n-1)} \times v_{ss}^{(n-1)} \right) \Big\} \\ &- \epsilon \Big\{ v_{0ssss} + 4 \left(v_s^{(n-1)} \cdot v_{sss}^{(n-1)} \right) v^{(n-1)} + 3 \left| v_{ss}^{(n-1)} \right|^2 v^{(n-1)} \Big\}, \end{aligned}$$

where $v^{(n-1)} = u^{(n-1)} + v_0$ and $u^{(n-1)} \in C(0, t; W_2^{4+m}(\boldsymbol{R})) \cap C^1(0, t; W_2^m(\boldsymbol{R}))$. Since the multiplicative inequality yields

$$\begin{aligned} &\left\| u^{(n-1)}(\cdot, t') - u^{(n-1)}(\cdot, t'') \right\|^{(3+m)} \\ &\leq c \left(\left\| u^{(n-1)}(\cdot, t') - u^{(n-1)}(\cdot, t'') \right\|^{(4+m)} \right)^{1-\theta} \left(\left\| u^{(n-1)}(\cdot, t') - u^{(n-1)}(\cdot, t'') \right\|^{(m)} \right)^{\theta} \\ &\leq c \left(\langle \| u^{(n-1)} \|^{(4+m)} \rangle_t + \langle \| u_t^{(n-1)} \|^{(m)} \rangle_t \right) |t' - t''|^{\theta} \end{aligned}$$

for $t', t'' \in [0, t]$, and $0 \leq \theta \leq 1/4$, $u^{(n)}$ is well defined for each n. Indeed, it follows from (2.3) and the imbedding theorem that

$$\begin{aligned} A_n &\equiv \langle \| u^{(n)} \|^{(4+m)} \rangle_t + \langle \| u_t^{(n)} \|^{(m)} \rangle_t \\ &\leq c \langle \| f^{(n-1)} \|^{(m)} \rangle_t^{(\beta)} \\ &\leq c \left(1 + \langle \| u^{(n-1)} \|^{(3+m)} \rangle_t^{(\beta)} \right)^2 \langle \| u^{(n-1)} \|^{(3+m)} \rangle_t^{(\beta)} + c \\ &\leq c t^{1/4-\beta} (1 + A_{n-1})^2 A_{n-1} + c. \end{aligned}$$

Here c depends on v_0. Choose $t = T_1$ so small that there exists a constant M independent of n and satisfying $c T_1^{1/4-\beta}(1 + M)^2 M + c \leq M$. Then we obtain $A_n \leq M$ if $A_{n-1} \leq M$. Hence $A_n \leq M$ with n arbitrary.

Setting $w^{(n)} = u^{(n)} - u^{(n-1)}$, we have

$$w_t^{(n+1)} = -\epsilon w_{ssss}^{(n+1)} + f^{(n)} - f^{(n-1)},$$

$$w^{(n+1)}(s, 0) = 0.$$

Again by (2.3) and the fact that A_{n-1} and A_n are bounded by M, we estimate $w^{(n+1)}$ on $[0, t]$, $0 < t \leq T_1$, as

$$\begin{aligned} &\langle \| w^{(n+1)} \|^{(4+m)} \rangle_t + \langle \| w_t^{(n+1)} \|^{(m)} \rangle_t \\ &\leq c t^{1/4-\beta} (1 + M)^2 \left(\langle \| w^{(n)} \|^{(4+m)} \rangle_t + \langle \| w_t^{(n)} \|^{(m)} \rangle_t \right). \end{aligned}$$

If we choose $T_0 \in (0, T_1]$ so small that $cT_0^{1/4-\beta}(1+M)^2 < 1$, then the sequence $\{u^{(n)}\}$ converges to the function $v - v_0$ in the space $C(0, T_0; W_2^{4+m}(\boldsymbol{R})) \cap C^1(0, T_0; W_2^m(\boldsymbol{R}))$, where v is a solution of (1.3) and (1.4).

The uniqueness of v can be easily proved. □

Next, we prove the following lemma, which implies that the length of the vortex filament is conserved.

LEMMA 3.1. *Let v be a solution of* (1.3) *and* (1.4) *such that* $(v - v_0) \in C(0, T; W_2^{4+m}(\boldsymbol{R})) \cap C^1(0, T; W_2^m(\boldsymbol{R}))$, $T > 0$. *Then*

$$|v| = 1 \tag{3.1}$$

holds for any $(s, t) \in \boldsymbol{R} \times [0, T]$.

Proof. Define the function $h(s, t)$ by

$$h(s,t) = |v|^2 - 1$$

for $s \in \boldsymbol{R}$, $0 \le t \le T$. Then we obtain

$$h_s = 2v \cdot v_s,$$

$$h_{ss} = 2(v \cdot v_{ss} + |v_s|^2),$$

$$h_{sss} = 2(v \cdot v_{sss} + 3v_s \cdot v_{ss}),$$

$$h_{ssss} = 2(v \cdot v_{ssss} + 4v_s \cdot v_{sss} + 3|v_{ss}|^2).$$

Using these relations, (1.4) multiplied by v, and (1.3), we have

$$h_t = a\{h_{sss} - 3(v \cdot v_{ss})h_s + 6(v_s \cdot v_{ss})h\} - \epsilon\{h_{ssss} + 8(v_s \cdot v_{sss})h + 6|v_{ss}|^2 h\},$$

$$h(s, 0) = 0.$$

For this linear system, we conclude that $h = 0$ is the only solution because of $\|h\| = 0$ yielded by the estimate $(d/dt)\|h\|^2 \le c\|h\|^2$, where c depends on $\langle \|v - v_0\|^{(4)} \rangle_T$ and $\|v_{0s}\|^{(3)}$. Hence (3.1) follows. □

Utilizing Lemma 3.1, we derive an a priori estimate for (1.4).

LEMMA 3.2. *Let v be as in Lemma* 3.1. *Then there exists a positive constant ϵ_0 depending only on T and $\|v_{0s}\|$ such that v for any $\epsilon \in (0, \epsilon_0]$ satisfies the estimate*

$$\langle \|v - v_0\|^{(4+m)} \rangle_T + \langle \|v_t\|^{(m)} \rangle_T \le c_a, \tag{3.2}$$

where c_a depends only on $\|v_{0s}\|^{(3+m)}$, ϵ_0, T, and $|a|$.

Proof. By the density theorem it is sufficient to prove (3.2) for an infinitely s-differentiable v with a compact support.

From (3.1), we have

$$v \cdot v_s = 0, \quad v \cdot D^n v = -\frac{1}{2} \sum_{k=1}^{n-1} {}_nC_k \, D^k v \cdot D^{n-k} v \quad (n \ge 2). \tag{3.3}$$

It also follows from (3.1) that on the point where $|v_s|$ is nonzero, the vectors v, $v_s/|v_s|$, $v \times v_s/|v_s|$ are the orthonormal ones in $\boldsymbol{R}^3$. Then

$$v_s \times D^n v = v_s \times \left[(v \cdot D^n v)v + \frac{\big\{(v \times v_s) \cdot D^n v\big\} v \times v_s}{|v_s|^2} \right]$$

holds for $n \geq 2$ and leads to

$$v_s \times D^n v = -(v \cdot D^n v) v \times v_s + \big\{(v \times v_s) \cdot D^n v\big\} v. \tag{3.4}$$

Clearly, (3.4) is also valid where $v_s = 0$.

Multiplying (1.4) by v_{ss}, integrating over $\boldsymbol{R}$, and using (3.3), we obtain

$$\begin{aligned} \frac{d}{dt}\|v_s(\cdot,t)\|^2 &= -2\epsilon\left(\|v_{sss}\|^2 + 4\int_{\mathbf{R}} |v_s|^2 v_s \cdot v_{sss} ds + 3\big\||v_s||v_{ss}|\big\|^2\right) \\ &\leq -\epsilon\|v_{sss}\|^2 + \epsilon c_0 \|v_s\|^{10}, \end{aligned}$$

where c_0 is a positive constant yielded by use of the multiplicative inequality and Young's. Let $r(t)$ be a solution of the scalar equation $dr/dt = \epsilon c_0 r^5$ with $r(0) = \|v_{0s}\|^2$. Then we solve it as $r(t) = (\|v_{0s}\|^{-8} - 4\epsilon c_0 t)^{-1/4}$ when $4\epsilon c_0 t < \|v_{0s}\|^{-8}$. Choosing ϵ_0 so small that

$$0 < \epsilon_0 < \left(4c_0 T\|v_{0s}\|^8\right)^{-1}, \tag{3.5}$$

we have

$$\|v_s(\cdot,t)\| \leq r(t)^{1/2} \leq c_* \tag{3.6}$$

on $[0,T]$ for all $\epsilon \in (0,\epsilon_0]$.

Next, by (3.3), (3.4), (3.6), and the multiplicative and Young's inequalities, we obtain

$$\begin{aligned} &\frac{d}{dt}\left(\|v_{ss}(\cdot,t)\|^2 - \frac{5}{4}\big\||v_s(\cdot,t)|^2\big\|^2\right) \\ &= -\int_{\mathbf{R}} (2v_{sss} \cdot v_{st} + 5|v_s|^2 v_s \cdot v_{st}) ds \\ &\leq \int_{\mathbf{R}} 3\left\{|v_s|^2 v_{ss} \cdot (v \times v_s)\right\}_s ds \\ &\quad + a\int_{\mathbf{R}} \left\{|v_s|^2|v_{ss}|^2 + 8(v_s \cdot v_{ss})^2 - 5|v_s|^2 v_s \cdot v_{sss}\right\}_s ds \\ &\quad - 2\epsilon\|v_{ssss}\|^2 + \epsilon c_*\left(\|v_{ssss}\|^{5/3} + \|v_{ssss}\|^{4/3}\right) \\ &\leq c_*. \end{aligned}$$

It yields

$$\begin{aligned} \|v_{ss}(\cdot,t)\|^2 &\leq \|v_{0ss}\|^2 - \frac{5}{4}\big\||v_{0s}|^2\big\|^2 + \frac{5}{4}\big\||v_s(\cdot,t)|^2\big\|^2 + c_* t \\ &\leq c_* + \frac{1}{2}\|v_{ss}(\cdot,t)\|^2 + c_*\|v_s(\cdot,t)\|^6 + c_* t, \end{aligned}$$

from which

$$\langle\|v_{ss}\|\rangle_T \leq c_* \tag{3.7}$$

follows.

In the same way, by boring calculation we can verify

$$\frac{d}{dt}\left(\|v_{sss}\|^2 - \frac{7}{2}\||v_s||v_{ss}|\|^2 - 14\|v_s \cdot v_{ss}\|^2 + \frac{21}{8}\||v_s|^3\|^2\right) \le c_*,$$

which yields

$$\langle\|v_{sss}\|\rangle_T \le c_*. \tag{3.8}$$

Here we made use of the formula

$$|v_s|^2(v \times v_{ss}) \cdot v_{ssss} = (v_s \cdot v_{ss})(v \times v_s) \cdot v_{ssss} - (v_s \cdot v_{ssss})(v \times v_s) \cdot v_{ss},$$

which is proved as (3.4).

Let $j = 4, 5, \ldots, 4+m$. Then, using (3.3), (3.4), and integration by parts, we can derive $(d/dt)\|D^j v\|^2 \le c_a\|D^j v\|^2 + c_a$ if $\langle\|v_s\|^{(j-2)}\rangle_T \le c_a$ is given. This fact, together with (3.6), (3.7), (3.8), and Gronwall's inequality, yields $\langle\|v_s\|^{(3+m)}\rangle_T \le c_a$. Hence we have $\langle\|(v - v_0)_s\|^{(3+m)}\rangle_T \le c_a$. The estimates $\langle\|v_t\|^{(m)}\rangle_T \le c_a$ and $\langle\|v - v_0\|\rangle_T \le c_a$ are easily obtained. □

From Proposition 3.1 and Lemmas 3.1 and 3.2, by the standard continuation argument, we have the following theorem.

THEOREM 3.1. *Let* $T > 0$, $v_{0s} \in W_2^{3+m}(\boldsymbol{R})$, *and* $a \in \boldsymbol{R}$. *Then for each* $\epsilon \in (0, \epsilon_0]$ *with* ϵ_0 *satisfying* (3.5), *there exists a unique solution* v *of* (1.3), (1.4) *such that* $(v - v_0) \in C(0, T; W_2^{4+m}(\boldsymbol{R})) \cap C^1(0, T; W_2^m(\boldsymbol{R}))$, (3.1), *and* (3.2) *hold.*

4. Solvability of (1.2) with (1.3).

Considering the limit $\epsilon \to 0$, we establish the following theorem. Its proof is based mainly on the method in [4, §3].

THEOREM 4.1. *Let* $v_{0s} \in W_2^{3+m}(\boldsymbol{R})$ *and* $a \in \boldsymbol{R}$. *Then there exists a unique solution* v *of* (1.2), (1.3) *such that* (3.1) *is satisfied,* $(v - v_0) \in C(0, T; W_2^{4+m}(\boldsymbol{R})) \cap C^1(0, T; W_2^{1+m}(\boldsymbol{R}))$ *if* $a \neq 0$, *and* $(v - v_0) \in C(0, T; W_2^{4+m}(\boldsymbol{R})) \cap C^1(0, T; W_2^{2+m}(\boldsymbol{R}))$ *if* $a = 0$ *with any* $T > 0$.

Proof. Let $v^\epsilon(s, t)$ be the solution of (1.3), (1.4) in Theorem 3.1. Subtracting (1.4) for $\epsilon = \epsilon''$ from that for $\epsilon = \epsilon'$ $(0 < \epsilon'' < \epsilon' \le \epsilon_0)$ and setting $z = v^{\epsilon'} - v^{\epsilon''}$, we estimate z as

$$\frac{d}{dt}\left(\|z\|^2 + \|z_s\|^2\right) \le c_a\left(\|z\|^2 + \|z_s\|^2\right) + \epsilon' c_a. \tag{4.1}$$

Here we used (3.2) and $v^{\epsilon''} \cdot z_s = v_s^{\epsilon'} \cdot v^{\epsilon''} = -v_s^{\epsilon'} \cdot z$. Since $z(s, 0) = 0$, it follows that $\langle\|z\|^{(1)}\rangle_T \le c_a(\epsilon')^{1/2}$. Thus $v^\epsilon - v_0$ converges strongly to some function $(w - v_0) \in C(0, T; W_2^1(\boldsymbol{R}))$ for $W_2^1(\boldsymbol{R})$-norm and uniformly in $t \in [0, T]$ as $\epsilon \to 0$. From this, we know $|w| = 1$.

On the other hand, (3.2) implies that $(w - v_0) \in L^\infty(0, T; W_2^{4+m}(\boldsymbol{R}))$ and that the weak convergence $(v^\epsilon - v_0) \to (w - v_0)$ holds in $W_2^{4+m}(\boldsymbol{R})$, uniformly for t. Clearly, $w(s, 0) = v_0$ and $\|w(\cdot, t) - v_0\|^{(4+m)} \le c_a$.

Since $G(v^\epsilon) \equiv v^\epsilon \times v_{ss}^\epsilon + (3/2)a\{v_{ss}^\epsilon \times (v^\epsilon \times v_s^\epsilon) + v_s^\epsilon \times (v^\epsilon \times v_{ss}^\epsilon)\}$ is bounded by c_a in $W_2^{2+m}(\boldsymbol{R})$ for every t, its subsequence (if necessary) weakly converges in $W_2^{2+m}(\boldsymbol{R})$ and uniformly for t. Let $\Phi = \Phi(s)$ be an infinitely differentiable vector function with a compact support. Then

$$\left|\left(G(v^\epsilon) - G(w), \Phi\right)^{(2+m)}\right|$$

$$= \left| \int_{\mathbf{R}} \left[v^\epsilon \times v_s^\epsilon - w \times w_s + \frac{3}{2} a \{ v_s^\epsilon \times (v^\epsilon \times v_s^\epsilon) \right.\right.$$
$$\left.\left. - w_s \times (w \times w_s) \} \right] \cdot \left\{ \sum_{j=0}^{2+m} (-1)^{j+1} D^{2j+1} \Phi \right\} ds \right|$$
$$\leq c_a \| v^\epsilon - w \|^{(1)},$$

where $(\cdot, \cdot)^{(m)}$ means the scalar product in $W_2^m(\boldsymbol{R})$, tends to zero uniformly in $t \in [0, T]$ as $\epsilon \to 0$. Hence the weak limit of $G(v^\epsilon)$ is equal to $G(w) \in L^\infty(0, T; W_2^{2+m}(\boldsymbol{R}))$.

Next, integrate (1.4) over $[t', t''] \subset [0, T]$. Then we get

$$v^\epsilon(s, t'') - v^\epsilon(s, t')$$
$$= \int_{t'}^{t''} \left[a v_{sss}^\epsilon + G(v^\epsilon) - \epsilon \{ v_{ssss}^\epsilon + 4(v_s^\epsilon \cdot v_{sss}^\epsilon) v^\epsilon + 3 |v_{ss}^\epsilon|^2 v^\epsilon \} \right] dt.$$

Taking its scalar product in $W_2^{1+m}(\boldsymbol{R})$ (resp. $W_2^{2+m}(\boldsymbol{R})$) with the above Φ if $a \neq 0$ (resp. $a = 0$) and letting $\epsilon \to 0$, we have

$$w(s, t'') - w(s, t') = \int_{t'}^{t''} \big(a w_{sss} + G(w) \big) dt.$$

Hence w is a solution of (1.2) and (1.3).

Let us verify the uniqueness of the solution w. Assume that there exist two solutions w' and w'' for (1.2) with the same data. Then for $z = w' - w''$, we obtain the same estimate with (4.1) but $\epsilon' = 0$. Hence $z = 0$ follows.

Utilizing (3.2), we derive $(d/dt) \big(\| v_{ss}^\epsilon \|^{(2+m)} \big)^2 \leq c_a$ from (1.4) for $\epsilon \in (0, \epsilon_0]$. Therefore, $\| v_{ss}^\epsilon(\cdot, t) \|^{(2+m)} \leq \| v_{0ss} \|^{(2+m)} + c_a t^{1/2}$ follows. By the limit $\epsilon \to 0$ we have $\| w_{ss}(\cdot, t) \|^{(2+m)} \leq \| v_{0ss} \|^{(2+m)} + c_a t^{1/2}$, which leads to $\limsup_{t \to 0} \| w_{ss}(\cdot, t) \|^{(2+m)} \leq \| v_{0ss} \|^{(2+m)} = \| w_{ss}(\cdot, 0) \|^{(2+m)}$. Since w_{ss} is weakly continuous in $W_2^{2+m}(\boldsymbol{R})$, it is strongly continuous in $W_2^{2+m}(\boldsymbol{R})$ at $t = 0$. As in [4], making use of the uniqueness of w and the reversibility of (1.2) in t, we can show $w_{ss} \in C(0, T; W_2^{2+m}(\boldsymbol{R}))$. Then it yields $(w - v_0) \in C(0, T; W_2^{4+m}(\boldsymbol{R})) \cap C^1(0, T; W_2^{1+m}(\boldsymbol{R}))$ if $a \neq 0$, $(w - v_0) \in C(0, T; W_2^{4+m}(\boldsymbol{R})) \cap C^1(0, T; W_2^{2+m}(\boldsymbol{R}))$ if $a = 0$. □

Since we have $c_a \leq c_*$ if $|a| \leq 1$ is assumed, the limit $a \to 0$ can be discussed in the same way as $\epsilon \to 0$.

COROLLARY. *In Theorem* 4.1, *the difference between the solution* v *for* $a \neq 0$ *and that for* $a = 0$ *converges to zero strongly in* $W_2^1(\boldsymbol{R})$ *and weakly in* $W_2^{4+m}(\boldsymbol{R})$, *uniformly in* t *as* $a \to 0$.

It should be noted that our method is also applicable when $a \in \boldsymbol{R}$ and the spatially periodic condition $v(s, t) = v(s + 1, t)$ is imposed.

5. Initial-boundary value problem. In this section, the domain of s is restricted to $J \equiv (-1, 1)$ and a is assumed to be equal to zero. As a boundary condition imposed on (1.2), we take

$$v_s(\pm 1, t) = 0. \tag{5.1}$$

Let V^m be the completion with respect to $\| \cdot \|_J^{(m)}$ of the space where every element g belongs to $C^\infty([-1, 1])$ and satisfies $D^{2j-1} g(\pm 1) = 0$ for $j = 1, 2, \ldots$. Then by the Fourier expansion and Parseval's equality, we obtain the following result.

LEMMA 5.1. *If $\epsilon > 0$, $u_0 \in V^{4+m}$, and $f \in C^{\beta}(0,T;V^m)$ for $T > 0$, $0 < \beta < 1$, then there exists a unique solution of* (2.1), (2.2), *and* $u_s(\pm 1,t) = u_{sss}(\pm 1,t) = 0$ *in* $C(0,T;V^{4+m}) \cap C^1(0,T;V^m)$. *Moreover, the following estimate is valid:*

$$\langle \|u\|_J^{(4+m)} \rangle_T + \langle \|u_t\|_J^{(m)} \rangle_T \le c\left(\|u_0\|_J^{(4+m)} + \langle \|f\|_J^{(m)} \rangle_T^{(\beta)}\right),$$

where c is independent of u_0 and f.

Let $g = (g_1, g_2, g_3) \in V^n$, $n \ge 2$. Then for $k = 1, 2, \dots, n-1$, we verify

$$\|D^k g\|_J^2 = -\int_J D^{k-1}g \cdot D^{k+1}g \, ds \le \|D^{k-1}g\|_J \, \|D^{k+1}g\|_J,$$

$$|D^k g(s)|^2 = \sum_{i=1}^{3} 2 \int_{s_i}^{s} D^k g_i \, D^{k+1} g_i \, ds \le 2\|D^k g\|_J \, \|D^{k+1} g\|_J.$$

Here s_i ($i = 1, 2, 3$) is a point on $[-1, 1]$ satisfying $D^k g_i(s_i) = 0$, whose existence is trivial for an odd k and is obtained from $\int_J D^k g \, ds = 0$ for an even k. From this, we know the validity of the multiplicative inequality for an element in V^n.

Therefore, using Lemma 5.1, we prove the following theorem for (1.4) with (1.3), (5.1), and

$$v_{sss}(\pm 1, t) = 0. \tag{5.2}$$

THEOREM 5.1. *Let $T > 0$, $v_0 \in V^{4+m}$, and $a = 0$. Then for each $\epsilon \in (0, \epsilon_0]$ with $0 < \epsilon_0 < (4c_0 T \|v_{0s}\|_J^8)^{-1}$, there exists a unique solution of* (1.3), (1.4), (5.1), (5.2) *such that $v \in C(0,T;V^{4+m}) \cap C^1(0,T;V^m)$ and* (3.1) *holds. Moreover, $\langle \|v\|_J^{(4+m)} \rangle_T + \langle \|v_t\|_J^{(m)} \rangle_T \le c_*$ is valid, where c_* depends only on v_0, T, and ϵ_0.*

Proof. The proof is divided into two parts. One is to establish the existence of a temporally local solution. It is done as in the proof of Proposition 3.1 because the s-derivatives of any odd order for $v \times v_{ss}$, $(v_s \cdot v_{sss})v$, $|v_{ss}|^2 v$ are equal to zero at $s = \pm 1$ if $D^{2j-1}v(\pm 1, t) = 0$ for $j = 1, 2, \dots$. The other is to derive (3.1) and the a priori estimate in the theorem, and we do so by the method in the proofs of Lemmas 3.1 and 3.2. □

In the same manner as in the proof of Theorem 4.1, we establish the following theorem.

THEOREM 5.2. *Let $v_0 \in V^{4+m}$ and $a = 0$. Then there exists a unique solution of* (1.2), (1.3), (5.1) *such that $v \in C(0,T;V^{4+m}) \cap C^1(0,T;V^{2+m})$ with any $T > 0$ and* (3.1) *is satisfied.*

Here we noted that (5.2) is formally derived from (1.2) with $a = 0$, (1.3), and (5.1), irrespective of the class of v. In fact, (3.1) is formally obtained because of $v \cdot v_t = 0$, and $v_{sss} = (v_{st} - v_s \times v_{ss}) \times v - 3(v_s \cdot v_{ss})v$ follows.

Remark. Our method is also useful to another initial-boundary value problem given by (1.2) with $a = 0$ for $s > 0$, (1.3), and the condition $v_s(0, t) = 0$.

REFERENCES

[1] L. S. Da Rios, *On the motion of an unbounded fluid with a vortex filament of any shape*, Rend. Circ. Mat. Palermo, 22 (1906), pp. 117–135 (in Italian).

[2] Y. Fukumoto and T. Miyazaki, *Three-dimensional distortions of a vortex filament with axial velocity*, J. Fluid Mech., 222 (1991), pp. 369–416.

[3] H. Hasimoto, *A soliton on a vortex filament*, J. Fluid. Mech., 51 (1972), pp. 477–485.

[4] T. Kato, *Nonstationary flows of viscous and ideal fluids in R^3*, J. Funct. Anal., 9 (1972), pp. 296–305.

[5] G. L. Lamb, Jr., *Solitons on moving space curves*, J. Math. Phys., 18 (1977), pp. 1654–1661.

[6] S. Mizohata, *Theory of Partial Differential Equations*, Cambridge University Press, Cambridge, 1973.

[7] D. W. Moore and P. G. Saffman, *The motion of a vortex filament with axial flow*, Philos. Trans. Roy. Soc. London Ser. A, 272 (1972), pp. 403–429.

[8] T. Nishiyama and A. Tani, *Solvability of the localized induction equation for vortex motion*, Comm. Math. Phys., 162 (1994), pp. 433–445.

[9] R. L. Ricca, *Rediscovery of Da Rios equations*, Nature, 352 (1991), pp. 561–562.

[10] ———, *Physical interpretation of certain invariants for vortex filament motion under LIA*, Phys. Fluids A, 4 (1992), pp. 938–944.

[11] R. Showalter, *Hilbert Space Methods for Partial Differential Equations*, Pitman, London, 1977.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 1024–1037, July 1996

SOLUTIONS FOR A TWO-DIMENSIONAL HYPERBOLIC-ELLIPTIC COUPLED SYSTEM*

GARY GANSER†, XIAOPING HU†, AND DENING LI†

Abstract. We study a two-dimensional hyperbolic-elliptic coupled system arising from the study of a gas fluidized bed model. The well-posedness of an initial-boundary value problem is discussed and the short-time existence of classical solutions is obtained.

Key words. two-dimensional hyperbolic-elliptic coupled system, fluidized bed

AMS subject classifications. Primary 35L50, 35L65, 35M05; Secondary 35J25

1. Introduction. The mathematical model describing a gas fluidized bed can be derived from the conservation of mass of each phase and the conservation of momentum. Let α denote the concentration by volume of particles and $\mathbf{v}_p$, $\mathbf{v}_g$ denote the velocities in the particle and gas phases. If the relatively small density of gas is neglected and the intrinsic density ρ of particles is assumed to be constant, then the governing equations are the following system [2, 3]:

$$(1.1)\quad \begin{cases} \partial_t\alpha + \nabla\cdot(\alpha\mathbf{v}_p) = 0, \\ \partial_t(1-\alpha) + \nabla\cdot(1-\alpha)\mathbf{v}_g = 0, \\ \rho\partial_t(\alpha\mathbf{v}_p) + \rho\nabla\cdot(\alpha\mathbf{v}_p\mathbf{v}_p) = -\alpha\nabla p_g - \rho\nu^2\nabla F(\alpha) - \rho\alpha\mathbf{g} + B(\alpha)(\mathbf{v}_g - \mathbf{v}_p), \\ (1-\alpha)\nabla p_g + B(\alpha)(\mathbf{v}_g - \mathbf{v}_p) = 0. \end{cases}$$

Here $\nu > 0$ is the terminal velocity of an isolated falling particle, $\mathbf{g}$ is the constant gravity vector with $g = |\mathbf{g}|$, $B(\alpha)$ is the drag coefficient, and $\rho\nu^2F(\alpha)$ models the pressure difference in the two phases. For simplicity, we will assume in this paper that (see [2, 3])

$$(1.2)\quad \begin{cases} B(\alpha) = \dfrac{g\rho}{\nu}\dfrac{\alpha}{(1-\alpha)^2}, \\ F'(\alpha) > 0,\ \ F''(\alpha) > 0, \ \ \text{for } \alpha > 0. \end{cases}$$

There have been many discussions of system (1.1) for the case of one space variable [2, 3]. In particular, a traveling-wave solution containing admissible shocks was given in [4]. The general case of more space variables has potentially more interesting features, e.g., spherical regions devoid of particles called bubbles [3]. Only numerical studies of (1.1) have been done in the higher-dimensional case. Analytical work such as linear studies [3] has been restricted to approximations of (1.1) or similar systems.

In this paper, for the two-dimensional flow, we introduce the stream function $\phi(x, y)$ such that

$$(1.3)\quad \begin{cases} \alpha v_{p1} + (1-\alpha)v_{g1} = \partial_y\phi, \\ \alpha v_{p2} + (1-\alpha)v_{g2} = -\partial_x\phi; \end{cases}$$

* Received by the editors March 16, 1994; accepted for publication (in revised form) November 21, 1994.

† Department of Mathematics, West Virginia University, Morgantown, WV 26506. The research of the first author was supported in part by Department of Energy grant DE-F605-88ER25067. The research of the second and third authors was supported in part by Office of Naval Research grant N00014-91-J-1291.

therefore, we can express $\mathbf{v}_g$ in terms of $\mathbf{v}_p$ and ϕ:

$$v_{g1} = \frac{\partial_y \phi - v_{p1}}{1-\alpha},$$

$$v_{g2} = \frac{-\partial_x \phi - v_{p2}}{1-\alpha}.$$

The term ∇p_g can be eliminated by the last two equations in (1.1). From the last equation in (1.1) and $\partial_{xy} p_g = \partial_{yx} p_g$, we obtain

$$\left[\frac{B(\phi_y - u)}{(1-\alpha)^2}\right]_y = \left[\frac{B(-\phi_x - v)}{(1-\alpha)^2}\right]_x.$$

Therefore, system (1.1) can be rewritten as the following system for the variables of particle concentration α, particle velocity $\mathbf{v}_p = (u, v)$, and the stream function ϕ:

$$(1.4)\quad \begin{cases} \alpha_t + (\alpha u)_x + (\alpha v)_y = 0, \\ (\alpha u)_t + (\alpha u^2)_x + (\alpha u v)_y = -\nu^2 \partial_x F(\alpha) + \dfrac{g\alpha(\phi_y - u)}{\nu(1-\alpha)^4}, \\ (\alpha v)_t + (\alpha u v)_x + (\alpha v^2)_y = -\nu^2 \partial_y F(\alpha) + \dfrac{g\alpha(-\phi_x - v)}{\nu(1-\alpha)^4} - g\alpha, \\ \triangle \phi = u_y - v_x - \dfrac{G'(\alpha)}{G(\alpha)}\left[\alpha_x(\phi_x + v) + \alpha_y(\phi_y - u)\right], \end{cases}$$

where $G(\alpha) = B(\alpha)(1-\alpha)^{-2}$. The equations in (1.4) form a quasi-linear hyperbolic-elliptic coupled system of the variables (α, u, v, ϕ) in the two-dimensional space.

In this paper, we discuss an initial-boundary value problem of (1.4), derive the well-posedness of the linearized equations, and establish the existence of the solution for the nonlinear problem.

Consider the infinite vertical fluidized bed bounded in $-a \le x \le a$ $(a > 0)$. Assume that at the initial time $t = 0$, the distribution of (α, u, v) is known. Define $\Omega = (-a,\ a) \times R$. We have the initial conditions

$$(1.5)\qquad (\alpha, u, v)(0, x, y) = (\alpha_0, u_0, v_0)(x, y), \qquad (x, y) \in \Omega.$$

At the boundaries $x = \pm a$, we have the physical condition

$$(1.6)\qquad u = 0, \qquad \phi_y = 0.$$

At $y = \pm\infty$, we assume that the fluidized bed is in a homogeneous state; the velocities of gas entering and leaving the fluidized bed are the same constant J:

$$(1.7)\qquad \phi_x = -J.$$

From (1.7), we may assume

$$\phi(t, x, y) = \psi(t, x, y) - Jx;$$

hence (1.7) becomes the condition for the new variable ψ:

$$(1.8)\qquad \psi_x(t, x, \pm\infty) = 0.$$

Obviously, ϕ (and ψ) can be uniquely determined up to a constant; hence without loss of generality, we may assume $\psi(0,x,\pm\infty)=0$. The condition for ϕ on $x=\pm a$ now translates into $\psi(t,\pm a,y)=0$.

To sum up, the initial-boundary problem for (α,u,v,ψ) that we discuss in this paper is as follows:

$$\left\{\begin{aligned} &\alpha_t+(\alpha u)_x+(\alpha v)_y=0,\\ &\alpha u_t+\alpha u u_x+\alpha v u_y=-\nu^2\partial_x F(\alpha)+\frac{g\alpha(\psi_y-u)}{\nu(1-\alpha)^4},\\ &\alpha v_t+\alpha u v_x+\alpha v v_y=-\nu^2\partial_y F(\alpha)+\frac{g\alpha(J-\psi_x-v)}{\nu(1-\alpha)^4}-g\alpha,\\ &\Delta\psi=u_y-v_x-\frac{G'(\alpha)}{G(\alpha)}\left[\alpha_x(\psi_x-J+v)+\alpha_y(\psi_y-u)\right].\end{aligned}\right. \tag{1.9}$$

$$(\alpha,u,v)(0,x,y)=(\alpha_0,u_0,v_0)(x,y),\qquad (x,y)\in\Omega. \tag{1.10}$$

$$u(t,\pm a,y)=0,\qquad \psi(t,\pm a,y)=0. \tag{1.11}$$

$$\psi(t,x,\pm\infty)=0,\qquad 0<t<\infty,\qquad -a\le x\le a. \tag{1.12}$$

Denote $U=(\alpha,u,v)$, $DU=(U,U_x,U_y)$, and $D\psi=(\psi,\psi_x,\psi_y)$; then system (1.9) can be abbreviated as the following:

$$\left\{\begin{aligned} &A_0(U)\partial_t U+A_1(U)\partial_x U+A_2(U)\partial_y U=H_1(U,D\psi),\\ &\Delta\psi+B_1(DU)\partial_x\psi+B_2(DU)\partial_y\psi=H_2(DU),\end{aligned}\right. \tag{1.13}$$

where

$$A_0(U)=\begin{pmatrix}1&0&0\\0&\mu(\alpha)&0\\0,&0&\mu(\alpha)\end{pmatrix},\qquad A_1(U)=\begin{pmatrix}u&\alpha&0\\\alpha&u\mu(\alpha)&0\\0,&0&u\mu(\alpha)\end{pmatrix},$$

$$A_2(U)=\begin{pmatrix}v&0&\alpha\\0&v\mu(\alpha)&0\\\alpha,&0&v\mu(\alpha)\end{pmatrix},\qquad \mu(\alpha)=\frac{\alpha^2}{\nu^2F'(\alpha)},$$

and B_1,B_2,H_1, and H_2 are sufficient smooth functions of their arguments in the relevant domain; the explicit forms of these functions bear no importance in the following discussion.

The first part of system (1.13) is a quasi-linear symmetric hyperbolic system for U and the second is an elliptic equation in (x,y) for ψ with linear principal part. In particular, the variable t appears in the second equation only as a parameter for ψ. We consider the solutions of (1.9)–(1.12) which are constants at $y=\pm\infty$. From (1.9) and (1.11), these constants are not arbitrary. They must satisfy the following conditions:

$$u=0,\qquad v=J-\nu(1-\alpha)^4. \tag{1.14}$$

Because the boundaries $x = \pm a$ are characteristic with respect to the hyperbolic system in (1.13), we will use the partially tangential Sobolev spaces.

For any nonnegative integer m, let $E_m(\Omega)$ denote the spaces

$$E_m = \{w \in L^2(\Omega), (x^2 - a^2)^i \partial_x^j \partial_y^k w \in L^2 \text{ for } 0 \le i \le j, k + 2j - i \le m\}$$

with norm

$$\|w\|_m^2 = \sum_{0 \le i \le j, k+2j-i \le m} \int_\Omega |(x^2 - a^2)^i \partial_x^j \partial_y^k w|^2 \, dx \, dy. \tag{1.15}$$

Denote $E_{m,T}$ as the space

$$E_{m,T} = \{w(t,x,y) : \partial_t^k w \in L^\infty([0,T], E_{m-k}), \text{ for } 0 \le k \le m\} \tag{1.16}$$

with the norm defined as

$$|||w|||_{m,T} = \sup_{0 \le t \le T} \left(\sum_{k=0}^{m} \|\partial_t^k w(t)\|_{m-k}^2 \right)^{\frac{1}{2}}. \tag{1.17}$$

In particular, for positive even numbers $2m$, the spaces E_{2m} coincide with the spaces $\tilde{E}_{2m}$ in [1]. For these spaces, we have similar embedding results.

LEMMA 1.1. *For the spaces E_{2m} and $E_{2m,T}$ and for $j \le m - 1$, we have the following embedding properties:*

$$E_{2m}(\Omega) \subset C^j(\Omega),$$

$$E_{2m,T} \subset C^j([0,T] \times \Omega).$$

The first part of the lemma is a special case of Proposition 4.1.2 of [1]; see also [6, 8]. The second part of the lemma follows directly from the definition in (1.16) for $E_{2m,T}$.

In this paper, we prove the existence and uniqueness of smooth solutions for the coupled system in (1.9) with the boundary and initial conditions in (1.10)–(1.12). To simplify the discussion, we will assume the initial data $U_0 \in C^\infty(\bar{\Omega})$. Obviously, certain compatibility conditions at $(t,x) = (0, \pm a)$ are necessary in order to obtain the existence of classical solutions. The zero-order compatibility condition is simply $U_0(\pm a, y) = 0$. The first-order compatibility condition is obtained by comparing $\partial_t u(0, \pm a, y) = 0$ with the value of $\partial_t v$ determined from the initial condition (1.11) and the equations in (1.13). The higher-order compatibility condition can be derived similarly. Since the value of ψ at $t = 0$ is not given explicitly and has to be found by solving the Dirichlet problem, we will use the following equivalent concept of the compatibility.

The initial data U_0 are called kth-order compatible if there is an approximate solution $\tilde{U}$, $\tilde{\psi}$ such that

(1) $(\tilde{U}, \tilde{\psi})$ satisfies the initial and boundary conditions in (1.10)–(1.12);

(2) $(\tilde{U}, \tilde{\psi})$ satisfies

$$\begin{cases} A_0(\tilde{U})\partial_t \tilde{U} + A_1(\tilde{U})\partial_x \tilde{U} + A_2(\tilde{U})\partial_y \tilde{U} - H_1(\tilde{U}, D\tilde{\psi}) = \tilde{F}_1, \\ \Delta\tilde{\psi} + B_1(D\tilde{U})\partial_x \tilde{\psi} + B_2(D\tilde{U})\partial_y \tilde{\psi} - H_2(D\tilde{U}) = \tilde{F}_2 \end{cases} \tag{1.18}$$

with

$$\partial_t^j \tilde{F}_1(0,x,y) = \partial_t^j \tilde{F}_2(0,x,y) = 0, \quad 0 \le j \le k-1. \tag{1.19}$$

Interested readers are also referred to [1, 4].

Let $\mathbf{G} \subset (0,1) \times R^2$ be an open set such that $\mathbf{G} \subset\subset (0,1) \times R^2$, $U \in E_{2m,T}$, and $D\psi \in E_{2m,T}$ with $U(t,x,y) \in \mathbf{G}$ for all $(t,x,y) \in [0,T] \times \Omega$. Then the main result is the following.

THEOREM 1.1. *Assume the initial value U_0 in* (1.10)–(1.12) *satisfies*

(i) $U_0 \in C^\infty(\bar{\Omega})$, *and*

(ii) U_0 *are 2mth-order compatible,* $\tilde{U} \in E_{2m+1,T}$, *and* $D\tilde{\psi} \in E_{2m,T}$ *with* $\tilde{U}(t,x,y) \in \mathbf{G} \subset\subset (0,1) \times R^2$ *for all* $(t,x,y) \in [0,T] \times \Omega$.

Then there is a $T > 0$ such that the equations in (1.9) *with the initial-boundary conditions in* (1.10)–(1.12) *have a unique classical solution* (U,ψ) *with* $U(t,x,y) \in C^1([0,T] \times \Omega)$ *and* $\psi(t,x,y) \in C\left([0,T], C^2(\Omega)\right)$. *Furthermore,*

$$U, D\psi \in E_{2m,T}. \tag{1.20}$$

The outline of the paper is as follows. In §2, we derive the estimate for the linearized problem of (1.9)–(1.12). Section 3 is devoted to linear iteration to establish the existence of the local solution. Finally, the uniqueness of the solution is proved in §4.

2. Estimate for the linearized problem. It is easy to see that $A_0(U)$ is uniformly positive definite:

$$\sigma I \le A_0(U) \le \sigma^{-1} I.$$

Let (V,φ) denote the perturbations of (U,ψ). Consider the linearization problem of (1.9)–(1.12) for (V,φ):

$$A_0(U)\partial_t V + A_1(U)\partial_x V + A_2(U)\partial_y V = F_1, \tag{2.1a}$$

$$\Delta\varphi + B_1(DU)\partial_x\varphi + B_2(DU)\partial_y\varphi = F_2, \tag{2.1b}$$

$$V(0,x,y) = 0, \tag{2.2}$$

$$\varphi(t,\pm a,y) = 0, \qquad V_2(t,\pm a,y) = 0, \qquad \varphi(t,x,\pm\infty) = 0, \tag{2.3}$$

where $F_1 \in E_{2m,T}$, F_2, $\partial_t F_2 \in E_{2m-1,T}$, and $\partial_t^j F_1(0,x,y) = 0$ for $0 \le j \le 2m-1$. It follows from (2.3) and $\partial_t^j F_1(0,x,y) = 0$ $(0 \le j \le 2m-1)$ that the smooth solutions $V(t,x,y)$ satisfy

$$\partial_t^j V(0,x,y) = 0 \text{ for } 0 \le j \le 2m. \tag{2.4}$$

For the smooth solutions of this linear problem, we have the following estimate.

THEOREM 2.1. *There are constants C_1 and C_2, depending only on the upper bound of $|||U|||_{2m,T}$, such that the smooth solutions (V,φ) of* (2.1)–(2.3) *satisfy the following energy estimate:*

$$|||V|||_{2m,T} \le C_1 T |||F_1|||_{2m,T}, \tag{2.5}$$

$$|||D\varphi|||_{2m,T} \le C_2\left(|||F_2|||_{2m-1,T} + |||\partial_t F_2|||_{2m-1,T}\right), \tag{2.6}$$

where $D\phi = (\varphi, \partial_x\varphi, \partial_y\varphi)$.

Proof. In the following, C denotes a constant depending on $|||U|||_{2m,T}$, but independent of $T << 1$.

The proof of (2.5) consists of two steps. First, we derive the purely tangential estimate. In order to treat the nondegenerate terms, we need to use the first two equations of (2.1a). Since $U(t,x,y) \in \mathbf{G}$, $\mu(\alpha) > 0$, and U satisfies (1.11), there is a constant $\delta > 0$ such that for some $\epsilon > 0$, which depends only on min $\{\alpha^2/\mu(\alpha)\}$ and $\|\partial_x\left(u \pm \alpha^2/\mu(\alpha)\right)\|_\infty$, we have

$$|u \pm \alpha^2/\mu(\alpha)| \ge \delta \text{ for } x \in [-a,\ -a+\epsilon] \cup [a-\epsilon,\ a],\ y \in R,\ t \in [0,\ T]. \tag{2.7}$$

For $j+k+l \le 2m$, denote $V_{j,k,l} = (x^2-a^2)^j \partial_x^j \partial_y^k \partial_t^l V$. From (2.1), we have

$$A_0\partial_t V_{j,k,l} + A_1\partial_x V_{j,k,l} + A_2\partial_y V_{j,k,l} = F_{j,k,l} \tag{2.8}$$

with

$$\begin{aligned} F_{j,k,l} &= (x^2-a^2)^j A_0 \partial_x^j\partial_y^k\partial_t^l(A_0^{-1}F_1) \\ &+ A_0[A_0^{-1}A_1\partial_x V_{j,k,l} - (x^2-a^2)^j\partial_x^j\partial_y^k\partial_t^l(A_0^{-1}A_1\partial_x V)] \\ &+ A_0[A_0^{-1}A_2\partial_y V_{j,k,l} - (x^2-a^2)^j\partial_x^j\partial_y^k\partial_t^l(A_0^{-1}A_2\partial_y V)]. \end{aligned} \tag{2.9}$$

Let $A = (A_0, A_1, A_2)$ with $\nabla \cdot A = (A_0)_t + (A_1)_x + (A_2)_y$. Taking inner product of (2.8) with $V_{j,k,l}$ in $(x,y) \in \Omega$ and integrating by parts, we have

$$\partial_t\langle V_{j,k,l}, A_0 V_{j,k,l}\rangle + \langle A_1 V_{j,k,l}, V_{j,k,l}\rangle_{x=-a}^{x=a} = \langle V_{j,k,l}, (\nabla\cdot A)\, V_{j,k,l}\rangle + 2\langle F_{j,k,l}, V_{j,k,l}\rangle.$$

The boundary terms at $x = \pm a$ are

$$\langle V_{j,k,l}, A_1 V_{j,k,l}\rangle = 2(x^2-a^2)^{2j}\alpha\partial_x^j\partial_y^k\partial_t^l V_1 \partial_x^j\partial_y^k\partial_t^l V_2.$$

They are all zeroes because for $j = 0$, it is zero by the boundary condition in (2.3) for V, and for $j > 0$, it is zero since $(x^2 - a^2) = 0$. Applying Gronwall's inequality and (2.4), we derive for $t \in [0,\ T]$

$$\|V_{j,k,l}(t)\|_0 \le \sigma\langle V_{j,k,l}, A_0 V_{j,k,l}\rangle^{\frac{1}{2}} \le C\int_0^t \|F_{j,k,l}(\tau)\|_0\, d\tau. \tag{2.10}$$

The $F_{j,k,l}$ terms in (2.9) are estimated as follows. Obviously, for the first and the third terms of (2.9), we have

$$\|(x^2-a^2)^j A_0\partial_x^j\partial_y^k\partial_t^l(A_0^{-1}F_1)\|_0 \le C|||F_1|||_{2m,T}$$

and

$$\begin{aligned} &\|A_0[A_0^{-1}A_2\partial_y V_{j,k,l} - (x^2-a^2)^j\partial_x^j\partial_y^k\partial_t^l(A_0^{-1}A_2\partial_y V)]\|_0 \\ &\le C\sum_{j+k+l\le 2m}\|(x^2-a^2)^j\partial_x^j\partial_y^k\partial_t^l V(t)\|_0. \end{aligned}$$

It remains to estimate

$$A_0[A_0^{-1}A_1\partial_x V_{j,k,l} - (x^2-a^2)^j\partial_x^j\partial_y^k\partial_t^l(A_0^{-1}A_1\partial_x V)]. \tag{2.11}$$

In the interior domain away from the boundaries $x = \pm a$, the term in (2.11) is estimated in the same way as we estimate the third term of (2.9). We need only to consider the estimate of (2.11) near $x = \pm a$. (2.11) is the sum of two parts. The first part consists of the terms such that the order of x-derivatives applied on V is at most equal to the order of the factor $(x^2 - a^2)$. This part can be controlled by

$$\sum_{j+k+l\le 2m} \|(x^2 - a^2)^j \partial_x^j \partial_y^k \partial_t^l V\|_0. \tag{2.12}$$

The second part consists of the terms of the form

$$(x^2 - a^2)^j \partial_x^{j+1} \partial_y^{k_1} \partial_t^{l_1} V \quad \text{with} \quad k_1 + l_1 < 2m - j. \tag{2.13}$$

Since u is the only nonzero element in the third row of $A_0^{-1}A_1$, the third component in (2.13) is controlled by (2.12). For the first two components in (2.13), because of (2.7) we can express V_1 and V_2 in terms of a combination of $\partial_t V$, $\partial_y V$, and F_1. Consequently, the first two components of $(x^2 - a^2)^j \partial_x^{j+1} \partial_y^{k_1} \partial_t^{l_1} V$ can be expressed as a linear combination of

$$(x^2-a^2)^j \partial_x^j \partial_y^{k_1+1} \partial_t^{l_1} V, \qquad (x^2-a^2)^j \partial_x^j \partial_y^{k_1} \partial_t^{l_1+1} V, \qquad (x^2-a^2)^j \partial_x^j \partial_y^{k_1} \partial_t^{l_1} F_1, \tag{2.14}$$

and the later can be controlled by

$$\|(x^2 - a^2)^j \partial_x^j \partial_y^k \partial_t^l V\|_0 \quad \text{and} \quad \|(x^2 - a^2)^j \partial_x^j \partial_y^k \partial_t^l F_1\|_0 \tag{2.15}$$

To summarize, we obtain

$$\|F_{j,k,l}(t)\|_0 \le C \left[|||F_1|||_{2m,T} + \sum_{j+k+l\le 2m} \|(x^2 - a^2)^j \partial_x^j \partial_y^k \partial_t^l V\|_0 \right]. \tag{2.16}$$

Taking summation in (2.10) for all $j + k + l \le 2m$, we have

$$\sum_{j+k+l\le 2m} \|V_{j,k,l}(t)\|_0 \le C \int_0^t \left[\sum_{j+k+l\le 2m} \|V_{j,k,l}(\tau)\|_0 + |||F_1|||_{2m,T} \right] d\tau. \tag{2.17}$$

From (2.4) and applying Gronwall's inequality, we obtain for $t \in [0, T]$

$$\sum_{j+k+l\le 2m} \|V_{j,k,l}(t)\|_0 \le CT|||F_1|||_{2m,T}. \tag{2.18}$$

Next, we prove the nontangential part in (2.5). We prove the following by induction on $0 \le r \le m$:

$$\sum_{0\le j-i\le r, k+l+2j-i\le 2m} \|(x^2 - a^2)^i \partial_x^j \partial_y^k \partial_t^l V(t)\|_0 \le CT|||F_1|||_{2m,T}. \tag{2.19}$$

The case $r = 0$ is simply (2.18). Assume (2.19) is true for $0, 1, \ldots, r - 1$. The estimate in (2.19) is trivial in the interior domain; hence we need only to consider the proof of (2.19) near the boundary. At the boundaries $x = \pm a$, for $2r + i + k + l \le 2m$, consider the first two components of $(x^2 - a^2)^i \partial_x^{i+n} \partial_y^k \partial_t^l V$. By applying the operator

$(x^2-a^2)^i\partial_x^{i+r-1}\partial_y^k\partial_t^l$ to the first two equations in (2.1a), these first two components can be expressed as a linear combination of the following terms:

$$(x^2-a^2)^i\partial_x^{i+r-1}\partial_y^{k+1}\partial_t^l V, \qquad (x^2-a^2)^i\partial_x^{i+r-1}\partial_y^k\partial_t^{l+1} V, \qquad (x^2-a^2)^i\partial_x^{i+r-1}\partial_y^k\partial_t^l F_1.$$

By induction, the first two terms can be controlled by $T|||F_1|||_{2m,T}$ and the third term can be controlled by $T\|(x^2-a^2)^i\partial_x^{i+n-1}\partial_y^k\partial_t^{l+1}F_1\|_0$, which in turn is controlled by $T|||F_1|||_{2m,T}$.

Furthermore, because $2r+i+k+l \le 2m$, we can also apply the operators $(x^2-a^2)^i\partial_x^{i+r-1}\partial_y^{k+1}\partial_t^l$ or $(x^2-a^2)^i\partial_x^{i+r-1}\partial_y^k\partial_t^{l+1}$ to the first two equations in (2.1a) to obtain the estimate for

$$\|(x^2-a^2)^i\partial_x^{i+r}\partial_y^{k+1}\partial_t^l V_{1,2}\|_0 + \|(x^2-a^2)^i\partial_x^{i+r}\partial_y^k\partial_t^{l+1} V_{1,2}\|_0. \tag{2.20}$$

For the third component, $(x^2-a^2)^i\partial_x^{i+r}\partial_y^k\partial_t^l V_3$, we consider the third equation of (2.1a) as an equation for the variable V_3. Applying the operator $(x^2-a^2)^i\partial_x^{i+n}\partial_y^k\partial_t^l$ to this equation, we obtain an equation for the variable $(x^2-a^2)^i\partial_x^{i+r}\partial_y^k\partial_t^l V_3$:

$$(\partial_t + u\partial_x + v\partial_y)(x^2-a^2)^i\partial_x^{i+r}\partial_y^k\partial_t^l V_3 = G. \tag{2.21}$$

In addition to the standard right-side term, $(x^2-a^2)^i\partial_x^{i+r}\partial_y^k\partial_t^l F_1$, the term G in (2.21) contains other two parts. The first part comes from the commutator

$$[(x^2-a^2)^i\partial_x^{i+r}\partial_y^k\partial_t^l, u\mu^{-1}(\alpha)\partial_x]V_3,$$

which is the linear combination of the following terms:

$$(x^2-a^2)^i\partial_x^{i+r}\partial_y^k\partial_t^l V_3, \quad (x^2-a^2)^{i+1}\partial_x^{i+r+1}\partial_y^{k-1}\partial_t^l V_3, \quad (x^2-a^2)^{i+1}\partial_x^{i+r+1}\partial_y^k\partial_t^{l-1} V_3. \tag{2.22}$$

The first term in (2.22) is the standard lower-order term for (2.21). The other two terms in (2.22) can be treated by further induction on $k+l$ by noticing that these terms do not appear for $k+l=0$.

The second part in G comes from the commutators

$$[(x^2-a^2)^i\partial_x^{i+r}\partial_y^k\partial_t^l, v\mu^{-1}(\alpha)\partial_y]V_3 \tag{2.23}$$

and

$$(x^2-a^2)^i\partial_x^{i+r}\partial_y^k\partial_t^l[\alpha\mu^{-1}(\alpha)\partial_y V_1]. \tag{2.24}$$

The terms in (2.23) are either standard lower-order terms relative to (2.21) or terms where x-derivatives are one order lower than $i+r$. The latter can be controlled by induction assumption. The term in (2.24) is obviously controlled by (2.20).

Since $u=0$ at the boundaries $x=\pm a$, no boundary condition is required to obtain the energy estimate for equation (2.21) from integration by parts. Finally, applying Gronwall's inequality, we obtain

$$\begin{aligned} &\|(x^2-a^2)^i\partial_x^{i+r}\partial_y^k\partial_t^l V_3\|_0 \\ &\quad \le CT\left\{\sum_{k_1+l_1\le 2m-i-2r+1}\|(x^2-a^2)^i\partial_x^{i+r}\partial_y^{k_1}\partial_t^{l_1}V_1\|_0 + |||F_1|||_{2m,T}\right\}. \end{aligned} \tag{2.25}$$

This completes the proof of induction on r and consequently (2.5).

We now turn to the proof of (2.6). Applying the operator $(x^2-a^2)^i\partial_x^j\partial_y^k\partial_t^l$ to (2.1b) for $0 \le i \le j$, we have

$$\begin{aligned}\Delta(x^2-a^2)^i\partial_x^j\partial_y^k\partial_t^l\varphi + B_1(DU)\partial_x(x^2-a^2)^i\partial_x^j\partial_y^k\partial_t^l\varphi \\ + B_2(DU)\partial_y(x^2-a^2)^i\partial_x^j\partial_y^k\partial_t^l\varphi = \tilde{F}_2,\end{aligned} \tag{2.26}$$

where

$$\begin{aligned}\tilde{F}_2 = (x^2-a^2)^i\partial_x^j\partial_y^k\partial_t^l F_2 + \left[B_1(DU)\partial_x, (x^2-a^2)^i\partial_x^j\partial_y^k\partial_t^l\right]\varphi \\ + (x^2-a^2)^i\left[B_2(DU)\partial_y, \partial_x^j\partial_y^k\partial_t^l\right]\varphi + \left[\partial_x^2, (x^2-a^2)^i\partial_x^j\partial_y^k\partial_t^l\right]\varphi.\end{aligned} \tag{2.27}$$

For the special case $(j,k,l) = (0,0,l)$ with $l < 2m$, from (2.26) and the fact that $\partial_t^l\varphi = 0$ on the boundaries, we can use the classical result for second-order elliptic Dirichlet problems [5] to obtain

$$\|\partial_t^l D\varphi\|_{H^2} \le C\|\tilde{F}_2\|_0 \le C\left\{\sum_{l_1<l}\|\partial_t^{l_1}D\varphi\|_0 + \|\partial_t^l F_2\|_0\right\},$$

which implies that

$$\sum_{l<2m}\|\partial_t^l D\varphi\|_0 \le C|||F_2|||_{2m-1,T}. \tag{2.28}$$

For $(j,k,l) = (0,0,2m)$, we have from (2.26)–(2.28)

$$\|\partial_t^{2m}D\varphi\|_0 \le \|\partial_t^{2m}\varphi\|_{H^1} \le C\|\tilde{F}_2\|_{H^{-1}} \le C\left\{\|\partial_t^{2m}F_2\|_{H^{-1}} + |||F_2|||_{2m-1,T}\right\}. \tag{2.29}$$

From (2.29) and (2.30), it follows that

$$\sum_{l\le 2m}\|\partial_t^l D\varphi\|_0 \le C\left\{\|\partial_t^{2m}F_2\|_{H^{-1}} + |||F_2|||_{2m-1,T}\right\}. \tag{2.30}$$

Next, we estimate $(x^2-a^2)^i\partial_x^j\partial_y^k\partial_t^l D\varphi$. By directly differentiating equation (2.1b), it is easy to see that the estimate of $(x^2-a^2)^i\partial_x^j\partial_y^k\partial_t^l D\varphi$ can be reduced to the estimate of $(x^2-a^2)^i\partial_x^{j-1}\partial_y^{k+1}\partial_t^l D\varphi$. Therefore, we need only to consider the special case $j=0$. We will perform induction on $k+l$. By (2.30), we will always assume $k>0$.

Consider (2.26) with $i=j=0$ and $k+l=r$. Because of the boundary condition

$$\partial_y^k\partial_t^l\varphi|_{\partial\Omega} = 0, \tag{2.31}$$

similarly as in deriving (2.28), we obtain

$$\sum_{k+l=r+1,k>0}\|\partial_y^k\partial_t^l D\varphi\|_0 \le C\|\tilde{F}_2\|_0 \le C\sum_{k+l\le r}\left(\|\partial_y^k\partial_t^l D\varphi\|_0 + \|\partial_y^k\partial_t^l F_2\|_0\right).$$

By induction assumption, we have

$$\sum_{k+l\le r}\|\partial_y^k\partial_t^l D\varphi\|_0 \le C\left\{\|\partial_t^{2m}F_2\|_{H^{-1}} + |||F_2|||_{2m-1,T}\right\}. \tag{2.32}$$

Therefore,

$$\sum_{i\le j,k+l+2j-i\le 2m}\|(x^2-a^2)^i\partial_x^j\partial_y^k\partial_t^l D\varphi\|_0 \le C\left\{\|\partial_t^{2m}F_2\|_{H^{-1}} + |||F_2|||_{2m-1,T}\right\}. \tag{2.33}$$

This gives (2.6).

Finally, the dependency of the constants on $|||U|||_{2m,T}$ comes from the Nirenberg inequality and the Banach algebra property of the space $E_{2m,T}$ for $m \geq 1$; see [1, 6]. This completes the proof of Theorem 2.1.

Remark 1. In Theorem 2.1, the term $|||\partial_t F_2|||_{2m-1,T}$ is included in (2.6) to account for the derivative in the t direction because the solution of Laplace equation has no regularizing effect on the t-derivative. However, if F_2 has the special form $F_2 = DG_2$ in (2.1b), then (2.6) can be replaced simply by

$$|||D\varphi|||_{2m,T} \leq C_2|||G_2|||_{2m,T}. \tag{2.34}$$

This is because, in this case, the regularizing effect of inverse Laplacian cancels the one-order space derivative on G_2 and the two sides of (2.34) have the same regularity in the t direction. We will use this fact in the proof of Theorem 4.1.

3. Linear iteration. In this section, we will use linear iteration to prove the existence of classical solutions for the nonlinear problem in (1.10)–(1.12) in Theorem 1.1.

Assume the initial data $U_0 \in C^\infty(\bar{\Omega})$ are $2m$th-order compatible, and $(\tilde{U}, \tilde{\psi})$ are the approximate solutions such that $\tilde{U} \in E_{2m+1,T}$ and $D\tilde{\psi} \in E_{2m,T}$ with $\tilde{U}(t,x,y) \in \mathbf{G}_1 \subset\subset \mathbf{G}_2 \subset\subset \mathbf{G} = (0,1) \times R^2$ for all $(t,x,y) \in [0, T] \times \Omega$. From the embedding property of the space E_{2m}, there is an $\epsilon_0 > 0$ so small such that if

$$|||U - \tilde{U}|||_{2m,T} \leq \epsilon_0, \tag{3.1}$$

then we have $U(t,x,y) \in \bar{\mathbf{G}}_2$.

Our solution is constructed through the following iteration scheme:

$$\begin{cases} A_0(U^{(k)})\partial_t U^{(k+1)} + A_1(U^{(k)})\partial_x U^{(k+1)} + A_2(U^{(k)})\partial_y U^{(k+1)} = H_1(U^{(k)}, D\psi^{(k)}), \\ \Delta\psi^{(k+1)} + B_1(DU^{(k)})\partial_x\psi^{(k+1)} + B_2(DU^{(k)})\partial_y\psi^{(k+1)} = H_2(DU^{(k)}), \end{cases} \tag{3.2}$$

$$U^{(k+1)}(0,x,y) = U_0(x,y), \tag{3.3}$$

$$u^{(k+1)}(t, \pm a, y) = 0, \qquad \psi^{(k+1)}(t, \pm a, y) = 0, \qquad \psi^{(k+1)}(t, x, \pm\infty) = 0 \tag{3.4}$$

for $k = 0, 1, 2, \dots$ with

$$U^{(0)}(t,x,y) = \tilde{U}(t,x,y), \qquad \psi^{(0)}(t,x,y) = \tilde{\psi}(t,x,y).$$

Here, $U^{(k+1)}, \psi^{(k+1)}$ can be solved in any interval $[0, T_k]$ in which

$$U^{(k)}(t,x,y) \in \mathbf{G}_2 \quad \text{for } (t,x,y) \in [0, T_k] \times \Omega$$

and

$$U^{(k)}(t,x,y) \in C^1([0, T_k] \times \Omega), \qquad \psi^{(k)}(t,x,y) \in C\big([0, T_k], C^1(\Omega)\big).$$

For this sequence of $(U^{(k)}, \psi^{(k)})$, we also have the following result.

LEMMA 3.1. *The sequence of $(U^{(k)}, \psi^{(k)})$ has zero traces at $t = 0$ up to order $2m$:*

$$\begin{aligned} &\partial_t^j \left(U^{(k+1)} - \tilde{U}\right)(0,x,y) = 0, \quad 0 \leq j \leq 2m, \\ &\partial_t^j \left(\psi^{(k+1)} - \tilde{\psi}\right)(0,x,y) = 0, \quad 0 \leq j \leq 2m-1 \end{aligned} \tag{3.5}$$

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

Proof. It follows from (1.18) and (3.2) that for $k = 0, 1, 2, 3, \ldots$

$$\tag{3.6} \begin{cases} A_0(U^{(k)})\partial_t \left[U^{(k+1)} - U^{(0)}\right] + A_1(U^{(k)})\partial_x \left[U^{(k+1)} - U^{(0)}\right] \\ \quad + A_2(U^{(k)})\partial_y \left[U^{(k+1)} - U^{(0)}\right] \\ = H_1(U^{(k)}, D\psi^{(k)}) - H_1(U^{(0)}, D\psi^{(0)}) - \tilde{F}_1 + H_3(U^{(0)}, U^{(k)}), \\ \Delta\left[\psi^{(k+1)} - \psi^{(0)}\right] + B_1(DU^{(k)})\partial_x \left[\psi^{(k+1)} - \psi^{(0)}\right] \\ \quad + B_2(DU^{(k)})\partial_y \left[\psi^{(k+1)} - \psi^{(0)}\right] \\ = H_2(DU^{(k)}) - H_2(DU^{(0)}) - \tilde{F}_2 + H_4(U^{(0)}, U^{(k)}, \psi^{(0)}), \end{cases}$$

where

$$\tag{3.7} \begin{cases} H_3 = \left[A_0(U^{(0)}) - A_0(U^{(k)})\right]\partial_t U^{(0)} + \left[A_1(U^{(0)}) - A_1(U^{(k)})\right]\partial_x U^{(0)} \\ \qquad + \left[A_2(U^{(0)}) - A_2(U^{(k)})\right]\partial_y U^{(0)} - \tilde{F}_1, \\ H_4 = \left[B_1(DU^{(0)}) - B_1(DU^{(k)})\right]\partial_x \psi^{(0)} + \left[B_2(DU^{(0)}) - B_2(DU^{(k)})\right]\partial_y \psi^{(0)} - \tilde{F}_2. \end{cases}$$

For $k = 0$, (3.6) becomes simply

$$\tag{3.8} \begin{cases} A_0(U^{(0)})\partial_t \left[U^{(1)} - U^{(0)}\right] + A_1(U^{(0)})\partial_x \left[U^{(1)} - U^{(0)}\right] \\ \quad + A_2(U^{(0)})\partial_y \left[U^{(1)} - U^{(0)}\right] = -\tilde{F}_1, \\ \Delta\left[\psi^{(1)} - \psi^{(0)}\right] + B_1(DU^{(0)})\partial_x \left[\psi^{(1)} - \psi^{(0)}\right] \\ \quad + B_2(DU^{(0)})\partial_y \left[\psi^{(1)} - \psi^{(0)}\right] = -\tilde{F}_2. \end{cases}$$

Since $\partial_t^j \tilde{F}_2(0, x, y) = 0$ $(0 \le j \le 2m - 1)$, taking the derivative with respect to t in the second equation of (3.8), we have

$$\tag{3.9} \partial_t^j \left[\psi^{(1)} - \psi^{(0)}\right](0, x, y) = 0 \quad (0 \le j \le 2m - 1).$$

Since both $U^{(1)}$ and $U^{(0)}$ satisfy (1.10), we have $\left(U^{(1)} - U^{(0)}\right)(0, x, y) = 0$. Therefore, we derive inductively on j from (3.6), (3.9), and $\partial_t^j \tilde{F}_1 = 0\,(0 \le j \le 2m - 1)$ that

$$\partial_t^j (U^{(1)} - U^{(0)})(0, x, y) = 0 \quad \text{for } 0 \le j \le 2m.$$

This proves (3.5) for $k = 0$. The cases for $k > 0$ can be easily proved inductively by noticing that the terms H_3 and H_4 in (3.7) have zero traces up to order $2m - 1$ by induction assumption.

The following lemma establishes the existence of iteration sequence in a common time interval.

LEMMA 3.2. *There are constants $T_0 > 0$ and $\ell > 0$ such that the iteration sequence $\{U^{(k)}, \psi^{(k)}\}$ satisfies*

$$\tag{3.10} |||U^{(k)} - \tilde{U}|||_{2m,T_0} \le \epsilon_0, \qquad |||D(\psi^{(k)} - \tilde{\psi})|||_{2m,T_0} \le \ell$$

for all $k \in \mathbf{Z}^+$. The constants T_0 and ℓ depend only on the approximate solution $(\tilde{U}, \tilde{\psi})$ (which in turn can be determined from the initial data U_0) up to order $|||\tilde{U}|||_{2m+1,T}$, $|||D\tilde{\psi}|||_{2m,T}$.

Proof. This can be shown inductively on k. For simplicity, we denote $U^{(k)}$ by U, $\psi^{(k)}$ by ψ, $U^{(k+1)} - U^{(0)}$ by V, and $\psi^{(k+1)} - \psi^{(0)}$ by Ψ. We apply (2.5) and (2.6) to system (3.6). Noticing that the right sides of (3.6) have zero traces at $T = 0$ up to order $2m - 1$ and that the constants C_1 and C_2 in (2.5) and (2.6) depend only on $|||U|||_{2m,T}$, we obtain

$$|||V|||_{2m,T} \le CT, \qquad |||D\Psi|||_{2m,T} \le C', \tag{3.11}$$

where the constants C and C' depend only on

$$\begin{aligned} C &= C(|||\tilde{F}_1||||_{2m,T}, |||U - U^{(0)}|||_{2m,T}, |||D(\psi - \psi^{(0)})|||_{2m,T}), \\ C' &= C'(|||\tilde{F}_2||||_{2m-1,T}, |||U - U^{(0)}|||_{2m,T}, |||D(\psi - \psi^{(0)})|||_{2m,T}). \end{aligned} \tag{3.12}$$

$|||\tilde{F}_1||||_{2m,T}$ and $|||\tilde{F}_2||||_{2m-1,T}$ depend only on $|||\tilde{U}|||_{2m+1,T}$ and $|||D\tilde{\psi}|||_{2m,T}$. By induction assumption, $|||U|||_{2m,T}$ and $|||D\psi|||_{2m,T}$ are uniformly bounded by (3.10). Therefore, T_0 can be easily decided from (3.11) and depends only on $|||\tilde{U}|||_{2m+1,T}$ and $|||D\tilde{\psi}|||_{2m,T}$. This completes the proof of the lemma.

Next lemma establishes the convergence of the solution sequence above.

LEMMA 3.3. *There are positive constants* T_1, $0 < T_1 \le T_0$, ρ_1, ρ_2, $\rho_1 + \rho_2 < 1$, *and* $\tilde{C}$ *such that the functions* $\{U^{(k)}, \psi^{(k)}\}$ *obtained from* (3.1)–(3.4) *satisfy*

$$|||U^{(k+2)} - U^{(k+1)}|||_{2,T_1} \le \rho_1 |||U^{(k+1)} - U^{(k)}|||_{2,T_1} + \rho_2 |||U^{(k)} - U^{(k-1)}|||_{2,T_1} \tag{3.13}$$

and

$$|||D\psi^{(k+1)} - D\psi^{(k)}|||_{2,T_1} \le \tilde{C} |||U^{(k)} - U^{(k-1)}|||_{2,T_1}, \tag{3.14}$$

$k = 1, 2, 3, \ldots$.

Proof. Denote $V^{(k)} = U^{(k+1)} - U^{(k)}$ and $\Psi^{(k)} = \psi^{(k+1)} - \psi^{(k)}$, $k = 0, 1, 2, \ldots$. It follows from (3.2) that

$$\begin{aligned} &A_0(U^{(k)})\partial_t V^{(k)} + A_1(U^{(k)})\partial_x V^{(k)} + A_2(U^{(k)})\partial_y V^{(k)} \\ &\quad = H_1(U^{(k)}, D\psi^{(k)}) - H_1(U^{(k-1)}, D\psi^{(k-1)}) - \left(A_0(U^{(k)}) - A_0(U^{(k-1)})\right) \partial_t U^{(k)} \\ &\quad\quad - \left(A_1(U^{(k)}) - A_1(U^{(k-1)})\right) \partial_x U^{(k)} - \left(A_2(U^{(k)}) - A_2(U^{(k-1)})\right) \partial_y U^{(k)}, \end{aligned} \tag{3.15}$$

$$\begin{aligned} &\Delta\Psi^{(k)} + B_1(DU^{(k)})\partial_x \Psi^{(k)} + B_2(DU^{(k)})\partial_y \Psi^{(k)} \\ &\quad = H_2(DU^{(k)}) - H_2(DU^{(k-1)}) - \left(B_1(DU^{(k)}) - B_1(DU^{(k-1)})\right) \partial_x U^{(k)} \\ &\quad\quad - \left(B_2(DU^{(k)}) - B_2(DU^{(k-1)})\right) \partial_y U^{(k)}. \end{aligned} \tag{3.16}$$

By Lemma 3.2 and the Banach algebra property of the space $E_{m,T}$, there exists $C = C(\bar{\mathbf{G}}_2)$ such that, for any $0 < T \le T_0$,

$$\begin{aligned} &|||H_1(U^{(k)}, D\psi^{(k)}) - H_1(U^{(k-1)}, D\psi^{(k-1)})|||_{2,T} \\ &\quad \le C(\bar{\mathbf{G}}_2) \left(|||V^{(k-1)}|||_{2,T} + |||D\Psi^{(k-1)}|||_{2,T}\right), \end{aligned}$$

$$\begin{aligned} &|||\left(A_0(U^{(k)}) - A_0(U^{(k-1)})\right) \partial_t U^{(k)}|||_{2,T} + |||\left(A_1(U^{(k)}) - A_1(U^{(k-1)})\right) \partial_x U^{(k)}|||_{2,T} \\ &\quad + |||\left(A_2(U^{(k)}) - A_2(U^{(k-1)})\right) \partial_y U^{(k)}|||_{2,T} \le C(\bar{\mathbf{G}}_2) |||V^{(k-1)}|||_{2,T}, \end{aligned}$$

$$\begin{aligned} &|||D(H_2(DU^{(k)}) - H_2(DU^{(k-1)}))|||_{0,T} \\ &\quad + |||D\left((B_1(DU^{(k)}) - B_1(DU^{(k-1)}))\partial_x U^{(k)}\right)|||_{0,T} \\ &\quad + |||D\left((B_2(DU^{(k)}) - B_2(DU^{(k-1)}))\partial_y U^{(k)}\right)|||_{0,T} \le C(\bar{\mathbf{G}}_2) |||V^{(k-1)}|||_{2,T}. \end{aligned}$$

Thus, applying (2.5) and (2.6) to (3.15) and (3.16), we have

$$|||D\psi^{(k+1)}(t) - D\psi^{(k)}|||_{2,T} \leq C_2(\bar{\mathbf{G}}_2)|||V^{(k-1)}|||_{2,T}, \tag{3.17}$$

$$\begin{aligned} &|||U^{(k+1)}(t) - U^{(k)}(t)|||_{2,T} \\ &\leq C_1(\bar{\mathbf{G}}_2)T\left(|||V^{(k-1)}|||_{2,T} + |||D\Psi^{(k-1)}|||_{2,T}\right) \\ &\leq C_1(\bar{\mathbf{G}}_2)T|||V^{(k-1)}|||_{2,T} + C_1(\bar{\mathbf{G}}_2)C_2(\bar{\mathbf{G}}_2)T|||V^{(k-2)}|||_{2,T} \end{aligned} \tag{3.18}$$

Therefore, in order to prove Lemma 3.3, we need only to choose $\tilde{C} = C_2(\bar{\mathbf{G}}_2)$, $\rho_2 = C_1(\bar{\mathbf{G}}_2)C_2(\bar{\mathbf{G}}_2)T_1$, and $\rho_1 = C_1(\bar{\mathbf{G}}_2)T_1$ with $T_1 < \left(C_1(\bar{\mathbf{G}}_2) + C_1(\bar{\mathbf{G}}_2)C_2(\bar{\mathbf{G}}_2)\right)^{-1}$.

Now, we are ready to prove the existence part of Theorem 1.1.

Proof of Theorem 1.1. It follows from (3.13) and (3.14) that

$$\sum_{k=1}^{\infty} |||U^{(k+1)} - U^{(k)}|||_{2,T_1} < \infty$$

and

$$\sum_{k=1}^{\infty} |||D\psi^{(k+1)} - D\psi^{(k)}|||_{2,T_1} < \infty.$$

Therefore, there exist $U \in E_{2,T_1}$ and ψ with $D\psi \in E_{2,T_1}$ such that

$$\lim_{k\to\infty} \left(|||U^{(k)} - U|||_{2,T_1} + |||D\psi^{(k)} - D\psi|||_{2,T_1}\right) = 0. \tag{3.19}$$

The $E_{m,T}$ space interpolation inequalities [1] implies that there exists a constant $C = C(m)$ such that for any $s \in [2, 2m]$,

$$||V||_s \leq C||V||_2^{\frac{2m-s}{2m-2}} ||V||_{2m}^{\frac{s-2}{2m-2}}, \tag{3.20}$$

where $||\cdot||_s$ is the norm of the space E_s. Thus from Lemma 3.2 and Lemma 3.3, we have that, for any $s < 2m$

$$\lim_{k\to\infty} \left(|||U^{(k)} - U|||_{s,T_1} + |||D\psi^{(k)} - D\psi|||_{s,T_1}\right) = 0.$$

If we choose $s > 2$, the imbedding property in Lemma 1.1 for the space $E_{m,T}$ implies that $U^{(k)} \to U$ in $C\left([0,\ T_2], C^1(\Omega)\right)$ and $\psi^{(k)} \to \psi$ in $C\left([0,\ T_1], C^2(\Omega)\right)$. From (3.1), we have

$$\begin{aligned} \partial_t U^{(k+1)} =& A_0^{-1}(U^{(k)})H_1(U^{(k)}, D\psi^{(k)}) - A_0^{-1}(U^{(k)})A_1(U^{(k)})\partial_x U^{(k+1)} \\ &- A_0^{-1}(U^{(k)})A_2(U^{(k)})\partial_y U^{(k+1)}, \end{aligned}$$

thus $\partial_t U^{(k)} \to \partial_t U$ in $C([0, T_1], C(\Omega))$. Letting $k \to \infty$ in (3.2), we obtain that $U(t,x,y)$ and $\psi(t,x,y)$ satisfy (1.9)–(1.12) in the classical sense. This completes the proof of Theorem 1.1.

4. Uniqueness. The solution we obtained in §3 is unique in the following sense.

THEOREM 4.1. *Let* (U_i, ψ_i) $(i = 1, 2)$ *with* $U_i \in E_{2m,T}$, $D\psi_i \in E_{2m,T}$ *be the solutions of the coupled system in* (1.9)–(1.12) *with the same initial value* $U_0(x,y) \in E_{2m}(\Omega)$, *which are* 2*m*th*-order compatible. If* $U_i(t,x,y) \in \mathbf{G}_2$ *for* $(t,x,y) \in [0,\ T] \times \Omega$, *then* $U_1(t,x,y) \equiv U_2(t,x,y)$, $\psi_1(t,x,y) \equiv \psi_2(t,x,y)$.

Proof. Setting $U = U_1 - U_2$ and $\psi = \psi_1 - \psi_2$, we have

$$\left\{\begin{array}{l} A_0(U_1)\partial_t U + A_1(U_1)\partial_x U + A_2(U_1)\partial_y U = W_1, \\ \Delta\psi + B_1(DU_1)\partial_x\psi + B_2(DU_1)\partial_y\psi = W_2 \end{array}\right. \tag{4.1}$$

with

$$\left\{\begin{array}{l} \psi(t,-a,y) = \psi(t,a,y) = 0, \;\; \psi(t,x,-\infty) = \psi(t,x,\infty) = 0, \\ u_2(t,-a,y) = u_2(t,a,y) = 0, \\ U(0,x,y) = 0. \end{array}\right. \tag{4.2}$$

Here

$$\begin{aligned} W_1 = H_1(U_1, D\psi_1) - H_1(U_2, D\psi_2) - \big(A_0(U_1) - A_0(U_2)\big)\partial_t U_2 \\ - \big(A_1(U_1) - A_1(U_2)\big)\partial_x U_2 - \big(A_2(U_1) - A_2(U_2)\big)\partial_y U_2, \end{aligned} \tag{4.3}$$

$$\begin{aligned} W_2 = H_2(DU_1) - H_2(DU_2) - \big(B_1(DU_1) - B_1(DU_2)\big)\partial_x\psi_2 \\ - \big(B_2(DU_1) - B_2(DU_2)\big)\partial_y\psi_2. \end{aligned} \tag{4.4}$$

By the Nirenberg inequalities and Taylor's theorem, there are constants C_1 and C_2, such that, for $t \in [0, T]$,

$$|||W_1|||_{2m,t} \le C_1 \left(|||U|||_{2m,t} + |||D\psi|||_{2m,t}\right), \tag{4.5}$$

$$|||W_2|||_{2m-1,t} \le C_2|||U|||_{2m,t}. \tag{4.6}$$

From (2.34) in Remark 1 and (4.6), it follows that there exists C' such that

$$|||D\psi|||_{2m,t} \le C'|||U|||_{2m,t}. \tag{4.7}$$

Again it follows from (2.5), (4.5), and (4.7) that there exists C'' such that

$$||U||_{2m,t} \le C''t|||U|||_{2m,t}. \tag{4.8}$$

The inequality (4.8) implies that for $t \ll 1$, we have $|||U|||_{2m,t} = 0$ and consequently $|||D\psi|||_{2m,t} = 0$. Then by the standard continuation argument, we obtain $(U, \psi) \equiv (0,0)$ for all $t \in [0, T]$. This completes the proof.

REFERENCES

[1] S. Alinhac, *Existence d'ondes de rarefaction pour des systemes quasi-lineaires multidimensionnels*, Comm. Partial Differential Equations, 14 (1989), pp. 173–230.

[2] A. J. F. Davidson, R. Clift, and D. Harrison, *Fluidization*, 2nd ed., Academic Press, New York, 1985.

[3] D. A. Drew, *Mathematical modeling of two-phase flow*, Ann. Rev. Fluid Mech., 15 (1983), pp. 261–291.

[4] G. Ganser and J. Lightbourne, *Oscillatory traveling waves in a hyperbolic model of a fluidized bed*, Chem. Engrg. Sci., 46 (1991), pp. 1339–1347.

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

[6] D. Li, *Rarefaction and shock waves for multi-dimensional hyperbolic conservation laws*, Comm. Partial Differential Equations, 16 (1991), pp. 425–450.

[7] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Springer-Verlag, Berlin, New York, Heidelberg, 1984.

[8] J. Rauch, *Symmetric positive systems with boundary characteristic of constant multiplicity*, Trans. Amer. Math. Soc., 291 (1985), pp. 167–187.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 1038–1048, July 1996

009

EXISTENCE AND BLOW-UP OF SOLUTIONS TO TWO-PHASE NONEQUILIBRIUM PROBLEMS*

ZHICHENG GUAN[†] AND XU-JIA WANG[†]

Abstract. In this paper, we deal with the one-dimensional Stefan problem

$$u_t - u_{xx} = \Gamma'(t)\delta(x - \Gamma(t)) \text{ in } I\!R \times I\!R^+, u(x,0) = u_0(x)$$

with a kinetic condition $\Gamma'(t) = f(u)$ on the free boundary $F = \{(x,t), x = \Gamma(t)\}$, where $\delta(x)$ is the Dirac function. We show that if $|f(u)| \leq M\, e^{\gamma|u|}$ for some $M > 0$ and $\gamma \in (0, \frac{1}{4})$, then there exists a global solution to the above problem. We also give an example to show that the solution may blow up in finite time if $f(u) \geq Ce^{|u|}$ for some $C > 0$.

Key words. free boundary, Stefan problem, existence

AMS subject classifications. 35K15, 35R35

1. Introduction.

In this paper, we deal with the following one-dimensional Stefan problem with kinetic condition on the free boundary:

$$(1.1)\qquad \begin{cases} u_t - u_{xx} = 0 & \text{in } Q\backslash F, \\ u^-(\Gamma(t),t) = u^+(\Gamma(t),t) & \text{on } F, \\ u_x^-(\Gamma(t),t) - u_x^+(\Gamma(t),t) = \Gamma'(t) & \text{on } F, \\ \Gamma'(t) = f(u),\ \Gamma(0) = b & \text{on } F, \\ u(x,0) = u_0(x), & \end{cases}$$

where $Q = I\!R \times I\!R^+$ and $F = \{(x,t); x = \Gamma(t)\}$ is the free boundary.

Problem (1.1) with $Q = (0,1) \times (0,T]$ has been studied by many authors [3, 9, 10], and the case when $Q = I\!R \times I\!R^+$ has been studied by Yin [11]. In both cases, the local existence and uniqueness of solutions have been established. But to obtain the global existence of solutions, the condition that $f(u)u \leq C(1 + u^2)$ was imposed in [3, 9, 10].

Here we are interested in the existence of global solutions without the restriction above. Recently one of the authors [8] obtained the global existence of solutions to the problem (1.1) with $Q = (0,1) \times (0,\infty)$ under the following conditions:

(i) $f(u) = u^{2m}$,

(ii) $\Gamma(0) = \Gamma_0 \geq 1 - K_0^{1-2m}(1 - \frac{1}{2m})^{2m}/(2m-1)$,

where $m > \frac{1}{2}$ and $K_0 = \max\{\|u_0(x)\|_{L^\infty(0,1)}, \|u(0,t)\|_{L^\infty(0,\infty)}, \|u(1,t)\|_{L^\infty(0,\infty)}\}$.

In this paper, we will prove that if $|f(u)| \leq M\, e^{\gamma|u|}$ for some $M > 0$ and $\gamma \in (0, \frac{1}{4})$, then there exists a global solution to the problem (1.1). If $f(u) \geq \delta e^{|u|}$ for some $\delta > 0$, we will give an example to show that the solution may blow up in finite time.

This paper is arranged as follows. In §1, we give the finite-time blow-up example. In §2, we prove the global existence result for the problem (1.1).

2. A blow-up example.

For any given $T > 0$, we construct $f(u)$ and $\Gamma(t)$ so that the solution u of problem (1.1) blows up at $t = T$.

*Received by the editors June 25, 1993; accepted for publication (in revised form) November 11, 1994. This research was supported by the NNSF of China and the NSF of Zhejiang Province.

[†]Department of Mathematics, Zhejiang University, Hangzhou 310027, People's Republic of China.

Let $\Gamma(t) \in C^2[0,T)$ be given; we consider the problem

$$\text{(2.1)} \qquad \begin{cases} u_t - u_{xx} = 0 & \text{in } Q\backslash F, \\ u^-(\Gamma(t),t) = u^+(\Gamma(t),t) & \text{on } F, \\ u_x^-(\Gamma(t),t) - u_x^+(\Gamma(t),t) = \Gamma'(t) & \text{on } F, \\ u(x,0) = 0 \text{ and } u(x,t) \to 0 & \text{as } |x| \to \infty. \end{cases}$$

Problem (2.1) is uniquely solvable. In the sense of distributions, the solution of (2.1) is equivalent to the solution of

$$\text{(2.2)} \qquad \begin{cases} u_t - u_{xx} = \Gamma'(t)\delta(x - \Gamma(t)) & \text{in } I\!R \times (0,T), \\ u(x,0) = 0, \end{cases}$$

where $\delta(x)$ is the Dirac function. The solution of (2.2) is given by

$$\text{(2.3)} \qquad u(x,t) = \int_0^t \int_{-\infty}^{\infty} \Gamma'(s)\delta(\xi - \Gamma(s))K(x-\xi, t-s)d\xi ds,$$

where

$$K(x,t) = \begin{cases} \frac{1}{\sqrt{4\pi t}} e^{-\frac{x^2}{4t}}, & t > 0, \\ 0, & t \leq 0. \end{cases}$$

Hence

$$\text{(2.4)} \qquad \begin{aligned} u(x,t) &= \int_0^t \frac{\Gamma'(s)}{\sqrt{4\pi(t-s)}} e^{-\frac{|x-\Gamma(s)|^2}{4(t-s)}} ds \\ &= \int_0^t \frac{\Gamma'(t-s)}{\sqrt{4\pi s}} e^{-\frac{|x-\Gamma(t-s)|^2}{4s}} ds. \end{aligned}$$

For $u(x,t)$ given above, one can verify that (see [1, Chap. 14])

$$\text{(2.5)} \qquad \lim_{x\to\Gamma(t)^-} \frac{\partial u(x,t)}{\partial x} = \frac{1}{2}\Gamma'(t) + \int_0^t \frac{\partial K}{\partial x}(\Gamma(t) - \Gamma(s), t-s)\Gamma'(s)\, ds,$$

$$\text{(2.6)} \qquad \lim_{x\to\Gamma(t)^+} \frac{\partial u(x,t)}{\partial x} = -\frac{1}{2}\Gamma'(t) + \int_0^t \frac{\partial K}{\partial x}(\Gamma(t) - \Gamma(s), t-s)\Gamma'(s)\, ds.$$

Hence the function $u(x,t)$ in (2.4) is a solution of (1.1). Note that for any $t > 0$, the maximum of $|u(x,t)|$ in $I\!R \times \underline{(0,T)}$ is attained on F.

Now let $\Gamma(t) = -\delta\sqrt{T-t}$, where $0 < \delta \leq \frac{1}{2}$ is a constant. Denote $\varepsilon = T - t$. We have

$$\text{(2.7)} \qquad \Gamma'(t-s) = \frac{\delta}{2\sqrt{\varepsilon+s}}, \quad \Gamma''(t-s) = \frac{\delta}{4(\varepsilon+s)^{3/2}}.$$

Let $F(t,s) = \frac{1}{4s}|\Gamma(t) - \Gamma(t-s)|^2 = \frac{\delta^2}{4s}|\sqrt{\varepsilon+s} - \sqrt{\varepsilon}|^2$. Then

$$\text{(2.8)} \qquad F(t,s) = \frac{\delta^2}{4} \frac{\sqrt{\varepsilon+s} - \sqrt{\varepsilon}}{\sqrt{\varepsilon+s} + \sqrt{\varepsilon}} < \frac{\delta^2}{4},$$

$$\frac{\partial}{\partial t}F(t,s)=\frac{\delta^2}{4s}(\sqrt{\varepsilon+s}-\sqrt{\varepsilon})\left(\frac{1}{\sqrt{\varepsilon}}-\frac{1}{\sqrt{\varepsilon+s}}\right)\le\frac{\delta^2}{4}\frac{1}{\sqrt{\varepsilon}\sqrt{\varepsilon+s}}. \tag{2.9}$$

From (2.4), we therefore obtain

$$\begin{aligned} u(\Gamma(t),t)&=\int_0^t\frac{\Gamma'(t-s)}{\sqrt{4\pi s}}e^{-F(t,s)}\,ds\\ &\ge\frac{e^{-\delta^2/4}}{\sqrt{4\pi}}\int_0^t\frac{\Gamma'(t-s)}{\sqrt{s}}\,ds\\ &=\frac{e^{-\delta^2/4}}{\sqrt{4\pi}}\int_0^t\frac{\delta}{2\sqrt{s}\sqrt{\varepsilon+s}}\,ds\\ &=\frac{e^{-\delta^2/4}}{4\sqrt{\pi}}\int_0^{t/\varepsilon}\frac{\delta}{\sqrt{s}\sqrt{1+s}}\,ds\\ &=\frac{\delta e^{-\delta^2/4}}{2\sqrt{\pi}}\log\left(\sqrt{1+\frac{t}{\varepsilon}}+\sqrt{\frac{t}{\varepsilon}}\right)\\ &\ge\frac{\delta}{16}\log\left(1+\frac{t}{\varepsilon}\right)=\frac{\delta}{16}\log\frac{T}{T-t}; \end{aligned} \tag{2.10}$$

and

$$\begin{aligned} u(\Gamma(t),t)&\le\frac{1}{\sqrt{4\pi}}\int_0^t\frac{\Gamma'(t-s)}{\sqrt{s}}\,ds\\ &=\frac{1}{4\sqrt{\pi}}\int_0^{t/\varepsilon}\frac{\delta}{\sqrt{s}\sqrt{1+s}}\,ds\\ &=\frac{\delta}{2\sqrt{\pi}}\log\left(\sqrt{1+\frac{t}{\varepsilon}}+\sqrt{\frac{t}{\varepsilon}}\right)\\ &\le\begin{cases}\delta\cdot\sqrt{\frac{t}{T-t}} & \text{if } t\le T/2\\ \delta\cdot\log\frac{T}{T-t} & \text{if } t\ge T/2.\end{cases} \end{aligned} \tag{2.11}$$

From (2.10), it follows that $u(\Gamma(t),t)\to\infty$ as $t\to T$. Next, we show that $u(\Gamma(t),t)$ is strictly increasing in $t\in(0,T)$. We have

$$\begin{aligned} \frac{d}{dt}u(\Gamma(t),t)&=\int_0^t\frac{\Gamma''(t-s)}{\sqrt{4\pi s}}\,e^{-F(t,s)}\,ds-\int_0^t\frac{\Gamma'(t-s)}{\sqrt{4\pi s}}\,F_te^{-F(t,s)}\,ds\\ &\quad+\frac{\Gamma'(0)}{\sqrt{4\pi t}}\,e^{-|\Gamma(t)-\Gamma(0)|^2/4t}=:I_1+I_2+I_3, \end{aligned} \tag{2.12}$$

where $I_3>0$. By (2.7) and (2.8), we have

$$\begin{aligned} I_1&\ge\frac{e^{-1/4}}{\sqrt{4\pi}}\int_0^t\frac{\Gamma''(t-s)}{\sqrt{s}}\,ds\\ &=\frac{e^{-1/4}}{8\sqrt{\pi}}\int_0^t\frac{\delta}{\sqrt{s}(\varepsilon+s)^{3/2}}\,ds\\ &=\frac{\delta e^{-1/4}}{8\varepsilon\sqrt{\pi}}\int_0^{t/\varepsilon}\frac{1}{\sqrt{s}(1+s)^{3/2}}\,ds\\ &=\frac{\delta e^{-1/4}}{4\varepsilon\sqrt{\pi}}\sqrt{\frac{t}{t+\varepsilon}}=\frac{\delta e^{-1/4}}{4\varepsilon\sqrt{\pi}}\sqrt{\frac{t}{T}}. \end{aligned} \tag{2.13}$$

From (2.9), we have

$$
\begin{aligned}
|I_2| &\le \int_0^t \frac{|\Gamma'(t-s)F_t|}{\sqrt{4\pi s}}\,ds \\
&\le \int_0^t \frac{\delta^2|\Gamma'(t-s)|}{8\sqrt{\pi}\sqrt{s}\sqrt{\varepsilon}\sqrt{\varepsilon+s}}ds \\
&\le \frac{\delta^3}{16\sqrt{\pi}}\int_0^t \frac{1}{(\varepsilon+s)\sqrt{\varepsilon}\sqrt{s}}ds \\
&= \frac{\delta^3}{16\varepsilon\sqrt{\pi}}\int_0^{t/\varepsilon} \frac{1}{\sqrt{s}(1+s)}\,ds \\
&= \frac{\delta^3}{8\varepsilon\sqrt{\pi}}\operatorname{arctg}\sqrt{\frac{t}{T-t}},
\end{aligned}
\tag{2.14}
$$

where $\varepsilon = T - t$. Note that since $\sin\theta \ge \frac{2}{\pi}\theta$ for $\theta \in (0, \frac{\pi}{2})$, we have $\sqrt{\frac{t}{T}} \ge \frac{2}{\pi}\operatorname{arctg}\sqrt{\frac{t}{T-t}}$. Hence if $0 < \delta \le \frac{1}{2}$, by (2.13) and (2.14) we have $I_1 > I_2$, and so $\frac{d}{dt}u(\Gamma(t), t) > 0$ for $t \in (0, 1)$.

Since $u(\Gamma(t), t)$ is strictly increasing and tends to infinity as $t \to T$, we can define $f(u) : I\!R^+ \to I\!R^+$ by

$$
f(u(\Gamma(t),t)) = \Gamma'(t) = \frac{\delta}{2\sqrt{T-t}}.
\tag{2.15}
$$

For the function $f(u)$ defined above, we therefore conclude that the solution $u(x,t)$ blows up at time $t = T$.

Remark 2.1. Since $\Gamma'(t)$ and $u(\Gamma(t), t)$ are positive and strictly increasing, it follows that $f(u)$ is also positive and increasing in $(0, T)$. From (2.10) and (2.11), it is easy to see that $f(u)$ is of exponential growth as $u \to +\infty$. From (2.10) and (2.15), we have

$$
f(u) \le \frac{\delta}{2\sqrt{T}}e^{8u/\delta};
\tag{2.16}
$$

by (2.11) and (2.15), we have

$$
f(u) \ge \frac{\delta}{2\sqrt{T}}e^{u/2\delta} \text{ for } u \ge \delta,
\tag{2.17}
$$

where $\delta \in (0, \frac{1}{2}]$. Moreover, $\lim_{u\to 0} f(u) = \frac{\delta}{2\sqrt{T}}$.

3. Global existence. Let us begin with the local existence and regularity for solutions of (1.1). Suppose $u(x,t)$ is a solution of (1.1). Then $u(x,t)$ satisfies in the sense of distributions that

$$
\begin{cases}
u_t - u_{xx} = \Gamma'(t)\delta(x - \Gamma(t)) & \text{in } I\!R \times I\!R^+, \\
u(x,0) = u_0(x).
\end{cases}
$$

Hence

(3.1)

$$
\begin{aligned}
u(x,t) &= \int_{-\infty}^{\infty} K(x-\xi,t)u_0(\xi)d\xi + \int_0^t\int_{-\infty}^{\infty} \Gamma'(s)\delta(\xi-\Gamma(s))K(x-\xi,t-s)d\xi ds \\
&= \int_{-\infty}^{\infty} K(x-\xi,t)u_0(\xi)d\xi + \int_0^t \Gamma'(s)K(x-\Gamma(s),t-s)ds,
\end{aligned}
$$

where $K(x,t)$ was as defined in §2. We have the following lemma.

LEMMA 3.1. *Suppose* $u_0(x) \in C(I\!R) \cap L^\infty(I\!R)$ *and* $f(y)$ *is locally Lipschitz continuous. Then the problem* (1.1) *admits a local solution* $u(x,t) \in C^0(Q_T) \cap C^\infty(Q_T^\pm)$ *and* $\Gamma(t) \in C^1[0,T]$ *for some* $T > 0$, *where* $Q_T^+ = Q_T \cap \{x > \Gamma(t)\}$, $Q_T^- = Q_T \cap \{x < \Gamma(t)\}$, $Q_T = I\!R \times (0,T]$.

Proof. Let

$$A = \{\Gamma(t) \in C^1[0,T];\quad \Gamma(0) = b, \Gamma'(b) = f(u_0(b)),\ \text{and}\ \|\Gamma\|_{C^1[0,T]} \le M\},$$

where $M > 1$ and $T < 1$ are positive constants to be determined. For any $\Gamma(t) \in A$, let $u(x,t)$ be the function defined by (3.1). Then $u(\Gamma(t),t) \in C[0,T]$. Direct computation gives

$$|u_1(\Gamma_1(t),t) - u_2(\Gamma_2(t),t)| \le Ct^{1/2}\|\Gamma_1 - \Gamma_2\|_{C^1[0,T]},$$

where C is independent of t.

Let $G:\ \Gamma \in A \to \widetilde{\Gamma}(t) \in A$ be a mapping defined by

$$\widetilde{\Gamma}(t) = b + \int_0^t f(u(\Gamma(s),s))\,ds.$$

Then $\widetilde{\Gamma}(0) = b$, $\widetilde{\Gamma}'(0) = f(u_0(b))$, and $\widetilde{\Gamma} \in C^1[0,T]$. If M is large enough and T is small enough, by the Lipschitz continuity of f, it is easy to check that $G(A) \subset A$ and G is a contraction mapping. Hence G has a fixed point which gives a local solution to the problem (1.1). □

Remark 3.1. The fact that G is a contraction mapping implies the uniqueness of solutions to (1.1). Let

$$A = \{\Gamma \in C^1[0,T] \cap C^{1,1}(0,T];\quad \Gamma(0) = b, \Gamma'(0) = f(u_0(b)),\ \text{and}\ \|\Gamma\|_{\widetilde{C}^{1,1}(0,T]} < M\},$$

where

$$\|\Gamma\|_{\widetilde{C}^{1,1}(0,T]} = \|\Gamma\|_{C^1[0,T]} + \sup_{0<t<t+\delta\le T} t\delta^{-1}|\Gamma'(t+\delta) - \Gamma(t)|.$$

Then more careful computation (see Lemma 14.2.8 in [1]) shows that G is also a contraction mapping provided T is small enough. Hence the local solution satisfies $\Gamma \in C^{1,1}(0,T]$ and (by the intermediate Schauder estimates) $u \in C_{x,t}^{1+\alpha,(1+\alpha)/2}(\overline{Q_{T,\varepsilon}^\pm})$ for any $\varepsilon \in (0,T)$ and $\alpha \in (0,1)$, where $Q_{T,\varepsilon}^\pm = Q_T^\pm \cap \{t > \varepsilon\}$.

The main result of this section is the following theorem.

THEOREM 3.1. *Suppose* $u_0(x) \in C(I\!R)$ *is bounded and* $f(t)$ *is locally Lipschitz continuous. If there exist constants* $M > 0$ *and* $\gamma \in (0, \frac{1}{4})$ *so that*

$$|f(u)| \le M\,e^{\gamma|u|}, \tag{3.2}$$

then there exists a global solution to the problem (1.1).

Proof. For any $T > 0$, we will prove

$$u(x,T) \le M(T) =: M_0 + 1 + \frac{4}{\varepsilon}M(1+T)^2 e^{k^*+2}, \tag{3.3}$$

where $M_0 = \|u_0\|_{L^\infty(I\!R)}$,

$$\varepsilon = (1-4\gamma)/8, \tag{3.4}$$

and

$$k^* = \max(k_1^*,\ k_2^*,\ k_3^*)$$

with

$$k_1^* = \frac{48}{\varepsilon^3}(M_0 + T + 4),\quad k_2^* = \frac{1}{\varepsilon\gamma}\left|\log\frac{\varepsilon\gamma}{2}\right|,\quad k_3^* = \frac{4}{1-4\gamma}(M_0 + T + 4 + M(1+T)).$$

Note that (3.3) implies $u(x,t) \le C_1(1 + t^2 e^{C_2 t})$ for some C_1 and C_2 depending only on M_0, ε, γ, and M.

To prove (3.3), we argue by contradiction. Suppose (3.3) is false at $T > 0$. By the maximum principle, we have

$$|u(\Gamma(T), T)| \ge M(T). \tag{3.5}$$

The first integral on the right-hand side of (3.1) is the solution of the Cauchy problem

$$\begin{cases} u_t - u_{xx} = 0 \text{ in } I\!R \times I\!R^+, \\ u(x,0) = u_0(x). \end{cases}$$

Hence it is bounded with bound M_0. For the second integral, we have

$$\begin{aligned} \int_0^t \Gamma'(s) K(\Gamma(t) - \Gamma(s), t-s) ds &= \frac{1}{\sqrt{4\pi}} \int_0^t \frac{\Gamma'(s)}{\sqrt{t-s}} e^{-(\Gamma(t)-\Gamma(s))^2/4(t-s)} ds \\ &= \frac{-1}{\sqrt{\pi}} \int_0^t e^{-(\Gamma(t)-\Gamma(s))^2/4(t-s)} d\frac{\Gamma(t)-\Gamma(s)}{2\sqrt{t-s}} \\ &+ \frac{1}{4\sqrt{\pi}} \int_0^t \frac{\Gamma(t)-\Gamma(s)}{(t-s)^{3/2}} e^{-(\Gamma(t)-\Gamma(s))^2/4(t-s)} ds. \end{aligned} \tag{3.6}$$

Since

$$\int_{-\infty}^{\infty} e^{-t^2} dt = \sqrt{\pi}, \tag{3.7}$$

the first integral on the right-hand side of (3.6) is bounded. We therefore conclude that

$$|w(T)| \ge 4\sqrt{\pi}(M(T) - M_0 - 1),$$

where

$$w(t) = \int_0^t \frac{\Gamma(t)-\Gamma(s)}{(t-s)^{3/2}} e^{-(\Gamma(t)-\Gamma(s))^2/4(t-s)} ds. \tag{3.8}$$

Note that the integrand in (3.8) satisfies

$$\frac{|\Gamma(t)-\Gamma(s)|}{(t-s)^{3/2}} e^{-(\Gamma(t)-\Gamma(s))^2/4(t-s)} < \frac{1}{t-s}. \tag{3.9}$$

Let

$$E = \{t \in [0,T];\ \ \forall\ \ s < t, |\Gamma(t) - \Gamma(s)| \le \sqrt{t-s}/(4 + \log^2(t-s))\}.$$

By the C^1 continuity of Γ, it follows that $[0,\delta_0] \subset E$ for some $\delta_0 > 0$ small. Let $D = [0,T]\backslash E$. For any $t \in D$, let

$$\delta(t) = \inf\{\varepsilon \in (0,t); |\Gamma(t) - \Gamma(t-\varepsilon)| > \sqrt{\varepsilon}/(4+\log^2 \varepsilon)\}.$$

Since $\Gamma(t) \in C^1[0,T]$, it follows that D is an open subset of $[0,T]$, $\delta(t)$ is positive and upper semicontinuous on D, and for any $\tau \in (0,T)$, there exists a constant $C_\tau > 0$ so that

$$\delta(t) \geq C_\tau \quad \text{for any } t \in (0,\tau). \tag{3.10}$$

Moreover,

$$|\Gamma(t) - \Gamma(t-\delta(t))| = \sqrt{\delta(t)}/(4+\log^2 \delta(t)) \quad \text{for } t \in D. \tag{3.11}$$

For any $t \in E$, we have

$$|w(t)| \leq \int_0^t \frac{|\Gamma(t)-\Gamma(s)|}{(t-s)^{3/2}} ds \leq \int_0^t \frac{1}{(t-s)(4+\log^2(t-s))} ds \leq 1.$$

Hence by (3.1), (3.6), and (3.7), we have

$$|u(\Gamma(t),t)| \leq M_0 + 2 \quad \text{for} \quad t \in E. \tag{3.12}$$

For any $t \in D$, by (3.9) we have

$$\begin{aligned} |w(t)| &= \left| \left(\int_0^{t-\delta(t)} + \int_{t-\delta(t)}^t \right) \left(\frac{|\Gamma(t)-\Gamma(s)|}{(t-s)^{3/2}} e^{-(\Gamma(t)-\Gamma(s))^2/4(t-s)} \right) ds \right| \\ &\leq \int_0^{t-\delta(t)} \frac{1}{t-s} ds + \int_{t-\delta(t)}^t \frac{1}{(t-s)(4+\log^2(t-s))} ds \\ &\leq 1 + \log T - \log \delta(t) \leq 1 + T - \log \delta(t). \end{aligned} \tag{3.13}$$

Since the first integral of the right-hand side of (3.1) is bounded with bound M_0, from (3.6) we have

$$\int_0^T \frac{|\Gamma'(s)|}{\sqrt{t-s}} ds \geq \sqrt{4\pi}(|u(\Gamma(T),T)| - M_0) \geq \sqrt{4\pi}(M(T) - M_0 - 1). \tag{3.14}$$

Let

$$\beta = \int_0^T e^{2\gamma(1+\varepsilon)|u(\Gamma(s),s)|} ds.$$

By (3.2), we have

(3.15)
$$\begin{aligned} &\int_0^T \frac{|\Gamma'(s)|}{\sqrt{T-s}} ds \leq M \int_0^T \frac{e^{\gamma|u(\Gamma(s),s)|}}{\sqrt{T-s}} ds \\ &\leq M \left(\int_0^T (T-s)^{-(1+\varepsilon)/(1+2\varepsilon)} ds \right)^{(1+2\varepsilon)/2(1+\varepsilon)} \left(\int_0^T e^{2\gamma(1+\varepsilon)|u(\Gamma(s),s)|} ds \right)^{1/2(1+\varepsilon)} \\ &\leq \frac{1+2\varepsilon}{\varepsilon} M T^{\varepsilon/2(1+\varepsilon)} \left(\int_0^T e^{2\gamma(1+\varepsilon)|u(\Gamma(s),s)|} ds \right)^{1/2(1+\varepsilon)} \\ &\leq \frac{1+2\varepsilon}{\varepsilon} M \, T^{\varepsilon/2(1+\varepsilon)} \beta^{1/2(1+\varepsilon)}. \end{aligned}$$

Combining (3.14) and (3.15) yields

$$\beta > 4(T+1)e^{k^*+2}. \tag{3.16}$$

For any integer $k \geq 0$, let

$$A_k = \{t \in [0,T];\ \ k \leq |u(\Gamma(t),t)| \leq k+1\}.$$

Then

$$\sum_{k=0}^{\infty} \text{mes}(A_k)e^{2\gamma(1+\varepsilon)(k+1)} \geq \beta.$$

Let

$$a_k = e^{2\gamma(1+\varepsilon)(k+1)}\text{mes}(A_k)/\beta.$$

Then

$$\sum_{k=0}^{\infty} a_k \geq 1,$$

and by (3.16),

$$\sum_{k=0}^{k^*} a_k \leq \sum_{k=0}^{k^*} \frac{\text{mes}(A_k)}{\beta} e^{2\gamma(1+\varepsilon)(k+1)} \leq \frac{\text{mes}(A_k)}{\beta} e^{2\gamma(1+\varepsilon)(k^*+2)} \leq \frac{1}{2}.$$

Hence

$$\sum_{k=k^*+1}^{\infty} a_k \geq \frac{1}{2}.$$

Since $k^* \geq k_2^* = \frac{1}{\varepsilon\gamma}|\log\frac{\varepsilon\gamma}{2}|$, we have

$$\sum_{k=k^*+1}^{\infty} e^{-\varepsilon\gamma k} \leq \frac{1}{2}.$$

Hence there exists a $k > k^*$ so that $a_k \geq e^{-\varepsilon\gamma k}$.

For the integer k determined above, by $a_k \geq e^{-\varepsilon\gamma k}$ we have

$$e^{2\gamma(1+\varepsilon)(k+1)}\text{mes}(A_k) > \beta\, e^{-\varepsilon\gamma k},$$

which implies

$$\text{mes}(A_k) > \beta\, e^{-2\gamma(1+2\varepsilon)(k+1)}. \tag{3.17}$$

From (3.12), we have $A_k \subset D$ for $k \geq k^*$. From (3.1), (3.6), (3.7), and (3.13), we have

$$|u(\Gamma(t),t)| \leq \|u_0\|_{L^\infty} + 2 + |w(t)| \leq M_0 + T + 3 - \log\delta(t).$$

Since $k \leq |u(\Gamma(t),t)| \leq k+1$ in A_k, by (3.13) it follows that

$$\delta(t) \leq e^{-k+M'} \quad \text{for} \ \ t \in A_k, \tag{3.18}$$

where $M' = M_0 + T + 3$.

Let $t^0 = \sup\{t \in A_k\}$, and $t_0 = t^0 - \delta(t^0)$. By the closedness of A_k, we have $t^0 \in A_k$. For any $j \geq 1$, we define t^j and t_j inductively by letting $t^j = \sup\{t \leq t_{j-1}, t \in A_k\}$ and $t_j = t^j - \delta(t^j)$. Since A_k is closed, it follows that $t^j \in A_k$. By (3.10), the above procedure finishes after finitely many steps, that is, there is a $j_0 > 0$ so that $(0, t_{j_0}) \cap A_k = \emptyset$.

We therefore obtain a sequence of intervals (t_j, t^j) which satisfies the following properties:

(i) $t_j \geq t^{j+1}$;

(ii) $A_k \subset \cup_{j=0}^{j_0}[t_j, t^j]$;

(iii) $t^j \in A_k$.

By (i) and (3.11), we conclude

$$\int_0^T |\Gamma'(s)|ds \geq \sum_{j=0}^{j_0} |\Gamma(t^j) - \Gamma(t_j)| = \sum \sqrt{\delta(t^j)}/(4 + \log^2 \delta(t^j)).$$

Noticing that $k > k_1^* = \frac{48}{\varepsilon^3}(M_0 + T + 4)$, we have $\sqrt{\delta(t^j)}/(4 + \log^2 \delta(t^j) \geq [\delta(t^j)]^{(1+\varepsilon)/2}$. By (ii), (3.18), and (3.17), it follows that

$$\begin{aligned}\int_0^T |\Gamma'(s)|ds &\geq \sum(\delta(t^j))^{(1+\varepsilon)/2} \\ &\geq \inf(|\delta(t^j)|^{-(1-\varepsilon)/2}) \sum \delta(t^j) \\ &\geq \inf(|\delta(t^j)|^{-(1-\varepsilon)/2}) \text{mes} A_k \\ &\geq e^{(k-M')(1-\varepsilon)/2} \text{mes} A_k \\ &\geq e^{(k-M')(1-\varepsilon)/2} \cdot e^{-2\gamma(1+2\varepsilon)(k+1)}\beta \\ &\geq e^{(\frac{1}{2}-2\gamma-2\varepsilon)k-M'-1}\beta.\end{aligned}$$

On the other hand,

$$\begin{aligned}\int_0^T |\Gamma'(s)|ds &\leq M \int_0^T e^{\gamma|u(\Gamma(s),s)|} ds \\ &\leq M \left(\int_0^T ds\right)^{(1+2\varepsilon)/2(1+\varepsilon)} \left(\int_0^T e^{2\gamma(1+\varepsilon)|u(\Gamma(s),s)|} ds\right)^{1/2(1+\varepsilon)} \\ &\leq M T^{(1+2\varepsilon)/2(1+\varepsilon)} \beta^{1/2(1+\varepsilon)} \leq M(1+T)\beta^{1/2(1+\varepsilon)}.\end{aligned}$$

We obtain

$$e^{(\frac{1}{2}-2\gamma-2\varepsilon)k-M'-1}\beta \leq M(1+T)\beta^{1/2(1+\varepsilon)}.$$

But $k \geq k^* \geq \frac{4}{1-4\gamma}(M_0 + T + 4 + M(1+T))$, so we reach a contradiction because of $\beta > 1$. Hence (3.3) holds, which completes the proof. □

Remark 3.2. It is interesting to compare Theorem 3.1 with the finite-time blow-up example in §2. Theorem 3.1 asserts that (1.1) possesses a global solution provided $f(u) \leq Me^{\gamma|u|}$ for some $M > 0$ and $\gamma \in (0, \frac{1}{4})$, while in the example of §2, if we choose $\delta = \frac{1}{2}$ and let T be large enough, then the solution of (2.1) blows up in finite time, though $f(u) \leq \frac{1}{\sqrt{T}}e^{16u}$ by (2.16).

Remark 3.3. The condition $\gamma \in (0, \frac{1}{4})$ in Theorem 3.1 may be improved, but we are unable to do so by the method employed above.

It is not hard to see that the result in Theorem 3.1 is also true if the domain $Q = I\!R \times I\!R^+$ in problem (1.1) is replaced by $Q = (0,1) \times (0,T]$. Let us consider the problem

$$(3.19) \qquad \begin{cases} u_t - u_{xx} = 0 & \text{in } Q\backslash F \\ u^-(\Gamma(t),t) = u^+(\Gamma(t),t) & \text{on } F, \\ u_x^-(\Gamma(t),t) - u_x^+(\Gamma(t),t) = \Gamma'(t) & \text{on } F, \\ \Gamma'(t) = f(u),\ \Gamma(0) = b \in (0,1) & \text{on } F, \\ u(x,0) = u_0(x) & \text{and } \ u(0,t) = u(1,t) = 0. \end{cases}$$

The problem (3.19) has been studied in [8], where the author proved the global existence of solutions under conditions (i) and (ii) stated in the introduction. Here we have the following theorem.

THEOREM 3.2. *Suppose* $u_0(x) \in C[0,1]$, $f(u)$ *is locally Lipschitz continuous, and* $|f(u)| \leq Me^{\gamma|u|}$ *for some* $M > 0$ *and* $\gamma \in (0, \frac{1}{4})$. *Then there is a global solution to the problem* (3.19).

By global solution, we mean either $0 < \Gamma(t) < 1$ for $t \in [0,T]$ and $\Gamma(t) \in C^{1,1}[0,T]$, or there exists a $t_0 \in (0,T]$ so that the solution exists up to the moment $t = t_0$, $\lim_{t\to t_0} \Gamma(t)$ is 0 or 1 and $\Gamma(t) \in C^{1,1}[0,t_0)$.

The proof is similar to that of Theorem 3.1 with (3.1) replaced by

$$(3.20) \quad \begin{aligned} u(x,t) &= \int_0^1 (\theta(x-\xi,t) - \theta(x+\xi,t))u_0(\xi)d\xi \\ &\quad + \int_0^t \int_0^1 \Gamma'(s)\delta(\xi - \Gamma(s))[\theta(x-\xi,t-s) - \theta(x+\xi,t-s)]d\xi ds \\ &= \int_0^1 (\theta(x-\xi,t) - \theta(x+\xi,t))u_0(\xi)d\xi \\ &\quad + \int_0^t \Gamma'(s)[\theta(x-\Gamma(s),t-s) - \theta(x+\Gamma(s),t-s)]ds, \end{aligned}$$

where

$$\theta(x,t) = \sum_{m=-\infty}^{\infty} K(x+2m,t)$$

and $K(x,t)$ is the fundamental solution of the heat equation $u_t - u_{xx} = 0$. We omit the proof here.

Acknowledgment. We would like to express our thanks to the referee for helpful comments.

REFERENCES

[1] J. R. CANNON, *The One-Dimensional Heat Equation*, Addison–Wesley, Reading, MA, 1984.

[2] S. R. CORIELL AND R. L. PARKER, *Interface kinetics and the stability of the shape of a solid sphere glowing from the melt*, in Proc. International Conference on Crystal Growth, Boston, 1966, pp. 20–24.

[3] A. B. CROWLEY, *Some remarks on a nonequilibrium solidification problem*, in Free and Moving Boundary Problems, K. H. Hoffmann and J. Sprekels, eds., Pitman, Boston, 1989.

[4] L. C. EVANS, *Free boundary problem: the flow of the two immiscible fluids is one-dimensional porous medium*, Indiana Univ. Math. J., 27 (1978), pp. 93–111.

[5] A. FASANO AND M. PRIMICERIO, EDS., *Free boundary problems, theory and applications*, Pitman, Boston, MA, 1983.

[6] A. Fasano, M. Primicerio, S. D. Howison, and J. R. Ockendon, *Some remarks on the regularization of supercooled one-phase Stefan problems in one dimension*, Quart. Appl. Math., 48 (1990), pp. 153–168.

[7] A. Friedman and J. B. McLeod, *Blowup of positive solutions of semilinear heat equations*, Indiana Univ. Math. J., 34 (1985), pp. 425–477.

[8] Z. Guan, *On the global solutions to the two phase nonequilibrium problems*, Meccanica, to appear.

[9] A. Visitin, *Stefan problem with a kinetic condition at the free boundary*, Ann. Mat. Pura Appl., 146 (1987), pp. 97–122.

[10] W. Xie, *The Stefan problem with a kinetic condition at the free boundary*, SIAM J. Math. Anal., 21 (1990), pp. 362–373.

[11] H.-M Yin, *Blowup and global existence for a non-equilibrium phase change process*, preprint.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 1049–1056, July 1996

GLOBAL UNIQUENESS IN THE IMPEDANCE-IMAGING PROBLEM FOR LESS REGULAR CONDUCTIVITIES[*]

RUSSELL M. BROWN[†]

Abstract. If $L_\gamma = \operatorname{div} \gamma \nabla$ is an elliptic operator with scalar coefficient γ, we show that we can recover the coefficient γ from the Dirichlet-to-Neumann map under the assumption that γ has only $3/2 + \epsilon$ derivatives. Previously, the best result required γ to have two derivatives.

Key words. inverse problem, Dirichlet-to-Neumann map, impedance imaging, Besov space

AMS subject classification. 35R30

Let $\Omega \subset \mathbf{R}^n$, $n \geq 3$, be a bounded open set and let $L_\gamma = \operatorname{div} \gamma \nabla$ be an elliptic operator on Ω with scalar coefficient γ. We let Λ_γ denote the Dirichlet-to-Neumann map $\Lambda_\gamma f = \gamma \partial u / \partial \nu$, where u is the solution to the Dirichlet problem $L_\gamma u = 0$ in Ω, $u = f$ on $\partial\Omega$. In 1987, Sylvester and Uhlmann [10] showed that if we restrict attention to γ which are sufficiently smooth, then the map $\gamma \to \Lambda_\gamma$ is injective. Nachman, Sylvester, and Uhlmann [8] showed that injectivity continues to hold if γ has two bounded derivatives. Extensions to slightly less smooth conductivities or the related Schrödinger equation are given in Chanillo [3] and Ramm [9]. Isakov [5] has established injectivity for conductivities with jump discontinuities.

Since the only smoothness assumption needed to define Λ_γ is that γ be measurable, it is reasonable to ask if the map $\gamma \to \Lambda_\gamma$ is injective under less restrictive hypotheses on γ. In this paper, we show that γ need have only $3/2 + \epsilon$ derivatives. There is no reason to believe that the result in this paper is optimal. We conjecture that the right smoothness assumption is that γ have one derivative. However, the methods presented here do not give this. To state our main result, we recall the standard space of Hölder-continuous functions $C^\alpha(\bar\Omega) = \{f : f : \Omega \to \mathbf{R}$ and $|f(x) - f(y)| \leq M|x-y|^\alpha$ for some $M > 0\}$.

THEOREM 1. *Let $\Omega \subset \mathbf{R}^n$, $n \geq 3$, be a bounded, Lipschitz domain. Then the map $\gamma \to \Lambda_\gamma$ is injective on the set $\{\gamma : \gamma > 0$ in $\bar\Omega$, $\nabla\gamma \in \cup_{\epsilon>0} C^{1/2+\epsilon}(\bar\Omega)\}$.*

The outline of our argument is the same as in [10]. We construct special solutions of $L_\gamma u = 0$ by studying a Schrödinger operator $\Delta - q$. The innovation here is that we consider potentials q which lie in a Besov space of negative order.

We begin by recalling the Besov spaces and some of their simple properties. We will use the monograph of Bergh and Löfstrom [2] as our reference for these spaces. For $s \in \mathbf{R}$ and $1 \leq p, q \leq \infty$, we let $B^s_{p,q}$ denote the Besov space of distributions. Roughly speaking, a distribution in $B^s_{p,q}$ has s derivatives in L^p. We recall that if $0 < s < 1$, $1 \leq p, q < \infty$, then $f \in B^s_{p,q}$ if and only if

$$(1) \qquad \|f\|_{L^p} + \left(\int_{\mathbf{R}^n} \left(\int_{\mathbf{R}^n} |f(x+h) - f(x)|^p \, dx \right)^{q/p} |h|^{-n-sq} \, dh \right)^{1/q}$$

is finite. Furthermore, the expression in (1) gives a norm on $B^s_{p,q}$. When $p = q = \infty$,

[*] Received by the editors July 13, 1994; accepted for publication October 13, 1994. This research was supported by NSF grant DMS-9305753 and by the NSF and the Commonwealth of Kentucky through the NSF–EPsCOR.

[†] Department of Mathematics, University of Kentucky, Lexington, KY 40506-0027 (rbrown@ms.uky.edu).

the limiting version of (1) is

$$\|f\|_{L^\infty} + \sup_{x\in\mathbf{R}^n,\ |h|\neq o} |h|^{-s}|f(x+h) - f(x)|.$$

This provides a norm for $B^s_{\infty,\infty}$ and thus, for $0 < s < 1$, $B^s_{\infty,\infty} = C^s(\mathbf{R}^n)$.

We also consider a scale of weighted Besov spaces $B^{s,\delta}_{p,q}$ defined for $\delta \in \mathbf{R}$ by

$$B^{s,\delta}_{p,q} = \{f :\ (1+|x|^2)^{\delta/2} f \in B^s_{p,q}\}$$

with the norm

$$\|f\|_{B^{s,\delta}_{p,q}} = \|(1+|x|^2)^{\delta/2} f\|_{B^s_{p,q}}.$$

We will use $B^{s,c}_{p,q}$ to denote the distributions in $B^s_{p,q}$ which are compactly supported and

$$B^{s,\mathrm{loc}}_{p,q} = \{f :\ \psi f \in B^s_{p,q}\ \text{ for each } \psi \in C^\infty_0(\mathbf{R}^n)\}.$$

We recall that $B^0_{2,2}$ is the usual Lebesgue space L^2 on $\mathbf{R}^n$. If follows that

$$B^{0,\delta}_{2,2} = \{f :\ (1+|x|^2)^{\delta/2} f \in L^2\}$$

is the weighted Lebesgue space L^2_δ used by Sylvester and Uhlmann. We also have that $B^1_{2,2}$ is the Sobolev space of functions having one derivative in L^2 and that

$$B^{1,\delta}_{2,2} = \{f :\ f,\ \ \nabla f \in L^2_\delta\}. \tag{2}$$

Next, we note that since $B^s_{2,2}$ and $B^{s,\delta}_{2,2}$ are isomorphic, we may identify the complex interpolation spaces

$$[B^{s_0,\delta}_{2,2} B^{s_1,\delta}_{2,2}]_\theta = B^{s_\theta,\delta}_{2,2}, \qquad 0 < \theta < 1,$$

where s_0, $s_1 \in \mathbf{R}$, $s_\theta = (1-\theta)s_0 + \theta s_1$ (see [2, Thm. 6.4.5]).

The reason for introducing the Besov spaces to be able to define products of (certain) distributions as bilinear maps between Besov spaces. This depends on the following elementary result regarding multiplication in Besov spaces.

PROPOSITION 2. (a) *If* $\|\phi\|_\infty + \|\nabla\phi\|_\infty \leq M$, *then for* $0 < s < 1$, $\delta \in \mathbf{R}$, $1 \leq p,\ q \leq \infty$,

$$\|\psi u\|_{B^{s,\delta}_{p,q}} \leq C\|u\|_{B^{s,\delta}_{p,q}} M,$$

where $C = C(n,p,q)$.

(b) *For* $0 < s < 1$,

$$\|uv\|_{B^s_{1,2}} \leq C\|u\|_{B^s_{2,2}}\|v\|_{B^s_{2,2}},$$

where $C = C(s,n)$.

We do not prove this proposition, but note that each result follows easily from the norm for $B^s_{p,q}$ given in (1). I thank Mike Frazier for telling me of part (b) of the above proposition.

Next, we give estimates for the operator G_ζ, which is the solution operator to the equation

$$\Delta u + 2\zeta \cdot \nabla u = f,$$

where $\zeta \in \mathbf{C}^n$.

We observe that G_ζ, defined by

$$G_\zeta f = \left(\frac{\hat{f}}{-|\xi|^2 + 2i\zeta \cdot \xi} \right)^{\vee}, \tag{3}$$

maps from $\mathcal{S}$ to $\mathcal{S}'$. Here we are using the Fourier transform defined by $\hat{f}(\xi) = \int_{\mathbf{R}^n} e^{-ix\cdot\xi}\, dx$. In [10], it is shown that if $\zeta \cdot \zeta = 0$, then $G_\zeta : L^2_{\delta+1} \to L^2_\delta$, $-1 < \delta < 0$, with the bound

$$\|G_\zeta f\|_{L^2_\delta} \leq \frac{C}{|\zeta|} \|f\|_{L^2_{\delta+1}}, \qquad |\zeta| > 1 \text{ and } -1 < \delta < 0. \tag{4}$$

We give a simple extension of this result to obtain mapping properties of G_ζ on $B^{s,\delta}_{2,2}$. Shortly before this paper was written, A. Nachman established related estimates for the operator G_ζ in two dimensions [7, Lem. 1.3].

THEOREM 3. *Let $\zeta \in \mathbf{C}^n$ satisfy $\zeta \cdot \zeta = 0$ and $|\zeta| > 1$. Then for $-1 < \delta < 0$ and $0 \leq s \leq 1/2$, the map G_ζ defined by* (3) *satisfies*

$$\|G_\zeta f\|_{B^{s,\delta}_{2,2}} \leq \frac{C}{|\zeta|^{1-2s}} \|f\|_{B^{-s,\delta+1}_{2,2}},$$

where $C = C(n, s, \delta)$.

Proof. We choose a function ϕ satisfying $\phi = 1$ on $\{\xi : |\xi| \leq 4|\zeta|\}$, $\operatorname{supp} \phi \subset \{\xi : |\xi| < 8|\zeta|\}$, and $|\nabla\phi| \leq C/|\zeta|$. For $u \in L^2_\delta$, we define

$$Tu = \nabla[(\phi\hat{u})^{\vee}].$$

We claim that

$$\|Tu\|_{L^2_\delta} \leq C|\zeta|\, \|u\|_{L^2_\delta}, \qquad -1 \leq \delta \leq 1. \tag{5}$$

When $\delta = 0$, this is elementary since T is a multiplier operator whose symbol is bounded by $C|\zeta|$. To obtain (5) when $\delta = 1$, note that

$$\|\hat{u}\|_{L^2} + \|\nabla\hat{u}\|_{L^2}$$

gives an equivalent norm on the weighted Lebesgue space L^2_1. Now

$$\nabla\widehat{Tu} = i\xi\phi\nabla\hat{u} + \hat{u}\nabla(i\xi\phi)$$

and hence

$$\|\nabla\widehat{Tu}\|_{L^2} \leq C(|\zeta| + 1)\, \|u\|_{L^2_1}.$$

If we recall that $|\zeta| \geq 1$, then (5) follows for $\delta = 1$. The estimate (5) follows by duality when $\delta = -1$ and by interpolation for the remaining values of δ. □

Next, define on operator S by

$$(Sf)^\wedge(\xi) = \frac{i\xi(1-\phi)\hat{f}}{(-|\xi|^2 + 2i\xi\cdot\zeta)}$$
$$= \frac{1}{|\xi|}\psi\hat{f},$$

where $\psi(\xi) = i\xi|\xi|(1-\phi)/(-|\xi|^2 + 2i\xi\cdot\zeta)$. The argument used to treat T shows that

$$f \to (\psi\hat{f})^\vee$$

is bounded on L^2_δ, $-1 \le \delta \le 1$, and the norm of this operator is bounded for $|\zeta| \ge 1$.

The fractional integral $f \to (|\xi|^{-1}\hat{f})^\vee$ maps $L^2_{\delta+1}$ to L^2_δ, $-1 < \delta < 0$, by the argument in [10, Lem. 3.1]. This gives

$$\|Sf\|_{L^2_\delta} \le C(n,\delta)\|f\|_{L^2_{\delta+1}}, \qquad -1 < \delta < 0.$$

Summarizing, we have $\nabla G_\zeta f = T(G_\zeta f) + Sf$ and hence

$$\|\nabla G_\zeta f\|_{L^2_\delta} \le C|\zeta|\,\|G_\zeta f\|_{L^2_\delta} + C\|f\|_{L^2_{\delta+1}}$$
$$\le C\|f\|_{L^2_{\delta+1}},$$

where the second inequality is (4).

Combining this with (5) and the characterization of $B^{1,\delta}_{2,2}$ in (2) gives

$$\|G_\zeta f\|_{B^{1,\delta}_{2,2}} \le C\|f\|_{B^{0,\delta+1}_{2,2}}, \qquad -1 < \delta < 0.$$

By duality, we have

$$\|G_\zeta f\|_{B^{0,\delta}_{2,2}} \le C\|f\|_{B^{-1,\delta+1}_{2,2}}, \qquad -1 < \delta < 0.$$

Interpolating between these estimates and (4) gives

$$\|G_\zeta f\|_{B^{s,\delta}_{2,2}} \le \frac{C}{|\zeta|^{1-s}}\|f\|_{B^{0,\delta+1}_{2,2}} \tag{6}$$

and

$$\|G_\zeta f\|_{B^{0,\delta}_{2,2}} \le \frac{C}{|\zeta|^{1-s}}\|f\|_{B^{-s,\delta+1}_{2,2}}, \tag{7}$$

where each inequality holds for $0 \le s \le 1$ and $-1 < \delta < 0$. Finally, interpolating between (6) and (7) gives the estimate of the theorem.

If g is a function on $\mathbf{R}^n$ satisfying

$$\lambda^{-1} < g < \lambda \tag{8}$$

for some $\lambda > 0$ and ∇g is bounded and compactly supported, then for $u \in C^\infty(\mathbf{R}^n)$, we may define a distribution $m_q(u)$ by

$$m_q(u)(v) = -\int_{\mathbf{R}^n} \nabla g \cdot \nabla\left(\frac{1}{g}uv\right)\, dx. \tag{9}$$

Formally, $q = g^{-1}\Delta g$ will be the potential in our Schrödinger operator and $m_q(u)$ is the product qu. Our main result on m_q is the following.

THEOREM 4. *Suppose that g is defined on $\mathbf{R}^n$, satisfies* (8), *and for some s, $0 < s < 1$, and $M > 0$, satisfies*

$$\|\nabla g\|_{B^{1-s}_{\infty,2}} \le M, \tag{10}$$

$$\operatorname{supp} \nabla g \subset \{x : \ |x| < M\}. \tag{11}$$

Then there exists $C = C(M, \lambda, s)$ so that the map m_q satisfies

$$\|m_q(u)\|_{B^{-s,\delta+1}_{2,2}} \le C\|u\|_{B^{s,\delta}_{2,2}}.$$

Before presenting the proof of this theorem, we note that if $\psi \in C_0^\infty(\mathbf{R}^n)$, then

$$\|\psi u\|_{B^s_{p,q}} \le C(\psi, \delta, s, p, q)\|u\|_{B^{s,\delta}_{p,q}}, \tag{12}$$

and if $u \in B^{s,c}_{2,2}$, with $\operatorname{supp} u \subset \{x : \ |x| < R\}$, then

$$\|u\|_{B^{s,\delta}_{2,2}} \le C(R, \delta)\|u\|_{B^s_{2,2}}. \tag{13}$$

In each case, the stated inequality follows by observing that if $\psi \in C_0^\infty(\mathbf{R}^n)$ and $r \in \mathbf{R}$, then $u \to (1 + |x|^2)^r \psi u$ is bounded on each Besov space.

Proof of Theorem 4. We prove the estimate of the theorem for u smooth, and then we may extend m_q to $B^{s,\delta}_{2,2}$ by density. Let $\psi = 1$ on $\operatorname{supp} \nabla g$ with $\psi \in C_0^\infty(\mathbf{R}^n)$. Then we write

$$\begin{aligned} |m_q(u)(\phi)| &= \left| \int \psi \nabla g \cdot \nabla \left(\psi^2 \frac{u\phi}{g} \right) dx \right| \\ &\le \|\psi \nabla g\|_{B^{1-s}_{\infty,2}} \|\nabla(\psi^2 g^{-1} u\phi)\|_{B^{s-1}_{1,2}}. \end{aligned} \tag{14}$$

We use the fact that $\partial/\partial x_i : B^s_{2,2} \to B^{s-1}_{2,2}$, Proposition 2(b), and then (13) to obtain

$$\begin{aligned} \|\nabla(\psi^2 g^{-1} u\phi)\|_{B^{s-1}_{1,2}} &\le C\|\psi^2 u\phi\|_{B^s_{1,2}} \\ &\le C\|\psi u\|_{B^s_{2,2}} \|\psi\phi\|_{B^s_{2,2}} \\ &\le C\|u\|_{B^{s,\delta}_{2,2}} \|\phi\|_{B^{s,-\delta-1}_{2,2}}. \end{aligned} \tag{15}$$

Using (12) and (15) in (14) gives that

$$|m_q(u)(\phi)| \le C\|\nabla g\|_{B^{1-s}_{\infty,2}} \|u\|_{B^{s,\delta}_{2,2}} \|\phi\|_{B^{s,-\delta-1}_{2,2}}$$

or that $m_q(u)$ is in the dual of $B^{s,-\delta-1}_{2,2}$, $(B^{s,-\delta-1}_{2,2})' = B^{-s,\delta+1}_{2,2}$ (see [2, Cor. 6.2.8] for the duals of unweighted Besov spaces). □

Remark. An examination of the above proof shows that in fact we have $m_q : B^{s,\mathrm{loc}}_{2,2} \to B^{s,c}_{2,2}$. We will use this in Corollary 6 to define $m_q(1)$.

Our next theorem considers solutions to the equation

$$\Delta\psi + 2\zeta\cdot\nabla\psi - m_q(\psi) = f.$$

THEOREM 5. *Let g satisfy* (8), (10), *and* (11) *and let $\zeta \in \mathbf{C}^n$ satisfy $\zeta\cdot\zeta = 0$. If $0 < s < 1/2$, $-1 < \delta < 0$, and $f \in B^{-s,\delta+1}_{2,2}$, then there exists $C_0 = C_0(\lambda, M, s, \delta, n)$ so that for $|\zeta| > C_0$, there exists a unique solution to*

$$\Delta\psi + 2\zeta\cdot\nabla\psi - m_q(\psi) = f, \qquad \psi \in B^{s,\delta}_{2,2}, \tag{16}$$

and this solution satisfies

$$\|\psi\|_{B^{s,\delta}_{2,2}} \le \frac{C}{|\zeta|^{1-2s}}\|f\|_{B^{-s,\delta+1}_{2,2}},$$

where $C = C(n, s, \delta, M, \lambda)$.

Proof. Consider the map $\psi \to G_\zeta(m_q(\psi))$. By Theorems 3 and 4, we have

$$\|G_\zeta(m_q(\psi))\|_{B^{s,\delta}_{2,2}} \le \frac{C}{|\zeta|^{1-s}}\|\psi\|_{B^{s,\delta}_{2,2}}.$$

Hence, if $|\zeta|$ is sufficiently large, then this map is a contraction on $B^{s,\delta}_{2,2}$.

From the uniqueness of solutions to $\Delta\psi + 2\zeta\cdot\nabla\psi = 0$, $\psi \in L^2_\delta$ (see [10, Cor. 3.4], [4, Thm. 7.1.27]), ψ satisfies (16) if and only if

$$\psi = G_\zeta(f) + G_\zeta(m_q(\psi)), \tag{17}$$

and by Theorem 3, $G_\zeta(f) \in B^{s,\delta}_{2,2}$. Thus the contraction-mapping principle implies solutions to (17) exist and are unique in $B^{s,\delta}_{2,2}$. □

Now we are ready to return to the study of $L_\gamma = \operatorname{div}\gamma\nabla$. It will be convenient to assume that γ is defined in all of $\mathbf{R}^n$ and satisfies for some $1/2 > \epsilon > 0$ and $R > 0$, $\lambda > 1$,

$$\lambda^{-1} < \gamma < \lambda, \tag{18}$$

$$\nabla\gamma \in B^{1/2+2\epsilon,c}_{\infty,\infty} \subset B^{1/2+\epsilon,c}_{\infty,2} \quad \text{for some } \epsilon > 0, \tag{19}$$

$$\gamma(x) = 1 \ \text{ if } \ |x| > R \tag{20}$$

The embedding in (19) follows easily from the definition of the $B^s_{p,q}$-norm [2, Def. 6.2.2]. Thus if $g = \sqrt{\gamma}$, g satisfies the hypotheses of Theorem 5 with $s = 1/2 - \epsilon$.

COROLLARY 6. *Suppose that γ satisfies* (18)–(20) *and $\zeta \in \mathbf{C}^n$ satisfies $\zeta\cdot\zeta = 0$ and $|\zeta| > C_0$. Then there exists a solution to $L_\gamma u = 0$ of the form*

$$u(x) = \gamma(x)^{-1/2}(1 + \psi(x))e^{x\cdot\zeta}, \qquad \psi \in B^{1/2-\epsilon,\delta}_{2,2}.$$

Furthermore, $D^2u \in L^2_{\mathrm{loc}}(\mathbf{R}^n)$.

Proof. Given γ, we construct m_q as in (9), with $g = \sqrt{\gamma}$. We let ψ be the solution of

$$\Delta\psi + 2\zeta\cdot\nabla\psi - m_q(\psi) = m_q(1)$$

from Theorem 5. Since $v = e^{x\cdot\zeta}(1+\psi)$ solves $\Delta v - m_q(v) = 0$ in $\mathcal{S}'$ and $m_q(v) \in B^{\epsilon-1/2,c}_{2,2}$, regularity theory for Δ implies $v \in B^{\epsilon+3/2,\text{loc}}_{2,2}$ and, in particular, $\nabla v \in L^2_{\text{loc}}$. Then a calculation shows that $u = \gamma^{-1/2}v$ solves $L_\gamma u = 0$, $\nabla u \in L^2_{\text{loc}}$. Finally, since γ is C^1, regularity theory for L_γ implies $\nabla^2 u \in L^2_{\text{loc}}$. □

THEOREM 7. *Suppose that $\partial\Omega$ is Lipschitz and $\Lambda_{\gamma_1} = \Lambda_{\gamma_2}$. If $\nabla\gamma_i \in C^{1/2+2\epsilon}(\bar\Omega)$ for some $\epsilon > 0$, then there exist extensions of γ_i to $\mathbf{R}^n$ so that, with $g_i = \sqrt{\gamma_i}$,*

$$\int_{\mathbf{R}^n} \nabla g_1 \cdot \nabla(g_1^{-1}\phi) = \int_{\mathbf{R}^n} \nabla g_2 \cdot \nabla(g_2^{-1}\phi), \quad \phi \in C_0^\infty(\mathbf{R}^n).$$

Proof. We begin by observing that since $\Lambda_{\gamma_1} = \Lambda_{\gamma_2}$ and $\partial\Omega$ is Lipschitz, we have $\gamma_1 = \gamma_2$ and $\nabla\gamma_1 = \nabla\gamma_2$ on $\partial\Omega$. This result was proven for smooth conductivities in [6] and for C^1 conductivities in Lipschitz domains by [1]. Thus we may extend γ_1 and γ_2 to $\mathbf{R}^n$ so that $\gamma_1 = \gamma_2$ in $\mathbf{R}^n \setminus \bar\Omega$ and satisfies (18)–(20).

We let u_1 and u_2, be solutions of $L_{\gamma_i}\, u_i = 0$, $\nabla u_i \in L^2(\Omega)$, $i = 1,2$. We let $v_i = \gamma_i^{1/2}u_i$ and obtain

$$\begin{aligned}\int_{\partial\Omega} u_2\Lambda_{\gamma_1}u_1 &= \int_\Omega \gamma_1\nabla(\gamma_1^{-1/2}v_1)\cdot\nabla(\gamma_1^{-1/2}v_2)\,dx \\ &= \int_\Omega -\nabla\gamma_1^{1/2}\cdot\nabla(\gamma_1^{-1/2}v_1v_2) + \nabla v_1\cdot\nabla v_2\,dx,\end{aligned}$$

where have used the facts that $\gamma_1 = \gamma_2$ on $\partial\Omega$ and the second equality depends on the product rule.

Reversing the roles of u_1 and u_2 gives

$$\int_{\partial\Omega} u_1\Lambda_{\gamma_2}u_2 = \int_\Omega -\nabla\gamma_2^{1/2}\cdot\nabla(\gamma_2^{-1/2}v_1v_2) + \nabla v_1\cdot\nabla v_2\,dx.$$

If we subtract these expressions and use the fact that Λ_{γ_2} is a symmetric operator, we have

$$\begin{aligned}\int_{\partial\Omega} u_1(\Lambda_{\gamma_1} - \Lambda_{\gamma_2})u_2 &= \int_\Omega -\nabla\gamma_1^{1/2}\cdot\nabla(\gamma_1^{-1/2}v_1v_2) \\ &\qquad + \nabla\gamma_2^{1/2}\cdot\nabla(\gamma_1^{-1/2}v_1v_2)\,dx.\end{aligned}\tag{21}$$

If we assume that u_1 and u_2 are defined in all of $\mathbf{R}^n$, then (21), our assumptions that $\Lambda_{\gamma_1} = \Lambda_{\gamma_2}$ and that $\gamma_1 = \gamma_2$ in $\mathbf{R}^n \setminus \bar\Omega$ give

$$0 = \int_{\mathbf{R}^n} -\nabla\gamma_1^{1/2}\cdot\nabla(\gamma_1^{-1/2}v_1v_2) + \nabla\gamma_2^{1/2}\cdot\nabla(\gamma_2^{-1/2}v_1v_2)\,dx.$$

To choose u_1 and u_2, we fix $k \in \mathbf{R}^n$ and then note that the argument in [10, p. 157] and the estimate of Theorem 5 allow us to construct sequences $u_1^{(n)}, u_2^{(n)}$ so that $L_{\gamma_i}u_i^{(n)} = 0$ and $v_1^{(n)}\cdot v_2^{(n)} \to e^{ix\cdot k}$ in $B^{1/2-\epsilon,\text{loc}}_{1,2}$ as $n \to \infty$. Hence we conclude that

$$\int \nabla\gamma_1^{1/2}\cdot\nabla(\gamma_1^{-1/2}e^{ix\cdot k}) = \int \nabla\gamma_2^{1/2}\cdot\nabla(\gamma_2^{-1/2}e^{ix\cdot k}), \qquad k \in \mathbf{R}^n.$$

This implies the conclusion of the theorem. □

PROPOSITION 8. *If the conclusion of Theorem* 7 *holds, then* $g_1 = g_2$.

Proof. We have

$$\int \nabla g_1 \cdot \nabla \left(\frac{1}{g_1}\phi\right) = \int \nabla g_2 \cdot \nabla \left(\frac{1}{g_2}\phi\right)$$

for all $\phi \in C_0^1(\mathbf{R}^n)$. Replace ϕ by $g_1 g_2 \psi$ and observe that this gives

$$\int g_1 g_2 \nabla(\log g_1 - \log g_2) \cdot \nabla\psi = 0.$$

In particular, if $\psi = \log \frac{g_1}{g_2}$, then $\log \frac{g_1}{g_2} = 0$ in $\mathbf{R}^n$. □

This proposition amounts to observing that the equality $g_1^{-1}\Delta g_1 = g_2^{-1}\Delta g_2$ implies that $\operatorname{div} g_1 g_2 \nabla \log(g_1/g_2) = 0$. I thank J. Tolle for showing me this argument, which is due to G. Alessandrini.

Proof of Theorem 1. Suppose that γ_1 and γ_2 are as in the Theorem and that $\Lambda_{\gamma_1} = \Lambda_{\gamma_2}$. Then we conclude that $\sqrt{\gamma}_1 = \sqrt{\gamma}_2$ from Theorem 7 and Proposition 8. □

REFERENCES

[1] G. ALESSANDRINI, *Singular solutions of elliptic equations and the determination of conductivity by boundary measurements*, J. Differential Equations, 84 (1990), pp. 252–272.

[2] J. BERGH AND J. LÖFSTRÖM, *Interpolation Spaces: An Introduction*, Springer-Verlag, Berlin, New York, Heidelberg, 1976.

[3] S. CHANILLO, *A problem in electrical prospection and an n-dimensional Borg–Levinson theorem*, Proc. Amer. Math. Soc., 108 (1990), pp. 761–767.

[4] L. HÖRMANDER, *The Analysis of Linear Partial Differential Operators* I, Springer-Verlag, Berlin, New York, Heidelberg, 1983.

[5] V. ISAKOV, *On uniqueness of recovery of a discontinuous conductivity coefficient*, Comm. Pure Appl. Math., 41 (1988), pp. 865–977.

[6] R. KOHN AND M. VOGELIUS, *Determining conductivity by boundary measurements*, Comm. Pure Appl. Math., 37 (1984), pp. 289–298.

[7] A. I. NACHMAN, *Global uniqueness for a two-dimensional inverse boundary value problem*, Ann. of Math., 143 (1996), pp. 71–96.

[8] A. I. NACHMAN, J. SYLVESTER, AND G. UHLMANN, *An n-dimensional Borg–Levinson theorem*, Comm. Math. Phys., 115 (1988), pp. 595–605.

[9] A. G. RAMM, *Completeness of products of solutions to pde and uniqueness theorems in inverse scattering*, Inverse Problems, 3 (1987), pp. L77–L82.

[10] J. SYLVESTER AND G. UHLMANN, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math., 125 (1987), pp. 153–169.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 1057–1069, July 1996

011

LORENZ EQUATIONS PART I: EXISTENCE AND NONEXISTENCE OF HOMOCLINIC ORBITS*

XINFU CHEN†

Abstract. The Lorenz equations are a system of three ordinary differential equations

$$x' = s(y - x), \quad y' = Rx - y - xz, \quad z' = xy - qz,$$

where s, R, and q are positive parameters. We show that this system has homoclinic orbits associated with the origin (i.e., orbits that tend to the origin as $t \to \pm\infty$) if and only if $s > (2q + 1)/3$. The method is based on Liapunov functions and a shooting argument used previously by Hastings and Troy in studying homoclinic orbits of the Lorenz equations.

Key words. Lorenz equations, homoclinic orbits, shooting methods, Liapunov functions, asymptotic behavior

AMS subject classification. 34C37

1. Introduction.

The Lorenz equations we studied here are a system of ordinary differential equations

$$\begin{cases} x' = s(y - x), \\ y' = Rx - y - xz, \\ z' = xy - qz, \end{cases} \tag{1.1}$$

where $' = \frac{d}{dt}$ and s, R, and q are positive parameters. This system was first presented in 1963 by E. N. Lorenz [11] in studying fluid convection in a two-dimensional layer heated from below. In the last decades, there has been an immense amount of interest generated by these equations due to the fact that for some parameter values, numerical computed solutions oscillate in the pseudorandom way which people call "chaotic." For more detailed description of the observed "chaotic" behavior and mathematical theories built upon (1.1), such as the geometric models of the Lorenz equations, Sil'nikov-type or homoclinic bifurcations, and averaging methods, see [3], [5], [13]–[16], [18], a review book of Sparrow [17], a mathematical textbook of Guckenheimer and Holmes [4], and the references therein.

Generally speaking, to apply certain well-developed theories to the concrete example of (1.1), certain hypotheses have to be verified. One way to achieve this would be by a computer simulation, but there are few rigorous results. Recently, Hastings and Troy [8], [9], Hassard, Hastings, Troy, and Zhang [7], and Mischaikow and Mrozek [12] built up mathematical theories for certain characterizations of the chaotic behavior of (1.1) and implemented rigorous arithmetic numerical schemes to verify the validity of their hypotheses, and hence lead to affirmative conclusions on the chaotic behavior they studied for the solutions of (1.1) for certain parameter values (s, R, q).

Among all the solutions of (1.1), a very special one is a homoclinic orbit associated with the origin, which is a solution having the property that it approaches the origin as $t \to \pm\infty$. In developing geometric model theories or bifurcation theories, homoclinic orbits play an essential role. An example of its fascination and importance can be read from the phase "homoclinic explosion," which is used to refer to the appearance

* Received by the editors March 9, 1994; accepted for publication (in revised form) November 8, 1994. This research was partially supported by National Science Foundation grant DMS-9200459 and the Alfred P. Sloan Research Fellowship.

† Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260.

of various kinds of chaotic behavior when parameters are perturbed from the values where there is a homoclinic orbit; see, for example, Sparrow [17] and Robinson [14]. Therefore, it is important to find these homoclinic orbits. The existence of homoclinic orbits associated with the origin for (1.1) can be seen from strong numerical evidence. The first pure mathematical proof is given by Hastings and Troy [10]. Using a shooting argument and a pioneering, though tedious, mathematical analysis on the nature of the solution when $(s, R, q) = (10, 1000, 1)$, they were able to show that for each (s, q) in some neighborhood of the point $(10, 1)$ there is an R in the interval $(1, 1000)$ such that the Lorenz equations have a homoclinic orbit associated with the origin. A rigorous numerical implementation of a similar method for the existence of a homoclinic orbit was used by Hassard and Zhang [6] to pin down the value of the parameter R. They showed that when $s = 10$ and $q = 8/3$, R is between 13.9265 and 13.927.

In this paper, we shall study the existence and nonexistence of homoclinic orbits of (1.1). In particular we prove the following theorem.

THEOREM 1.1. *Assume that s and q are given positive constants. Then the Lorenz equations* (1.1) *have at least one homoclinic orbit associated with the origin for some $R \in (0, \infty)$ if and only if $s > \frac{2q+1}{3}$.*

For the existence part, the proof is based on a shooting argument developed in [10], where the shooting parameter is R. In completing this shooting argument, one of the essential difficulties lies in the study of the behavior of the solution for large R. Impressed by the magnitude of the value $R = 1000$ taken in [10], here we replace their analysis for $R = 1000$ by studying the asymptotic behavior of the solution as $R \to \infty$. Though substantial information, compared to that obtained in [10] for $R = 1000$, was lost, the key properties needed to complete the shooting argument were not. The main advantage of our method is that we can establish the existence of homoclinic orbits for a large set of values of (s, q).

The proof of nonexistence is based solely on Liapunov functions. Fortunately, the result obtained compliments perfectly the existence result, as is seen from Theorem 1.1.

We shall prove the nonexistence part of Theorem 1.1 in §2 and the existence part in §§3 and 4.

Remark 1.2. As will be seen in the proof of Theorem 1.1, the homoclinic orbit established in this paper is the simplest one among all homoclinic orbits the Lorenz system could have; namely, its x coordinate does not change sign and has only one local extreme, so that in the x–x' phase plane, the orbit forms a single loop on the right or the left half plane.

In [1], we prove the existence of homoclinic orbits having the property that the x coordinate can change any prescribed number of times, and in each time interval where x does not change sign, x' can change an arbitrarily prescribed odd number of times (except in the first two intervals). Also we prove the existence of orbits having the property that the x coordinate changes sign infinitely many times and during each time interval where x does not change, x' changes a certain number of signs according to a prescribed double-bounded odd integer sequence.

In forthcoming paper [2], we shall prove the existence of homoclinic explosion for the Lorenz system for certain parameters; namely, we shall prove the existence of a horseshoelike Poincaré map.

2. Nonexistence of homoclinic orbits.

First, we recall a boundedness lemma originated by Lorenz [11].

LEMMA 2.1. *The region* $D := \{(x,y,z) \mid x^2 + \frac{1}{R}y^2 + \frac{1}{R}(z-(1+s)R)^2 \leq \frac{qR(1+s)^2+1}{\min\{2s,2,q\}}\}$ *is positively invariant; that is, if a trajectory enters* D *at some time, it will stay there forever. In addition, every trajectory will enter* D *in finite time.*

Proof. Set $V = x^2 + \frac{1}{R}y^2 + \frac{1}{R}(z-(1+s)R)^2 - \frac{qR(1+s)^2+1}{\min\{2s,2,q\}}$. Then along any trajectory of (1.1),

$$V' = -2sx^2 - \tfrac{2}{R}y^2 - \tfrac{q}{R}(z-(1+s)R)^2 - \tfrac{q}{R}z^2 + qR(1+s)^2 \leq -\min\{2s,2,q\}V - 1 - \tfrac{q}{R}z^2.$$

The assertion of the lemma thus follows. □

The following lemma shows that the case $R \in (0,1]$ is trivial.

LEMMA 2.2. *Assume that* $R \in (0,1]$. *Then for every positive* s *and* q, *every solution of* (1.1) *approaches, as* $t \to \infty$, $(0,0,0)$ *when* $R \in (0,1)$ *or* $(x^*, x^*, 0)$ *for some* $x^* \in (-\infty, \infty)$ *when* $R = 1$. *Consequently, there are no homoclinic orbits associated with any stationary points of* (1.1).

Proof. Define $V(x,y,z) = x^2 + sy^2 + sz^2$. Since $R \in (0,1]$, along any trajectory of (1.1),

$$V' = -2s\Big\{[x - \tfrac{1+R}{2}y]^2 + [1 - (\tfrac{1+R}{2})^2]y^2 + qz^2\Big\} \leq 0,$$

where the equal sign is taken only at the stationary points of the Lorenz system. The assertion of the lemma thus follows from a standard theory of Liapunov functions. □

In what follows, we shall always assume that

$$s > 0, \qquad q > 0, \qquad R > 1.$$

In this case, (1.1) has three stationary points located at

$$C^0 = (0,0,0), \qquad C_R^{\pm} = \left(\pm\sqrt{q(R-1)}, \sqrt{q(R-1)}, R-1\right).$$

When $R > 1$, the origin C^0 is nonstable and the linearized flow near the origin has three real eigenvalues,

$$(2.1) \quad \lambda_i^0 = \tfrac{1}{2}\Big\{-(s+1) - (-1)^i\sqrt{(s+1)^2 + 4s(R-1)}\Big\}\ (i = 1,2), \quad \lambda_3^0 = -q.$$

The eigenvectors corresponding to λ_1^0 and λ_2^0 lie on the x–y plane, whereas the eigenvalue λ_3 corresponds to the two trajectories being the positive and negative z-axis. Since λ_1^0 is positive and λ_2^0 and λ_3^0 are negative, at C^0, there is a two-dimensional stable manifold which contains the z-axis, and a one-dimensional unstable manifold $\gamma = \gamma^+ \cup \gamma^- \cup C^0$, where γ^- is the reflection of γ^+ across the z-axis and γ^+ initially points into the positive octant. In fact, the trace of γ^+ is a smooth curve starting from the origin having the asymptotic behavior $\gamma = (x, (1+\frac{\lambda_1^0}{s})x + O(x^2), \frac{\lambda_1^0+s}{s(2\lambda_1^0+q)}x^2 + O(x^3))$ for small positive x.

When $R > 1$, the characteristic function for the eigenvalues of the linearized flow near the stationary point C^+ and C^- is

$$\lambda^3 + (1+s+q)\lambda^2 + (s+R)q\lambda + 2sq(R-1) = 0.$$

This equation has a negative root, and, when the real part of the other two roots (which are complex conjugate to each other) are nonnegative, its imaginary part is nonzero. That is, either $C^{\pm}$ is stable or $C^{\pm}$ has a one-dimensional stable manifold and a two-dimensional unstable manifold on which the flow is of spiral type.

It is convenient to introduce

$$Q := z - \tfrac{1}{2s}x^2, \qquad \mu = 1 - \tfrac{q}{2s}.$$

Then, one can directly verify that the Lorenz equations (1.1) are equivalent to the following system:

$$(2.2) \qquad \begin{cases} x'' + (1+s)x' = s\Big\{(R-1) - Q - \tfrac{1}{2s}x^2\Big\}x \quad (= s(R-1-z)x), \\ Q' + qQ = \mu x^2. \end{cases}$$

To show that (2.2) has no homoclinic orbits associated with the origin when $s \le \frac{2q+1}{3}$, we considered three cases:

$$(1)\ 0 < s < \tfrac{q}{2}, \qquad (2)\ s = \tfrac{q}{2}, \qquad (3)\ \tfrac{q}{2} < s \le \tfrac{2q+1}{3}.$$

First we consider the case $0 < s < \frac{q}{2}$.

LEMMA 2.3. *Assume that* $0 < s < \frac{q}{2}$. *Then for any* $R > 1$, (1.1) *has no homoclinic orbits associated with any stationary points. In addition, as* $t \to \infty$, *every trajectory has a limit that is one of the stationary points of* (1.1).

Proof. Define $V = x'^2 + \frac{s}{2|\mu|q}Q'^2 + \frac{s}{2q}(x^2 - (R-1)q)^2$. Since $\mu = 1 - \frac{q}{2s} \le 0$, substituting Q' by $-qQ + \mu x^2$ yields

$$V = x'^2 + sx^2\Big[\tfrac{1}{4s}x^2 + Q - (R-1)\Big] - \tfrac{sq}{2\mu}Q^2 + \tfrac{s(R-1)^2 q}{2}.$$

It then follows by using the differential equations in (2.2) that along any trajectory,

$$\begin{aligned} V' &= 2x'\Big\{x'' + sx\Big[\tfrac{1}{2s}x^2 + Q - (R-1)\Big]\Big\} + \frac{s}{\mu}Q'\Big\{\mu x^2 - qQ\Big\} \\ &= -2(1+s)x'^2 + \frac{s}{\mu}Q'^2. \end{aligned}$$

Since $\mu < 0$, $V' \le 0$ for all $t \in (-\infty, \infty)$ and in addition, $V' = 0$ if and only if $x' = 0$ and $Q' = 0$. The assertions of the Lemma thus follow by the properties of the Liapunov function V. □

Next we consider the case $s = \frac{q}{2} > 0$.

LEMMA 2.4. *Assume that* $s = \frac{q}{2} > 0$. *Then the assertion of Lemma* 2.2 *holds.*

Proof. Note that when $s = \frac{q}{2}$, $\mu = 0$, so that $Q' = -qQ$. Consequently, $Q(t) = Q(0)e^{-qt}$. In particular, on γ^+, $Q(t) \equiv 0$ so that $x'' + (1+s)x' = sx[(R-1) - \frac{1}{2s}x^2]$. Set $V = x'^2 + \frac{1}{4}[x^2 - 2s(R-1)]^2$. Then $V' = -2(1+s)x'^2 \le 0$ and therefore γ^+ approaches C^+ as $t \to \infty$. (Note that the set $\{(x,y,z) | V < 0\}$ consists of two disjoint domains with C^+ and C^- in each of them.)

For any other trajectories, $Q(t) \to 0$ as $t \to \infty$, and one can show that every trajectory has to tend to one of the stationary points. Details are omitted. □

Finally, we consider the case $\frac{q}{2} < s \le \frac{2q+1}{3}$.

LEMMA 2.5. *Assume that* $0 < \frac{q}{2} < s \le \frac{2q+1}{3}$. *Then for any* $R > 1$, *the Lorenz equations* (1.1) *have no homoclinic orbits associated with the origin.*

Proof. We need only show that $\gamma^{\pm}$ does not approach the origin as $t \to \infty$. By symmetry, we need only consider γ^+.

First, notice that $\mu = 1 - \frac{q}{2s} > 0$, so that $Q' + qQ = \mu x^2 \geq 0$. It then follows that on γ^+, $Q(t) > 0$ for all t.

Define

$$V = x'^2 + sx^2\Big\{\tfrac{1}{4s}x^2 + Q - (R-1)\Big\} + (1+s)xx'.$$

Then, along γ^+,

$$\begin{aligned} V' &= 2x'\Big\{x'' + sx[\tfrac{1}{2s}x^2 + Q - (R-1)]\Big\} + sx^2Q' + (1+s)x'^2 + (1+s)xx'' \\ &= 2x'\Big\{-(1+s)x'\Big\} + sx^2[-qQ + \mu x^2] + (1+s)x'^2 \\ &\qquad + (1+s)x\Big\{-(1+s)x' + s[(R-1) - Q - \tfrac{1}{4s}x^2 - \tfrac{1}{4s}x^2]\Big\} \\ &= -(1+s)\Big\{x'^2 + (1+s)xx' + sx^2[\tfrac{1}{4s}x^2 + Q - (R-1)]\Big\} + \Big(\mu s - \tfrac{1+s}{4}\Big)x^4 - sqQx^2 \\ &= -(1+s)V + \tfrac{3}{4}\Big(s - \tfrac{2q+1}{3}\Big)x^4 - sqQx^2, \end{aligned}$$

where we have used the equations in (2.2) in the second equality, and the definition of V and the relation $\mu s - \frac{1+s}{4} = (1 - \frac{q}{2s})s - \frac{1+s}{4} = \frac{3}{4}(s - \frac{2q+1}{3})$ in the last equality.

Since $s \leq \frac{2q+1}{3}$ and $Q > 0$ on $\gamma^{\pm}$, the last equation yields

$$V'(t) + (1+s)V(t) \leq 0 \qquad \text{on } \gamma^+ \text{ for all } t \in (-\infty, \infty).$$

In particular, since $V(-\infty) = 0$, we have that $V(t) \leq 0$ for all t. Since Gronwall's inequality implies that $V(t) \leq V(0)e^{-(1+s)t}$ for all $t > 0$, we have that $|V(t)| \geq |V(0)|e^{-(1+s)t}$ for all $t > 0$.

We now claim that γ^+ cannot approach $C^0 = (0,0,0)$ as $t \to \infty$. In fact, if a trajectory is on the stable manifold of the origin and is not the z-axis, then for t sufficiently large, $x'' = -(1+s)x' + (R - 1 + o(1))x$, so that $x, x', x'' = O(e^{[\lambda_2^0 + o(1)]t})$ as $t \to \infty$. That is, if γ^+ approaches the origin as $t \to \infty$, then $V(t) = O(e^{2[\lambda_2 + o(1)]t})$. But this is impossible since $|V(t)| \geq |V(0)|e^{-(1+s)t}$ for all $t > 0$ and, from the expression of λ_2^0 in (2.1), $\lambda_2^0 < -(1+s)$. Therefore, γ^+ cannot approach the origin; namely, there is no homoclinic orbit associated with the origin. □

Summarizing Lemmas 2.2–2.5, we now can conclude the following theorem.

THEOREM 2.6. *Assume that* $0 < s \leq \frac{2q+1}{3}$. *Then for any* $R \in (0, \infty)$, *the Lorenz equations* (1.1) *have no homoclinic orbits associated with the origin.*

Remark. Lemmas 2.3 and 2.4 show that when $R > 1$ and $s \leq \frac{q}{2}$, $\gamma^{\pm}$ approaches $C^{\pm}$ as $t \to \infty$ and every other trajectory of (1.1) approaches one of the stationary points. However, we are unable to show the same conclusion for the case $\frac{q}{2} < s \leq \frac{2q+1}{3}$.

3. Existence of homoclinic orbits associated with the origin.

We now begin to show the existence part of Theorem 1.1; namely, we prove the following theorem.

THEOREM 3.1. *Assume that* $q > 0$ *and* $s > \frac{2q+1}{3}$. *Then there exists* $R \in (1 + \frac{q}{2s}, \infty)$ *such that the Lorenz equations* (1.1) *have a homoclinic orbit associated with the origin and lying on the half space* $\{(x,y,z) | x > 0\}$.

Clearly, Theorem 1.1 follows from Theorems 2.6 and 3.1.

In the sequel, we shall always assume that s, R, and q satisfy

$$q > 0, \qquad s > \frac{2q+1}{3}, \qquad R > 1.$$

Notice that the first two conditions imply that $\mu = 1 - \frac{q}{2s} > 0$.

The proof is based on a shooting argument first used by Hastings and Troy [10] and later used by Hassard and Zhang [6]; here R is the shooting parameter. In what follows, we shall use γ_R^+ to denote the trajectory of the stable manifold associated with the origin which initially enters the positive octant. Also we use x_R, y_R, z_R, Q_R to denote the solution with the trajectory γ_R^+. To set up the shooting argument, we define a set $\mathfrak{R}$ as follows:

$$\mathfrak{R} = \left\{ R > 1 \;\middle|\; \begin{array}{l} \text{(i) } \exists\, T_R^1 \text{ such that } x_R'(T_R^1) = 0,\ x_R''(T_R^1) < 0, \text{ and } x_R' > 0 \text{ in } (-\infty, T_R^1); \\ \text{(ii) } \exists\, T_R^2 > T_R^1 \text{ such that } x_R'(T_R^2) = -sx(T_R^2) \text{ and } x_R'' < -sx_R' \text{ in } [T_R^1, T_R^2]; \\ \text{(iii) } \exists\, T_R^3 > T_R^2 \text{ such that } x_R(T_R^3) = 0 \text{ and } x_R' < -sx_R \text{ in } (T_R^2, T_R^3]. \end{array} \right\}$$

Observe that $x_R' + sx_R = sy_R$, so that the second condition in the definition means that $y_R(T_R^2) = 0$ and $y_R' < 0$ in $[T_R^1, T_R^2]$, and the third condition means that $y_R < 0$ in $(T_R^2, T_R^3]$.

LEMMA 3.2. *The set $\mathfrak{R}$ is open.*

Proof. Assume that $\bar{R} \in \mathfrak{R}$. Since $x_{\bar{R}}''(T_{\bar{R}}^1) \neq 0$, $(x_{\bar{R}}' + sx_{\bar{R}})'(T_{\bar{R}}^2) \neq 0$, and $x_{\bar{R}}'(T_{\bar{R}}^3) \neq 0$, by the continuity of the trajectory γ_R^+ with respect to R and the implicit-function theorem, for all R sufficiently close to $\bar{R}$, there exist T_R^1, T_R^2, and T_R^3 satisfying the three conditions in the definition of $\mathfrak{R}$; that is, $R \in \mathfrak{R}$. Hence, $\mathfrak{R}$ is open. □

LEMMA 3.3. *Assume that $R \in (1, 1 + \frac{q}{2s}]$. Then the "box" $B := \{(x,y,z) \mid 0 < x < \sqrt{q}, 0 < y < \sqrt{q}, \frac{1}{2s}x^2 < z < 1\}$ is positively invariant. Since initially γ_R^+ stays in B, γ_R^+ stays in B for all $t \in (-\infty, \infty)$. Consequently, $R \notin \mathfrak{R}$ and there are no homoclinic orbits associated with the origin.*

Proof. The argument given below is adapted from [10].

To show that B is positive invariant, we need only check the direction of the vector field of (1.1) on the boundary of B.

On the bottom face of B, $Q = 0$. Since $Q' + qQ = \mu x^2 \geq 0$, we know that once Q is positive, it will remain positive. It then follows that no trajectory can exit B from the face $\{z = \frac{1}{2s}x^2\}$.

On the top face of B, $z = 1$. It follows that $z' = xy - qz = xy - q < 0$ except at $(x,y) = (\sqrt{q}, \sqrt{q})$. When $(x,y,z) = (\sqrt{q}, \sqrt{q}, 1)$, $y' = (R-2)\sqrt{q} < 0, z' = 0, z'' = \sqrt{q}y' < 0, x' = 0, x'' = (R-2)\sqrt{q} < 0$. Hence, no trajectory can exit from the top face of B. Similarly, one can show that no trajectory can exit B from the left face $\{x = 0\}$, the right face $\{x = \sqrt{q}\}$, and the front face $\{y = 0\}$ of B. It remains to consider the back face (without edges) of B.

On the back face $\{y = \sqrt{q}, 0 < x < \sqrt{q}, \frac{1}{2s}x^2 < z < 1\}$ of B, $y' = (R-z)x - \sqrt{q}$. Since $R \leq 1 + \frac{q}{2s}$ and $z > \frac{1}{2s}x^2$, $y' < x(1 + \frac{q}{2s} - \frac{1}{2s}x^2) - \sqrt{q} = (\sqrt{q} - x)(\frac{x(x+\sqrt{q})}{2s} - 1) \leq (\sqrt{q} - x)(\frac{q}{s} - 1) \leq 0$. We then also conclude that no trajectory can exit from B on the face $\{y = \sqrt{q}\}$.

In conclusion, B is positive invariant and $R \notin \mathfrak{R}$. Since an orbit cannot enter the origin from the inside of B, there are no homoclinic orbits associated with the origin. This completes the proof of the lemma. □

LEMMA 3.4. *There exists $R_0 \gg 1$ such that $[R_0, \infty) \in \mathfrak{R}$.*

This is a technical lemma, and we shall leave the proof to the next section. In fact, this is the major difference of our proof from that of Hastings and Troy [10] and Hassard and Zhang [6].

Assume for the moment that Lemma 3.4 has been proven. We can now complete the proof of Theorem 3.1. The argument below is essentially the same as that in [10] and [6]. Here we provide it for completeness, as well as for the reader's convenience.

Proof of Theorem 3.1. Define $R^* \equiv \inf\{R \mid R \in \mathfrak{R}\}$. From Lemmas 3.3 and 3.4, $R^* \in [1+\frac{q}{2s},\infty)$, so that $\gamma^+_{R^*}$ is well defined. Since $\mathfrak{R}$ is open, $R^* \notin \mathfrak{R}$. We now show that $\gamma^+_{R^*}$ is a homoclinic orbit, via the following steps:

1. there exists $T^1_{R^*}$ satisfying the first condition in the definition of $\mathfrak{R}$; namely, $x_{R^*} > 0$ in $(-\infty, T^1_{R^*})$, $x'_{R^*}(T^1_{R^*}) = 0$, and $x''_{R^*}(T^1_{R^*}) < 0$;
2. there exists $T^2_{R^*}$ satisfying the second condition in the definition of $\mathfrak{R}$; namely, $(x'_{R^*} + sx_{R^*})' < 0$ on $[T^1_{R^*}, T^2_{R^*}]$ and $(x_{R^*} + sx_{R^*})(T^2_{R^*}) = 0$;
3. $x_{R^*} > 0$ for all $t \in (-\infty,\infty)$, $x'_{R^*} < -sx_{R^*}$ for all $t \in (T^2_{R^*},\infty)$, and $\gamma^+_{R^*}$ approaches the origin as $t \to \infty$.

The first step. Since γ_{R^*} is on the unstable manifold of the origin, $x_{R^*} > 0$ and $x'_{R^*} > 0$ for all t sufficiently negative large. We claim that there is a first time such that $x'_{R^*} = 0$. In fact, if it is not true, then $x'_{R^*} > 0$ for all t, which implies that as $t \to \infty$, γ_{R^*} approaches the stationary point $C^+_{R^*} \equiv (\sqrt{q(R^*-1)}, \sqrt{q(R^*-1)}, R^*-1)$. (Recall that all the trajectories are bounded for positive t.) There are only two possibilities: (i) C_{R^*} is stable; (ii) C_{R^*} is nonstable with a two-dimensional unstable manifold on which the flow is of spiral type. In the first case, there is a small ball centered at C_{R^*} such that the vector fields of (1.1) intersect the boundary of the ball nontangentially and inwards. Consequently, when R is sufficiently close to R^*, the vector fields with the parameter R also intersect the boundary of the ball nontangentially and inwards, so that the ball is positively invariant. Since for all R sufficiently close to R^*, γ^+_R is close to $\gamma^+_{R^*}$ up to the time when γ^+_R enters the ball, which means that γ^+_R will be trapped into the ball before x_R reaches its first zero. Clearly, this violates the definition of R^*. In the second case, there are two small balls B_1 and B_2, both centered at C_{R^*} and $B_1 \subset B_2$, such that any trajectory starting from any point on the boundary of B_1 has to wind around $\gamma^+_{R^*}$ at least two times before it could possibly exit from the boundary of B_2. This implies that for R sufficiently close to R^*, any trajectory that hits B_1 has to wind around the stable manifold of the stationary point C^+_R at least one time before it could possibly hit the boundary of B_2, which means that x'_R has to change sign at least two times before x_R could possibly reach its first zero. Clearly, this is impossible for R in $\mathfrak{R}$. Therefore, there is a first time $T^1_{R^*}$ such that x'_{R^*} takes its first zero.

Since $T^1_{R^*}$ is the first zero of x'_{R^*}, $x'_{R^*} > 0$ in $(-\infty, T^1_{R^*})$ and $x''_{R^*}(T^1_{R^*}) \leq 0$. Therefore $Q''_{R^*} + qQ'_{R^*} = 2\mu x_{R^*}x'_{R^*} > 0$ in $(-\infty, T^1_{R^*})$, so that $Q'_{R^*} > 0$ on $(-\infty, T^1_{R^*}]$, and $z'_{R^*} = Q'_{R^*} + \frac{1}{s}x_{R^*}x'_{R^*} > 0$ in $(-\infty, T^1_{R^*}]$ as well. This excludes the possibility of $x''_{R^*}(T^1_{R^*})$ being zero, since otherwise $x'''_{R^*}(T^1_{R^*}) \geq 0$, and from the equation for x_{R^*}, one can derive that $z_{R^*} = R^* - 1$ and $z'_{R^*} \leq 0$, which is impossible. Hence $x''_{R^*}(T^1_{R^*}) < 0$. Consequently, $x'_{R^*} + sx_{R^*} > 0$ and $(x'_* + sx_{R^*})' < 0$ at T^1_*.

The second step. Since the solution of (1.1) is autonomous, we can shift t such that $T^1_R = 0$ for all $R \in \mathfrak{R} \cup R^*$. Under this normalization, since $x''_R(0) \neq 0$, by the Implicit Function Theorem, the solution of (1.1) is both smooth in time t and in the parameter R. We claim that $T^2_{R^*} = \liminf_{R\in\mathfrak{R}, R\searrow R^*} T^2_R$ is finite. In fact if $T^2_{R^*} = \infty$, then by the continuity of the trajectory in R and the definition of T^2_R, $(x'_{R^*} + sx_{R^*})' \leq 0$ and $(x'_{R^*} + sx_{R^*}) \geq 0$ for all $t > 0$. It then follows that $x'_{R^*} + sx_{R^*}$ is positive and monotonically decreasing in $[0,\infty)$, so that either x'_{R^*} reaches zero and then becomes positive at some time in $(0,\infty)$ before x_{R^*} reaches its first zero in $(-\infty,\infty)$ or $sx_{R^*} > -x'_{R^*} \geq 0$ for all $t > T^1_{R^*}$. The first case cannot happen since R^* is a limit point of $\mathfrak{R}$ in which x_R obtains its first zero before x'_R obtains its first zero in $(0,\infty)$. The second case implies that on the phase plane of x vis

x', (x_{R^*}, x'_{R^*}) approaches $(0,0)$ from the sector bounded by the positive x-axis and the line $x' + sx = 0$ (the possibility that x_{R^*} monotonically decreases to $\sqrt{q(R^* - 1)}$ can be excluded in a similar way as in the first step). This is impossible since the direction that a trajectory can approach the origin on the x–x' phase plane is $(1, \lambda_2^0)$ but $\lambda_2^0 < -(1+s)$. Hence, $T^2_{R^*} < \infty$.

By the definition of $T^2_{R^*}$, the continuity of solutions in R, and the second condition in the definition of $\Re$, $x'_{R^*} + sx_{R^*} = 0$ at $T^2_{R^*}$ and $(x'_{R^*} + sx_{R^*})' \le 0$ in $[0, T^2_{R^*}]$. If $(x'_{R^*} + sx_{R^*})' = 0$ at some point $T \in (0, T^2_{R^*})$, then $(x'_{R^*} + sx_{R^*})(T) > 0$ and $(x'_{R^*} + sx_{R^*})''(T) = 0$. From the differential equation $(x'+sx)'+(x'+sx) = s(R-z)x$, we then have that $z_{R^*}(T) < R^*$ and $z'_{R^*}(T) \ge 0$. In addition, from $z''+qz' = \frac{1}{s}[x(x'+sx)]' = x'(x'+sx)+x(x'+sx)' \le 0$ in $[T, T^2_{R^*}]$, we have that $z'_{R^*} < 0$ in $(T, T^2_{R^*}]$. Thus $z_{R^*}(T^2_{R^*}) < R^*$. However, at $T^2_{R^*}$, $sx_{R^*}(R^* - z_{R^*}) = (x'_{R^*}+sx_{R^*})'+(x'_{R^*}+sx_{R^*}) \le 0$, which implies that $z_{R^*}(T^2_{R^*}) \ge R^*$; we get a contradiction. This contradiction shows that $(x'_{R^*} + sx_{R^*}) < 0$ in $(0, T^2_{R^*})$. Now if $(x'_{R^*} + sx_{R^*})' = 0$ at $T^2_{R^*}$, then since by continuity and the definition of R^*, $x'_{R^*} + sx_{R^*}$ is positive in $[0, T^2_{R^*})$ and nonpositive in $[T^2_{R^*}, T^2_{R^*} + \delta)$ for some $\delta > 0$, we must have $(x_{R^*} + sx_{R^*})'' = 0$. Recall that this is equivalent to $y_{R^*} = y'_{R^*} = y''_{R^*} = 0$ at $T^2_{R^*}$, which, in turn, implies that $x_{R^*} = y_{R^*} = z_{R^*} = 0$; clearly this is impossible. Therefore, $T^2_{R^*}$ satisfies the second condition in the definition of $\Re$.

The final step. Now let $T^3_{R^*} = \liminf_{R \in \Re, R \searrow R^*} T^3_R$. We first claim that $T^3_{R^*} = \infty$. Assume that this is not true. Then $x(T^3_{R^*}) = 0$ and $x_{R^*} \ge 0$ for all $t \in (-\infty, T^3_{R^*}]$. From the continuity and the definition of R^*, $x'_{R^*} + sx_{R^*} \le 0$ for all $t \in [T^2_{R^*}, T^3_{R^*}]$. If $(x'_{R^*} + sx_{R^*})(T) = 0$ for some T in $(T^2_{R^*}, T^3_{R^*})$, then $(x'_{R^*} + sx_{R^*})'(T) = 0$, $(x'_{R^*} + sx_{R^*})''(T) \le 0$, and $x_{R^*}(T) > 0$. But from the equations $(x'_{R^*} + sx_{R^*})' + (x'_{R^*}+sx_{R^*})' = s(R^* - z_{R^*})x_{R^*}$ and $z'_{R^*} = \frac{1}{s}x_{R^*}(x'_{R^*}+sx_{R^*}) - qz_{R^*}$, this implies that $z_{R^*}(T) = R^*$ and $z'_{R^*}(T) < 0$. Consequently, $(x'_{R^*} + sx_{R^*})''(T) = -z'_{R^*}(T)x_{R^*}(T) > 0$, which is a contradiction. This contradiction shows that $x'_{R^*} + sx_{R^*} < 0$ in $(T^2_{R^*}, T^3_{R^*})$. Also, since $x'_{R^*}(T^3_{R^*}) = 0$ implies that the trajectory is the z-axis, we must have $x'(T^3_{R^*}) < 0$. Therefore, $T^3_{R^*}$ satisfies the third condition in the definition of $\Re$. Hence, from the first two steps, we have that $R^* \in \Re$, again a contradiction. This contradiction shows that $T^3_{R^*} = \infty$. Hence, we have that $x_{R^*} > 0$ for all t and, by a similar argument as above, $x'_{R^*} < -sx_{R^*} < 0$ for all $t > T^2_{R^*}$. Therefore, as $t \to \infty$, x_{R^*} and x'_{R^*} approach zero. That is, $\gamma^+_{R^*}$ approaches the origin.

This completes the proof of Theorem 3.1. □

4. Solutions for large *R*. To prove Lemma 3.3, we now study the behavior of the solutions of (1.1) for large R. To this end, we introduce a variation of a transformation which was first used by Robbins [13] and then by Swinnerton-Dyer [18] to establish, among other things, the existence of periodic solutions of (1.1) for large R. The transformation is as follows (assume that $q \ne 2s$):

$$\begin{aligned} &\tau = \sqrt{s(R-1)}\,t, && \varepsilon = (1+s)/\sqrt{s(R-1)}, \\ &\alpha = 4\mu s/(1+s), && \beta = q/(1+s), \\ &x_R(t) = 2\sqrt{s(R-1)}x_\varepsilon(\tau), && Q_R(t) = 4\mu\sqrt{s(R-1)}Q_\varepsilon(\tau). \end{aligned} \tag{4.1}$$

Under this transformation, the system (1.1) or (2.2) becomes

$$\begin{cases} \ddot{x}_\varepsilon = x_\varepsilon(1 - 2x_\varepsilon^2) - \varepsilon(\dot{x}_\varepsilon + \alpha x_\varepsilon Q_\varepsilon), \\ \dot{Q}_\varepsilon = x_\varepsilon^2 - \varepsilon\beta Q_\varepsilon, \end{cases} \tag{4.2}$$

where $\dot{} = \frac{d}{d\tau}$.

In what follows, we shall use the phase space $(x, \dot{x}, Q)$, so that a trajectory of (4.2) refers to its trace of points $(x_\varepsilon(\tau), \dot{x}_\varepsilon(\tau), Q_\varepsilon(\tau))$ for all $\tau \in (-\infty, \infty)$.

When $\varepsilon = 0$, the system (4.2) can be solved by quadrature. A solution of interest is given in the following lemma.

LEMMA 4.1. *When $\varepsilon = 0$, the system* (4.2) *has a unique trajectory γ_0^+ which approaches the origin as $\tau \to -\infty$ and lies in the first octant for sufficient negative large τ. Its corresponding solution is given by*

$$(4.3) \qquad X^0(\tau) = (\cosh\tau)^{-1}, \qquad Q^0(\tau) = 1 + \tanh\tau, \qquad t \in (-\infty, \infty).$$

The assertion of the Lemma follows by a direct calculation and details are omitted.

Note that in (4.3), we have normalized the solution such that X^0 obtains its maximum at $\tau = 0$.

One may notice that the stationary point $(0, 0, 0)$ of (4.2) changes from saddle type to degenerate type when ε is changed from positive to zero, so that the continuous dependence of the unstable manifold of the origin with respect to ε has to be rigorously verified. Hence, we provide the following lemma for completeness.

LEMMA 4.2. (1) *For every $\varepsilon \geq 0$,* (4.2) *has a unique trajectory γ_ε^+ which approaches the origin as $\tau \to -\infty$ and initially x_ε is positive.*

(2) *There exists a positive constant δ_1 such that for every $\varepsilon \in [0, 1]$, γ_ε^+ has the expansion*

$$x_\varepsilon = \xi, \qquad \dot{x}_\varepsilon = P(\varepsilon, \xi), \qquad Q_\varepsilon = Q(\varepsilon, \xi) \quad \textit{for all } \xi \in (0, \delta_1],$$

where $P(\varepsilon, \xi)$ and $Q(\varepsilon, \xi)$ are analytic functions of ξ and ε in an open neighborhood containing $[0, \delta_1] \times [0, 1]$. In addition, there exists a positive constant C_1 such that for all $\varepsilon \in [0, 1]$ and all $\xi \in [0, \delta_1]$, $P(\varepsilon, \xi)$ and $Q(\varepsilon, \xi)$ satisfy the following estimates:

$$(4.4) \qquad \begin{array}{ll} |P(\varepsilon, \xi) - \lambda_\varepsilon^P \xi| \leq C_1 \xi^2, & \textit{where } \lambda_\varepsilon^P := \sqrt{1 + (\varepsilon/2)^2} - \varepsilon/2, \\ |Q(\varepsilon, \xi) - \lambda_\varepsilon^Q \xi| \leq C_1 \xi^2, & \textit{where } \lambda_\varepsilon^Q := 1/(2\lambda_\varepsilon^P + \beta\varepsilon). \end{array}$$

(3) *There exists a small positive constant $\varepsilon_1 \in (0, 1]$ such that for all $\varepsilon \in [0, \varepsilon_1]$, the x_ε component of γ_ε^+ can obtain its first maximum. In addition, if one normalizes the solution $(X^\varepsilon, \dot{X}^\varepsilon, Q^\varepsilon)$ of* (4.2) *corresponding to γ_ε^+ such that X^ε obtains its first maximum at $\tau = 0$, then as a function of variables ε and τ, $(X^\varepsilon(\tau), \dot{X}^\varepsilon(\tau), Q^\varepsilon(\tau))$ is analytic in $[0, \varepsilon_1] \times (-\infty, \infty)$. Consequently, for all $T > 0$, there exists a constant $C_2(T)$ such that*

$$\|X^\varepsilon - X^0\|_{C^2([-T,T])} + \|Q^\varepsilon - Q^0\|_{C^1([-T,T])} \leq C_2(T)\varepsilon.$$

Proof. (1) When $\varepsilon > 0$, (4.2) is equivalent to (1.1) so that γ_ε^+ is well defined, whereas when $\varepsilon = 0$, γ_0^+ is given by Lemma 4.1.

(2) In a small neighborhood of the origin, γ_ε^+ can be obtained by solving the system of two ordinary differential equations:

$$\tfrac{d}{dx}P = x(1 - 2x^2)/P - \varepsilon(1 + \alpha Q/P), \qquad \tfrac{d}{dx}Q = x^2/P - \varepsilon\beta Q/P,$$

where x is an independent variable. One can verify that this new system has a convergent power series solution $P = P(\varepsilon, x) := \sum_{i=0}^\infty p_i(\varepsilon)x^i$, $Q = Q(\varepsilon, x) := \sum_{i=0}^\infty q_i(\varepsilon)x^i$ in some interval $[0, \delta_1]$ which is independent of $\varepsilon \in [0, 1]$, provided that one takes $p_0 = 0$, $p_1 = \lambda_\varepsilon^P$, $q_0 = q_1 = 0$, and $q_2 = \lambda_2^Q$, where λ_ε^P and λ_ε^Q are defined in (4.4), whereas all the other coefficients are uniquely determined by the differential equations. Clearly, this solution corresponds to γ_ε^+. In addition, by adding a dummy equation $\frac{d}{dx}\varepsilon = 0$, if necessary, one can show by the implicit-function theorem that $P(\varepsilon,\xi)$ and $Q(\varepsilon,x)$ are analytic in ε and x. This proves the second assertion of the lemma.

(3) From (2), the curve $(\delta_1, P(\varepsilon,\delta_1), Q(\varepsilon,\delta_1))$ is analytic in $\varepsilon \in [0,1]$. Taking this as initial conditions for (4.2) at $\tau = 0$, we have a solution $(\hat{x}_\varepsilon(\tau), \dot{\hat{x}}_\varepsilon(\tau), \hat{Q}_\varepsilon(\tau))$ which is analytic in both ε and τ. In addition, since at the point where $\hat{x}_0$ obtains its maximum, $\hat{x}_0'' = -1 \neq 0$, for all small ε, we can use the implicit-function theorem to solve, for T_ε, the equation $\hat{x}_\varepsilon'(T_\varepsilon) = 0$ near the place where $\hat{x}_0$ obtains its maximum. The function T_ε thus obtained is analytic in ε in $[0,\varepsilon_1]$ for some small positive ε_1. Shifting the time phase of the solution $(\hat{x}_\varepsilon, \dot{\hat{x}}_\varepsilon, \hat{Q}_\varepsilon)$ by T_ε, we then obtain the third assertion of the lemma. □

In what follows, we shall always use the representation $(X^\varepsilon, \dot{X}^\varepsilon, Q^\varepsilon)$ stated in Lemma 4.2 (3) for γ_ε^+, where $\varepsilon \in [0,\varepsilon_1]$.

LEMMA 4.3. *There exists a constant $\varepsilon_2 \in (0,\varepsilon_1]$ such that the following holds:*

(i) *For every $\varepsilon \in [0,\varepsilon_2]$, $\dot{X}^\varepsilon > 0$ in $(-\infty,0)$, $\dot{X}^\varepsilon(0) = 0$, and $\ddot{X}^\varepsilon(0) < -1/2$.*

(ii) *For every $\varepsilon \in (0,\varepsilon_2]$, there exists $T_\varepsilon^2 \in (0,\sqrt{\varepsilon}]$ such that*

$$(\dot{X}^\varepsilon + \tfrac{s\varepsilon}{1+s}X^\varepsilon)(T_\varepsilon^2) = 0, \qquad \tfrac{d}{d\tau}(\dot{X}^\varepsilon + \tfrac{s\varepsilon}{1+s}X^\varepsilon) < -1/4 \text{ in } [0,T_\varepsilon^2].$$

Proof. Since $\ddot{X}^0(0) = -1$, the first assertion follows from Lemma 4.2, whereas the second assertion follows by the continuity of the solution with respect to ε and τ and the implicit-function theorem. □

Observe that under the transformation (4.1), the quantity $x_R'+sx_R$ in the original variables becomes $\frac{2(1+s)^2}{\varepsilon^2}(\dot{x}_\varepsilon + \frac{s\varepsilon}{1+s}x_\varepsilon)$, so that when $R \in [\frac{1}{s}(\frac{1+s}{\varepsilon_2})^2, \infty)$, γ_R^+ defined in §3 satisfies the first two conditions in the definition of $\mathfrak{R}$. We now proceed to show that the third condition in the definition of $\mathfrak{R}$ is also satisfied by γ_R^+ for large R, or equivalently, by γ_ε^+ for small ε.

LEMMA 4.4. (1) *Define $H_\varepsilon = \dot{x}_\varepsilon^2 + x_\varepsilon^4 - x_\varepsilon^2$. Then, along any trajectory of* (4.2),

$$\tfrac{1}{2\varepsilon}\dot{H}_\varepsilon = -\dot{x}_\varepsilon^2 - \alpha Q_\varepsilon x_\varepsilon \dot{x}_\varepsilon, \qquad \textit{for all } \varepsilon \geq 0, \tau \in (-\infty,\infty). \tag{4.5}$$

(2) *Define $m := -\int_{-\infty}^{\infty} \left[(\dot{X}^0)^2 + \alpha Q^0 X^0 \dot{X}^0\right] d\tau$. Then $m = \frac{2}{1+s}\left[s - \frac{2q+1}{3}\right]$.*

(3) *For any $T > 0$, there exists $C_3(T) > 0$ such that for all $\varepsilon \in [0,\varepsilon_2]$,*

$$\left| \int_{-\infty}^{T} \left[(\dot{X}^\varepsilon)^2 + \alpha Q^\varepsilon X^\varepsilon \dot{X}^\varepsilon - (\dot{X}^0)^2 - \alpha Q^0 X^0 \dot{X}^0\right] d\tau \right| \leq C_3(T)\varepsilon.$$

Proof. The first two assertions follow from the differential equations in (4.2), the explicit expressions of X^0 and Q^0 in Lemma 4.1, the definition of α in (4.1), and direct computation; details are omitted. We now prove (3).

Let $T_\varepsilon(\delta_1) < 0$ be the time such that $X^\varepsilon(T_\varepsilon(\delta_1)) = \delta_1$. Since at $T_\varepsilon(\delta_1)$, $\dot{X}^\varepsilon > 0$, $T_\varepsilon(\delta_1)$ is analytic in ε. Hence the integral of $(\dot{X}^\varepsilon)^2 + \alpha Q^\varepsilon X^\varepsilon \dot{X}^\varepsilon$ from $[T_\varepsilon(\delta_1), T]$ differs from the integral of $(\dot{X}^0)^2 + \alpha Q^0 X^0 \dot{X}^0$ from $[T_0(\delta_1), T]$ by a quantity of order ε. Using the change of variables $\xi = X^\varepsilon(\tau)$, we have that

$$\int_{-\infty}^{T_\varepsilon(\delta_1)} \left[(\dot{X}^\varepsilon)^2 + \alpha Q^\varepsilon X^\varepsilon \dot{X}^\varepsilon\right] d\tau = \int_0^{\delta_1} \left[P(\varepsilon,\xi) + \alpha Q(\varepsilon,\xi)\xi\right] d\xi$$
$$= \int_0^{\delta_1} \left[P(0,\xi) + \alpha Q(0,\xi)\xi\right] d\xi + O(\varepsilon) = \int_{-\infty}^{T_0(\delta_1)} \left[(\dot{X}^0)^2 + \alpha Q^0 X^0 \dot{X}^0\right] d\tau + O(\varepsilon).$$

The third assertion of the lemma thus follows. □

We now are ready to prove the following lemma.

LEMMA 4.5. *Assume that* $s > \frac{2q+1}{3}$. *Then there exists* $\varepsilon_0 \in (0, \varepsilon_2]$ *such that for every* $\varepsilon \in (0, \varepsilon_0]$, *there exists* $T_\varepsilon^3 > T_\varepsilon^2$ *such that the following hold:*

$$X^\varepsilon(T_\varepsilon^3) = 0, \qquad \dot{X}^\varepsilon < 0\, in\, (0, T_\varepsilon^3], \qquad and\, (\dot{X}^\varepsilon + \tfrac{s\varepsilon}{1+s} X^\varepsilon) < 0\, in\, (T_\varepsilon^2, T_\varepsilon^3].$$

Proof. Since $s > \frac{2q+1}{3}$, m defined in Lemma 4.4 is positive. Let T_m be a large positive constant such that

$$\int_{T_m}^{\infty} \Big[(\dot{X}^0)^2 + \alpha |Q^0 X^0 \dot{X}^0| \Big]\, d\tau < \tfrac{1}{8} m, \tag{4.6}$$

$$\delta_m := X^0(T_m) < \min\{\tfrac{1}{8}, \tfrac{1}{12} m\}. \tag{4.7}$$

By the continuous dependence of the solution on ε, we can find a positive constant $\varepsilon_0 \in (0, \min\{1/\sqrt{2}, \varepsilon_2, m(8C_3(T_m))^{-1}\}]$ such that for all $\varepsilon \in [0, \varepsilon_0]$,

$$\dot{X}^\varepsilon < -\tfrac{s\varepsilon}{1+s} X^\varepsilon < 0 \qquad \text{in } (T_\varepsilon^2, T_m + 1], \tag{4.8}$$

$$|X^\varepsilon(T_m) - X^0(T_m)| + |Q^\varepsilon(T_m) - Q^0(T_m)| \le \tfrac{1}{2} \delta_m. \tag{4.9}$$

Denote the value of H_ε on γ_ε^+ by H^ε. Since $H^\varepsilon(-\infty) = 0$, integrating (4.5) along γ_ε^+ from $-\infty$ to T_m and using Lemma 4.4, we have that

$$\begin{aligned}
\tfrac{1}{2\varepsilon} H^\varepsilon(T_m) &= \int_{-\infty}^{T_m} \Big[-(\dot{X}^\varepsilon)^2 - \alpha Q^\varepsilon X^\varepsilon \dot{X}^\varepsilon \Big]\, d\tau \\
&= \int_{-\infty}^{\infty} \Big[-(\dot{X}^0)^2 - \alpha Q^\varepsilon X^0 \dot{X}^0 \Big]\, d\tau + \int_{T_m}^{\infty} \Big[(\dot{X}^0)^2 + \alpha Q^\varepsilon X^0 \dot{X}^0 \Big]\, d\tau \\
&\quad + \int_{-\infty}^{T_m} \Big[(\dot{X}^\varepsilon)^2 + \alpha Q^\varepsilon X^\varepsilon \dot{X}^\varepsilon - (\dot{X}^0)^2 - \alpha Q^0 X^0 \dot{X}^0 \Big]\, d\tau \\
&\ge m - \tfrac{1}{8} m - C_3(T_m)\varepsilon \ge \tfrac{6}{8} m \qquad \forall \varepsilon \in [0, \varepsilon_0].
\end{aligned}$$

Hence, $H^\varepsilon(T_m) \ge \frac{3}{2} m\varepsilon$ for all $\varepsilon \in [0, \varepsilon_0]$.

Now for all $\varepsilon \in (0, \varepsilon_0]$, define

$$T_\varepsilon^3 := \inf \Big\{ \tau > T_m \;\Big|\; X^\varepsilon > 0, \dot{X}^\varepsilon < 0 \text{ and } H^\varepsilon > m\varepsilon \text{ in } [T_m, \tau] \Big\}.$$

In view of (4.8), T_ε^3 is well defined and $T_\varepsilon^3 > T_m + 1$.

Since $\dot{X}^\varepsilon < 0$ in $[T_m, T_\varepsilon^3)$, we have that $0 \le X^\varepsilon(\tau) \le X^\varepsilon(T_m) \le \frac{3}{2}\delta_m \le \frac{3}{16}$ for all $\tau \in [T_m, T_\varepsilon^3]$. In addition, from the definition of H^ε, in $[T_m, T_\varepsilon^3]$,

$$-\dot{X}^\varepsilon = \sqrt{H^\varepsilon + (X^\varepsilon)^2[1 - (X^\varepsilon)^2]} \ge \sqrt{m\varepsilon + \tfrac{1}{2}(X^\varepsilon)^2}.$$

Therefore, we have that $-\dot{X} \ge \sqrt{m\varepsilon}$ in $[T_m, T_\varepsilon^3]$, which implies that $T_\varepsilon^3 \le X^\varepsilon(T_m)/\sqrt{m\varepsilon} < \infty$. Also we have that $-\dot{X}^\varepsilon > X^\varepsilon/\sqrt{2}$ in $[T_m, T_\varepsilon^3]$, which, together with (4.8) and the fact that $\varepsilon_0 \le 1/\sqrt{2}$, yields

$$\dot{X}^\varepsilon + \tfrac{s\varepsilon}{1+s} X^\varepsilon < 0 \text{ in } (T_\varepsilon^2, T_\varepsilon^3], \qquad \text{for all } \varepsilon \in [0, \varepsilon_0].$$

Hence, to finish the proof, we need only show that $X^\varepsilon(T_\varepsilon^3) = 0$.

At T_ε^3, there are only two possibilities: either $X^\varepsilon(T_\varepsilon^3) = 0$ or $H^\varepsilon(T_\varepsilon^3) = m\varepsilon$. We shall show that the second alternative will not happen by integrating the identity (4.5) from T_m to T_ε^3. To do this, we need an estimate for the maximum of $\dot{X}^\varepsilon$ in $[T_m, T_\varepsilon^3]$. (Though this can be directly verified by using Lemma 2.1, we provide another proof here for possible other applications.)

Since $Q^\varepsilon > 0$ for all $\tau \in (-\infty, \infty)$ and all $\varepsilon \geq 0$, $\dot{Q}^\varepsilon = -\varepsilon\beta Q^\varepsilon + (X^\varepsilon)^2 < (X^\varepsilon)^2$. Consequently, for all $\tau \in (T_m, T_\varepsilon^3]$,

$$\begin{aligned} Q^\varepsilon(\tau) &< Q^\varepsilon(T_m) + \int_{T_m}^{\tau} (X^\varepsilon)^2 = Q^\varepsilon(T_m) + \int_{X^\varepsilon(T_m)}^{X^\varepsilon(\tau)} \frac{X^\varepsilon}{\dot{X}^\varepsilon}\, dX^\varepsilon \\ &\leq Q^\varepsilon(T_m) + \int_0^{X^\varepsilon(T_m)} \frac{x}{\sqrt{m\varepsilon + x^2/2}} dx \leq Q(T_m) + \sqrt{2} X^\varepsilon(T_m) \leq 4, \end{aligned}$$

since $Q^\varepsilon(T_m) = Q^0(T_m) + \frac{1}{2}\delta_m \leq \frac{5}{2}$ and $X^\varepsilon(T_m) \leq \frac{3}{2}\delta_m < \frac{1}{2}$. Integrating (4.5) from T_m to $\tau \in [T_m, T_\varepsilon^3]$ then yields

$$\tfrac{1}{2\varepsilon}[H^\varepsilon(\tau) - H^\varepsilon(T_m)] \leq -\alpha \int_{T_m}^{\tau} Q^\varepsilon X^\varepsilon \dot{X}^\varepsilon \leq -4\alpha \int_{T_m}^{T_\varepsilon^3} X^\varepsilon \dot{X}^\varepsilon\, d\tau \leq 2\alpha (X^\varepsilon(T_m))^2 \leq \tfrac{1}{2}$$

for all $\varepsilon \in [0, \varepsilon_0]$, since $\alpha = \frac{4\mu s}{1+s} < 4$ and $X^\varepsilon(T_m) \leq \frac{3}{2}\delta_m \leq \frac{1}{4}$. It then follows from the definition of H^ε that $|\dot{X}^\varepsilon| \leq \sqrt{H^\varepsilon + X^2} \leq 1$ for all $\tau \in [T_m, T_\varepsilon^3]$.

Having the estimate for $|\dot{X}^\varepsilon|$ in $[T_m, T_\varepsilon^3]$, we can now integrate the identity (4.5) from T_m to T_ε^3 and using the fact that $-\alpha Q^\varepsilon X^\varepsilon \dot{X}^\varepsilon \geq 0$ in $(0, T_\varepsilon^3)$, obtain that

$$\begin{aligned} \tfrac{1}{2\varepsilon}\Big[H^\varepsilon(T_\varepsilon^3) - H^\varepsilon(T_m)\Big] &\geq -\int_{T_m}^{T_\varepsilon^3} (\dot{X}^\varepsilon)^2\, d\tau = -\int_{X^\varepsilon(T_m)}^{X^\varepsilon(T_\varepsilon^3)} \dot{X}^\varepsilon\, dX^\varepsilon \\ &\geq -X^\varepsilon(T_m) \geq -\tfrac{3}{2}\delta_m \qquad \forall \varepsilon \in [0, \varepsilon_0]. \end{aligned}$$

Therefore, $H^\varepsilon(T_\varepsilon^3) \geq H^\varepsilon(T_m) - 3\varepsilon\delta_m \geq \frac{3}{2}m\varepsilon - 3\varepsilon\delta_m \geq \frac{5}{4}m\varepsilon$ by the smallness of δ_m. Hence, we must have $X^\varepsilon(T_\varepsilon^3) = 0$. This completes the proof of the lemma. □

As mentioned earlier, since $X_R' + sX_R = \frac{2(1+s)^2}{\varepsilon^2}[\dot{x}_\varepsilon + \frac{s\varepsilon}{1+s}x_\varepsilon]$, Lemmas 4.3 and 4.5 imply that for R large enough, γ_R^+ satisfies the three conditions in the definition $\mathfrak{R}$; that is, $R \in \mathfrak{R}$. The assertion of Lemma 3.3 thus follows, thereby completing the proof of Theorem 3.1.

Acknowledgments. The author is grateful to Professor Stuart P. Hastings for bringing the author's attention to the fascinating Lorenz equations, demonstrating invaluable numerical calculations, and providing much appreciated direction through numerous communications.

Note added in proof. During this paper's production, the author learned of a previous result by G. A. Leonov [*Differential Equations*, 24 (1988), pp. 634–638; *Russian Math. Surveys*, 43 (1988), pp. 216–217]. He proved that when $s > \frac{2q+1}{3}$, the Lorenz system has a homoclinic orbit for some large R.

REFERENCES

[1] X. Chen, *Lorenz equations, part* II: *Existence of rotated homoclinic orbits and chaotic orbits*, Discrete Continuous Dynam. Sys., 2 (1996), pp. 121–140.

[2] X. Chen, *Lorenz equations, part* III: *Existence of homoclinic explosion*, preprint, University of Pittsburgh, Pittsburgh, PA, 1995.

[3] J. Guckenheimer, *A strange, strange attractor*, in The Hopf Bifurcation and Its Applications, J. E. Marsden and M. McCracken, eds., Springer-Verlag, Berlin, New York, 1976.

[4] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Appl. Math. Sci. 42, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1983.

[5] J. Guckenheimer and R. F. Williams, *Structural stability of the Lorenz attractor*, Inst. Hautes Études Sci. Publ. Math., 50 (1979), pp. 59–72.

[6] B. Hassard and J. Zhang, *Existence of a homoclinic orbit of the Lorenz system by precise shooting,* SIAM, J. Math. Anal., 25 (1994), pp. 179–196.

[7] B. Hassard, S. Hastings, W. Troy, and J. Zhang, *A computer proof that the Lorenz equations have "chaotic" solutions*, Appl. Math. Lett., to appear.

[8] S. P. Hastings and W. C. Troy, *A shooting approach to the Lorenz equations,* Bull. Amer. Math. Soc., 27 (1992), pp. 298–303.

[9] ———, *A shooting approach to chaos in the Lorenz equations,* J. Differential Equations, 126 (1996), to appear.

[10] ———, *A proof that the Lorenz equations have a homoclinic orbit*, J. Differential Equations, 113 (1994), pp. 166–188.

[11] E. N. Lorenz, *Deterministic non-periodic flows*, J. Atmos. Sci., 20 (1963), pp. 130–141.

[12] K. Mischaikow and M. Mrozek, *Chaos in the Lorenz equations: A computer-assisted proof,* Bull. Amer. Math. Soc., 32 (1995), pp. 66–72.

[13] K. A. Robbins, *Periodic solutions and bifurcation structure at a high R in the Lorenz model,* SIAM J. Appl. Math., 36 (1979), pp. 457–472.

[14] C. Robinson, *Homoclinic bifurcation to a transitive attractor of Lorenz type*, Nonlinearity, 2 (1989), pp. 495–518.

[15] M. R. Rychlik, *Lorenz attractors through Sil'nikov-type bifurcation, part* I, Ergodic Theory Dynamical Systems, 10 (1989), pp. 793–821.

[16] L. P. Sil'nikov, *A contribution to the problem of the structure of an extended neighborhood of a rough equilibrium state of saddle focus type*, Math. USSR-Sb., 10 (1970), pp. 91–102.

[17] C. Sparrow, *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*, Appl. Math. Sci. 41, Springer-Verlag, Berlin, New York, Heidelberg, 1982.

[18] H. P. F. Swinnerton-Dyer, *The method of averaging for some almost-conservative differential equations*, J. London Math. Soc., 22 (1980), pp. 534–542.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 1070–1083, July 1996

012

A GEOMETRIC APPROACH TO GLOBAL-STABILITY PROBLEMS*

MICHAEL Y. LI† AND JAMES S. MULDOWNEY‡

Abstract. A new criterion for the global stability of equilibria is derived for nonlinear autonomous ordinary differential equations in any finite dimension based on recent developments in higher-dimensional generalizations of the criteria of Bendixson and Dulac for planar systems and on a local version of the C^1 closing lemma of Pugh. The classical result of Lyapunov is obtained as a special case.

Key words. global stability, Bendixson criterion, compound equations

AMS subject classifications. 34D20, 34C35

1. Introduction.

Let the map $x \mapsto f(x)$ from an open subset $D \subset \mathbf{R}^n$ to $\mathbf{R}^n$ be such that each solution $x(t)$ to the differential equation

$$x' = f(x) \tag{1.1}$$

is uniquely determined by its initial value $x(0) = x_0$, and denote this solution by $x(t, x_0)$.

An equilibrium point $\bar{x} \in D$ of (1.1) is said to be *locally stable*—or simply *stable*—if, for each neighbourhood U of $\bar{x}$, there exists a neighbourhood V of $\bar{x}$ such that $x(t, V) \subset U$ for all $t > 0$; it is said to *attract points* in a neighbourhood W if $x(t, x_0) \to \bar{x}$ as $t \to \infty$ for each $x_0 \in W$. It is important to note that this convergence may not be uniform for $x_0 \in W$, and $\bar{x}$ may fail to be stable when it attracts points in a neighbourhood. This can be demonstrated by an example of $\bar{x}$ with a homoclinic orbit. However, if a stable $\bar{x}$ attracts points in a bounded set W, then the attraction is uniform with respect to $x_0 \in W$. In this case, $\bar{x}$ is said to attract W. We say $\bar{x}$ is *asymptotically stable* if it is stable and attracts a neighbourhood. The *basin of attraction* of $\bar{x}$ is the union of all points which it attracts. If $\bar{x}$ is asymptotically stable, its basin of attraction is an open subset of D and contains a neighbourhood of $\bar{x}$. An equilibrium $\bar{x}$ is said to be *globally asymptotically stable*—or simply *globally stable*—with respect to an open set D_1 if it is asymptotically stable and its basin of attraction contains D_1.

The local stability of an equilibrium $\bar{x}$ can be routinely verified by construction of a Lyapunov function in a small neighbourhood of $\bar{x}$ or by linearizing (1.1) at $\bar{x}$ if f is C^1. The primary interest of this paper is in the problem of global stability. Note that if $\bar{x}$ is globally stable with respect to D_1, then $\bar{x}$ is necessarily the only equilibrium in D_1 and there exists compact neighbourhood K of $\bar{x}$ such that each compact subset $F \subset D_1$ satisfies $x(t, F) \subset K$ for sufficiently large t. Such a K is called *absorbing* in D_1 for (1.1). An open set $D \subset \mathbf{R}^n$ is *simply connected* if each closed curve in D can be

* Received by the editors April 22, 1994; accepted for publication (in revised form) October 21, 1994. This research was supported by Natural Sciences and Engineering Research Council of Canada grant NSERC A7197.

† Centre de Recherches Mathématiques, Université de Montréal, Montréal, PQ H3C 3J7, Canada. Current address: Center for Dynamical Systems, Georgia Institute of Technology, Atlanta, GA 30332. The research of this author was partially supported by the Dissertation Fellowship at the University of Alberta and by the Natural Sciences and Engineering Research Council of Canada Postdoctoral Fellowship.

‡ Department of Mathematics, University of Alberta, Edmonton, AL T6G 2G1, Canada.

continuously deformed to a point within D. Without loss of generality, we formulate the problem as follows:

THE GLOBAL-STABILITY PROBLEM. *Assume that*

(H_1) *D is simply connected;*

(H_2) *there is a compact absorbing set $K \subset D$;*

(H_3) *$\bar{x}$ is the only equilibrium of* (1.1) *in D.*

Find conditions under which the global stability of $\bar{x}$ with respect to D is implied by its local stability.

The difficulty associated with this problem is largely due to the lack of practical tools. The method of Lyapunov functions is most commonly used (see [5], [6]); its application is often hindered by the fact that in many cases global Lyapunov functions are difficult to construct and there is practically no general approach to the construction of such functions. The application of the theory of monotone flows [8], [21] is an alternative which has been successfully implemented in recent years.

A new approach to the global-stability problem has emerged from a series of papers on higher-dimensional generalizations of the criteria of Bendixson and Dulac for planar systems and on so-called autonomous convergence theorems. Assume that (1.1) satisfies a condition in D which precludes the existence of periodic solutions and suppose that this condition is robust in the sense that it is also satisfied by ordinary differential equations that are C^1-close to (1.1); then every nonwandering point of (1.1) is an equilibrium, since otherwise, by the C^1 closing lemma of Pugh [19], [20], we can perturb (1.1) near such a nonequilibrium nonwandering point to get a periodic solution. As a special case, every omega limit point of (1.1) is an equilibrium. In the context of our global-stability problem, this implies that $\bar{x}$ attracts points in D. As a consequence its global stability is implied by the local stability.

Higher-dimensional generalizations of Bendixson's criterion have been obtained in papers of R. A. Smith [22] and J. S. Muldowney [17]. This was further developed and the generalized Dulac criteria derived in [11] based on the study of evolution of general surface functionals under (1.1). Smith used the fact that his condition has the required robustness to imply that all bounded trajectories converge to equilibria. Such results are called, after Smith [22], autonomous convergence theorems, and they are further explored in [12] and proved under the generalized Dulac conditions developed in [11]. In the special case that (1.1) has a unique equilibrium $\bar{x}$ in D, it is proved in [12] that each of these generalized Dulac conditions also implies the local stability of $\bar{x}$ and hence its global stability with respect to D. As shown in [12], the traditional method of Lyapunov functions can also be interpreted in this context.

Each of these conditions will be called a *Bendixson criterion* in this paper. Our main purpose is to introduce a new Bendixson criterion for (1.1) which is an extension of the generalized Dulac conditions in [11] and [12]. Roughly speaking, instead of requiring these generalized Dulac conditions to hold pointwise in D as in [11] and [12], we require that they hold after being averaged over time along all the trajectories. Because of this time average along trajectories, it is not always true that our Bendixson criterion is robust under small C^1 perturbations of f. We introduce the notion of robustness of a Bendixson criterion under *local* C^1 perturbations of f in the sense that it is also satisfied by g which is C^1-close to f and differs from f only near a point. We prove that our Bendixson criterion is robust in this weaker sense. Using a local version of Pugh's closing lemma [7], we develop the theory of [12] under weaker conditions.

For autonomous systems which possess the Poincaré–Bendixson property, a different method for proving global stability was recently used in [1] for a planar system associated with a chemostat model and in [13] for a three-dimensional competitive system arising from an epidemiological model. In this case, the key step is to rule out periodic trajectories. This was accomplished by proving that periodic trajectories are orbitally asymptotically stable whenever they exist using the stability criterion of Poincaré for planar systems (as in [1]) and its higher-dimensional generalizations (as in [13]) developed in [17].

The present paper is arranged as follows: in the next section, we establish the general framework; in §3, we introduce our Bendixson criterion in Theorem 3.1 and prove a new global-stability result in Theorem 3.5; in §4, as an example, we consider a global-stability problem arising in an epidemiological model.

2. A general principle for global stability. We begin by formulating the local version of the C^1 closing lemma of Pugh as in [7]. Let $|\cdot|$ denote a vector norm on $\mathbf{R}^n$ and the operator norm which it induces for linear mappings from $\mathbf{R}^n$ to $\mathbf{R}^n$. The distance between two functions f, $g \in C^1(D \to \mathbf{R}^n)$ such that $f - g$ has compact support is

$$|f-g|_{C^1} = \sup\left\{|f(x)-g(x)| + \left|\frac{\partial f}{\partial x}(x) - \frac{\partial g}{\partial x}(x)\right| \; : \; x \in D\right\}.$$

Here and throughout the paper, $\frac{\partial f}{\partial x}$ denotes the Jacobian matrix of a mapping f. A function $g \in C^1(D \to \mathbf{R}^n)$ is called a C^1 *local ϵ-perturbation* of f at $x_0 \in D$ if there exists an open neighbourhood U of x_0 in D such that the support $\text{supp}(f-g) \subset U$ and $|f-g|_{C^1} < \epsilon$. For such g, we consider the corresponding differential equation

$$x' = g(x). \tag{2.1}$$

A point $x_0 \in D$ is *wandering* for (1.1) if there exists a neighbourhood U of x_0 and $T > 0$ such that $U \cap x(t, U)$ is empty for all $t > T$. Thus, for example, any equilibrium, alpha limit point, or omega limit point is nonwandering.

LEMMA 2.1. *Let $f \in C^1(D \to \mathbf{R}^n)$. Suppose that x_0 is a nonwandering point for* (1.1) *and that $f(x_0) \neq 0$. Then, for each neighbourhood U of x_0 and $\epsilon > 0$, there exists a C^1 local ϵ-perturbation g of f at x_0 such that*

(1) $\text{supp}\,(f-g) \subset U$ *and*

(2) *the system* (2.1) *has a nonconstant periodic solution whose trajectory passes through x_0.*

A *Bendixson criterion* for (1.1) is a condition satisfied by f which precludes the existence of nonconstant periodic solutions to (1.1). A Bendixson criterion is said to be *robust under C^1 local perturbations of f at x_0* if, for each sufficiently small $\epsilon > 0$ and neighbourhood U of x_0, it is also satisfied by each C^1 local ϵ-perturbations g such that $\text{supp}(f-g) \subset U$.

Given in the following are some examples of Bendixson criteria.

(1) When $n = 2$, $D = \mathbf{R}^2$, the classical result of Bendixson states that if

$$\text{div}(f) < 0 \quad \text{in} \quad \mathbf{R}^2, \tag{2.2}$$

then (1.1) has no nonconstant periodic solutions. Bendixson's criterion (2.2) was later generalized to the Dulac criterion

$$\text{div}(\alpha f) < 0, \tag{2.3}$$

where $x \mapsto \alpha(x)$ is some scalar-valued function. Conditions (2.2) and (2.3) can be replaced by $\operatorname{div}(f) > 0$ and $\operatorname{div}(\alpha f) > 0$, respectively, so that it is not the sign but having constant sign throughout D that is important.

(2) Let W be the Euclidean unit ball in $\mathbf{R}^2$ and let $\overline{W}$ and ∂W be its closure and boundary, respectively. If $D \subset \mathbf{R}^n$, a function $\varphi \in \operatorname{Lip}(\overline{W} \to D)$ will be described as a *simply connected rectifiable* 2*-surface in* D; a function $\psi \in \operatorname{Lip}(\partial W \to D)$ is a *closed rectifiable curve in* D and will be called *simple* if it is one to one. Let $|\cdot|$ denote a vector norm in $\mathbf{R}^N$ as well as the matrix norm which it induces for $N \times N$ matrices. The *Lozinskiĭ measure* $\mu(E)$ of a $N \times N$ matrix E with respect to the norm $|\cdot|$ is defined as

$$\mu(E) = \lim_{h \to 0+} \frac{|I + hE| - 1}{h}$$

(see [3, p. 41]). Lozinskiĭ measures have been used for estimation of eigenvalues of matrices. They also arise in the stability analysis of linear differential systems when certain vector norm of solutions are used as Lyapunov functions. Readers are refered to [3] for their properties and applications. Consider a nonsingular $\binom{n}{2} \times \binom{n}{2}$ matrix-valued function $x \mapsto A(x)$ which is C^1 in D and a vector norm $|\cdot|$ on $\mathbf{R}^{\binom{n}{2}}$. Let μ be the Lozinskiĭ measure with respect to $|\cdot|$. Under assumptions (H_1) and (H_2), it is proved in [11] that if

$$\mu\left(A_f A^{-1} + A\frac{\partial f}{\partial x}^{[2]} A^{-1}\right) \leq -\delta < 0 \quad \text{on} \quad K, \tag{2.4}$$

then no simple closed rectifiable curve in D can be invariant with respect to (1.1). Here $A_f = (DA)(f)$ or, equivalently, A_f is the matrix obtained by replacing each entry a_{ij} in A by its directional derivative in the direction of f, $\frac{\partial a_{ij}}{\partial x}^* f$, and $\frac{\partial f}{\partial x}^{[2]}$ is the second additive compound matrix of $\frac{\partial f}{\partial x}$ (see [15], [17]). For readers unfamiliar with the Lozinskiĭ measure, the condition (2.4) is equivalent to assuming that $V(x, y) = |A(x)y|$ is a Lyapunov function whose derivative with respect to the $n + \binom{n}{2}$-dimensional system

$$\frac{dx}{dt} = f(x), \qquad \frac{dy}{dt} = \frac{\partial f}{\partial x}^{[2]}(x)\, y$$

is negative definite. It rules out not only periodic trajectories but also homoclinic trajectories and heteroclinic loops since each case gives rise to a simple closed rectifiable invariant curve.

Setting $A = I$ in (2.4) leads to the following condition:

$$\mu\left(\frac{\partial f}{\partial x}^{[2]}\right) < 0, \tag{2.5}$$

which was first obtained in [17]. If the norm $|\cdot|$ is such that $|y| = |y^* y|^{\frac{1}{2}}$, then calculation of the Lozinskiĭ measure μ in (2.5) according to [2] or [17] yields

$$\lambda_1 + \lambda_2 < 0, \tag{2.6}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are eigenvalues of $\frac{1}{2}(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial x}^*)$, a condition which was first established in [22]. Note that when $n = 2$, condition (2.5) is the classical Bendixson criterion. The criterion (2.4) provides the flexibility of a choice of $\binom{n}{2} \times \binom{n}{2}$ arbitrary functions in addition to the choice of vector norms $|\cdot|$ in deriving suitable conditions.

(3) Let $x \mapsto V(x)$ be a scalar-valued function which is C^1 in D. Then the condition

$$\frac{\partial V}{\partial x} f(x) < 0 \quad \text{if} \quad f(x) \neq 0 \tag{2.7}$$

is a Bendixson criterion since $V(x)$ is strictly decreasing along each solution of (1.1). Such a function is usually called a *global Lyapunov function* for (1.1).

Suppose that f satisfies a Bendixson criterion which is robust under C^1 local perturbations of f at all nonwandering points of (1.1) which are not equilibria. Then, for each C^1 local ϵ-perturbation g of f at such a nonwandering point, when ϵ is sufficiently small, (2.1) can not have any nonconstant periodic solutions. Therefore, Lemma 2.1 implies that every nonequilibrium point of (1.1) must be wandering. We thus have the following result.

PROPOSITION 2.2. *Suppose a Bendixson criterion for* (1.1) *is robust under* C^1 *local perturbations of* f *at all nonequilibrium nonwandering points to* (1.1). *Then every nonequilibrium point of* (1.1) *is wandering.*

Suppose $D = \mathbf{R}^n$ and all solutions to (1.1) are forwardly bounded. Then for each $x_0 \in \mathbf{R}^n$, $\omega(x_0)$ is nonempty and compact. If we assume that (1.1) has a unique equilibrium $\bar{x}$ in $\mathbf{R}^n$, then the conditions of Proposition 2.2 imply that $\omega(x_0) = \bar{x}$ for all $x_0 \in \mathbf{R}^n$. If $\bar{x}$ is also stable, then it is globally stable with respect to D. This provides a solution to the global-stability problem.

THEOREM 2.3 (global-stability principle). *Assume that*

(1) $D = \mathbf{R}^n$ *and all solutions to* (1.1) *are forwardly bounded;*

(2) $\bar{x} \in \mathbf{R}^n$ *is the unique equilibrium of* (1.1) *in* $\mathbf{R}^n$; *and*

(3) (1.1) *satisfies a Bendixson criterion that is robust under* C^1 *local perturbations of* f *at each nonwandering point* x_1 *for* (1.1) *such that* $f(x_1) \neq 0$.

Then $\bar{x}$ *is globally stable in* $\mathbf{R}^n$ *provided it is stable.*

If $D \subset \mathbf{R}^n$ is an open subset, results like Theorem 2.3 also hold under the assumption (H_2) that D contains a compact absorbing set K. In this case, the trajectory of each solution to (1.1) eventually enters and remains in K; it does not approach the boundary of D. Condition (3) of Theorem 2.3 implies that its omega limit set is the singleton $\{\bar{x}\}$. Therefore, we have the following local version of Theorem 2.3.

THEOREM 2.4. *Suppose that assumptions* (H_2) *and* (H_3) *hold and that* (1.1) *satisfies a Bendixson criterion that is robust under* C^1 *local perturbations of* f *at all nonequilibrium nonwandering points for* (1.1). *Then* $\bar{x}$ *is globally asymptotically stable with respect to* D *provided it is stable.*

In many cases, a Bendixson criterion would imply that the unique equilibrium $\bar{x}$ is locally stable. This is the case for conditions (2.4) and (2.7). The following result, which contains the classical global-stability result of Lyapunov (see [5]), was proved in [12].

THEOREM 2.5. *Under assumptions* (H_1), (H_2), *and* (H_3), $\bar{x}$ *is globally asymptotically stable in* D *provided that either* (2.4) *or* (2.7) *holds.*

Remark. Another generalization of the global stability result of Lyapunov is LaSalle's invariance principle, [10, Chap. 2, Thm. 6.4], in which properties of limit sets are obtained when (2.7) is replaced by the weak inequality $\frac{\partial V}{\partial x} f(x) \leq 0$ in D. For discussions on the relation of LaSalle's result with autonomous convergence theorems, we refer the reader to [12].

3. The quantity $\bar{q}_2$.

In this section, we develop some new Bendixson criteria which are robust under local C^1 perturbations. Assume that (1.1) has a compact absorbing set $K \subset D$. Then every solution $x(t, x_0)$ of (1.1) exists for all $t > 0$. The following quantities are well defined:

$$\bar{q}_2 = \limsup_{t\to\infty} \sup_{x_0\in K} \frac{1}{t}\int_0^t \mu(B(x(s,x_0)))\,ds, \tag{3.1}$$

where

$$B = A_f A^{-1} + A\frac{\partial f}{\partial x}^{[2]} A^{-1} \tag{3.2}$$

and $x \mapsto A(x)$ is a $\binom{n}{2} \times \binom{n}{2}$ matrix-valued function as in (2.4).

Let $\psi \in \mathrm{Lip}(\partial W \to D)$ be a simple closed rectifiable curve in D. Then

$$\Sigma(\psi, D) = \{\varphi \in \mathrm{Lip}(\overline{W} \to D) \ : \ \varphi(\partial W) = \psi(\partial W)\}$$

is nonempty since D is simply connected. Define a functional $\mathcal{S}$ on $\Sigma(\psi, D)$ by

$$\mathcal{S}\varphi = \int_{\overline{W}} \left| A(\varphi)\,\frac{\partial\varphi}{\partial u_1} \wedge \frac{\partial\varphi}{\partial u_2} \right|. \tag{3.3}$$

From Proposition 2.2 of [11] and the fact that $|A^{-1}(x)|$ is uniformly bounded for x in any compact subset of D, for each compact $F \subset D$, there exists $\delta > 0$ such that

$$\mathcal{S}\varphi \geq \delta \tag{3.4}$$

for all $\varphi \in \Sigma(\psi, D)$ such that $\varphi(\overline{W}) \subset F$. Let $\varphi_t = x(t,\varphi)$. Then $y_i(t) = \frac{\partial\varphi_t}{\partial u_i}$, $i = 1, 2$, are solutions of the linear variational equation of (1.1)

$$y'(t) = \frac{\partial f}{\partial x}(\varphi_t(u))\,y(t) \tag{3.5}$$

and $z(t) = \frac{\partial\varphi_t}{\partial u_1} \wedge \frac{\partial\varphi_t}{\partial u_2}$ a solution of the second compound equation of (3.5) (see [15], [17]),

$$z'(t) = \frac{\partial f}{\partial x}^{[2]}(\varphi_t(u))\,z(t). \tag{3.6}$$

Straightforward differentiation shows that $w(t) = A(\varphi_t)\,\frac{\partial\varphi_t}{\partial u_1} \wedge \frac{\partial\varphi_t}{\partial u_2}$ satisfies the differential equation $w'(t) = B(\varphi_t(u))\,w(t)$ with B given in (3.2). Suppose $\bar{q}_2 < 0$. Let $2\epsilon_0 = -\bar{q}_2 > 0$. Then there exists $T > 0$ such that, for $t > T$ and $x_0 \in K$,

$$\int_0^t \mu(B(x(s,x_0)))ds \leq -\epsilon_0\, t.$$

From a property of Lozinskiĭ measure,

$$\begin{aligned}
\mathcal{S}\varphi_t &= \int_{\overline{W}} \left| A(\varphi_t)\,\frac{\partial\varphi_t}{\partial u_1} \wedge \frac{\partial\varphi_t}{\partial u_2} \right| \\
&\leq \int_{\overline{W}} \left| A(\varphi)\frac{\partial\varphi}{\partial u_1} \wedge \frac{\partial\varphi}{\partial u_2} \right| \exp\left(\int_0^t \mu(B(\varphi_s(u))\,ds\right) \\
&\leq \mathcal{S}\varphi\, \exp(-\epsilon_0\, t).
\end{aligned}$$

(see [3, p. 41]). Therefore, $\mathcal{S}\varphi_t \to 0$ as $t \to \infty$. This contradicts (3.4) if ψ is invariant with respect to (1.1) since in this case, $\varphi_t \in \Sigma(\psi, D)$ and $\varphi_t(\overline{W}) \subset K$ for all sufficiently large t. We thus have established the following result.

THEOREM 3.1. *Under assumptions* (H_1) *and* (H_2), *if*

$$\bar{q}_2 < 0, \tag{3.7}$$

then no simple closed rectifiable curve in D can be invariant with respect to (1.1). *In particular,* (3.7) *is a Bendixson criterion for* (1.1).

Remarks. (1) When $A = I$ and $|\cdot|$ is chosen as the euclidean norm, the corresponding $\bar{q}_2$ is related to a quantity q_2 defined by Temam ([23, p. 277]) in the context of evolution equations in an infinite-dimensional Hilbert space. Temam defines q_2 over a compact invariant set, whereas $\bar{q}_2$ is defined over a compact absorbing set. Suppose K is a compact absorbing set in D. Then its omega limit set $F = \omega(K)$ is the maximal compact invariant set in D, which is usually called the *global attractor* of (1.1) in D. Suppose q_2 is defined over F. Then it is proved in [23, Chap. V, Prop. 2.1 and Thm. 3.3] that $q_2 < 0$ implies that the Hausdorff dimension of F is less than two. By a result of Smith [22, Thm. 5], E contains no simple closed piecewise smooth invariant curves. In particular, (1.1) has no nonconstant periodic solutions. This shows that $q_2 < 0$ is a Bendixson criterion. Since the global attractor E is not necessarily preserved under a local C^1 perturbation of (1.1) at a nonwandering point, the criterion $q_2 < 0$ may not be robust under such perturbations.

(2) Condition (3.7) is clearly implied by (2.4).

Let $x_0 \in D$ be a nonwandering point such that $f(x_0) \neq 0$. Then, for each sufficiently small neighbourhood U of x_0, there exists $t_1 > 0$ such that $x(t_1, U) \cap U = \emptyset$, and $x(t, U) \cap U \neq \emptyset$ for some $t > t_1$. The following quantities are then well defined.

$$\begin{aligned}\tau(U;\ x_0) = \inf\{\, t > 0\ :\ & x(t,U) \cap U \neq \emptyset, \\ & \text{and } \exists\ t_1 < t \text{ such that } x(t_1, U) \cap U = \emptyset\,\}\end{aligned}$$

and

$$\tau(x_0) = \sup\{\,\tau(U;\ x_0)\ :\ U \text{ is a sufficiently small neighbourhood of } x_0\}; \tag{3.8}$$

when x_0 is an equilibrium, $\tau(x_0)$ is defined to be zero. We call $\tau(x_0)$ the *minimum return time* at the nonwandering point x_0. It follows from the continuous dependence on initial conditions that x_0 is an equilibrium if and only if $\tau(x_0) = 0$. In fact, the following is true.

LEMMA 3.2. *Let x_0 be nonwandering. A solution $x(t, x_0)$ to* (1.1) *is periodic if and only if $\tau(x_0)$ is finite, in which case $\tau(x_0)$ is the minimum period.*

Proof. From the definition, if $\tau(x_0) < \infty$, there exist sequences $x_k \to x_0$, $t_k \to \tau(x_0)$ ($k \to \infty$) such that $x(t_k, x_{2k}) = x_{2k+1}$, which implies that $x(t, x_0)$ is a periodic solution of period $\tau(x_0)$. Thus the least period T satisfies $0 \leq T \leq \tau(x_0)$. Conversely, if $x(t, x_0)$ is a periodic solution of period T, then $\tau(U, x_0) \leq T$ for all sufficiently small neighbourhoods U of x_0 since $x(T, x_0) = x_0 \in U$. Thus $\tau(x_0) \leq T$. The lemma follows from these two observations. □

PROPOSITION 3.3. *Suppose $\tau(x_0) = +\infty$. Then the condition $\bar{q}_2 < 0$ is robust under C^1 local perturbations of f at x_0.*

Proof. Let $\delta = -\bar{q}_2 > 0$. Since K is absorbing, there exists $T > 1$ such that $x(t, K) \subset K$ if $t > T$ and

$$\int_{t_2}^{t_1} \mu\left(B(x(s, x_1))\right) ds \leq -\frac{\delta(t_1 - t_2)}{2} \tag{3.9}$$

for all $t_1, t_2 \geq 0$ such that $t_1 - t_2 > T$ and all $x_1 \in K$. The assumption $\tau(x_0) = +\infty$ implies that $f(x_0) \neq 0$ and $\tau(U; x_0) > T$ for all sufficiently small neighbourhoods U

of x_0. Let Π be a $n-1$-dimensional transversal to the vector $f(x_0)$ at x_0 and U_1 be a sufficiently small ball in Π centered at x_0. Consider the flow box

$$\Sigma = \{x(t, U_1) \ : \ -\alpha \leq t \leq \alpha\}$$

generated by the evolution of the ball $U_1 \subset \Pi$ along the solutions of (1.1) for a small time interval $[-\alpha, \alpha]$ (see Figure 1). Let $\Gamma_+ = x(\alpha, U_1)$ and $\Gamma_- = x(-\alpha, U_1)$. By taking the ball $U_1 \subset \Pi$ and $\alpha > 0$ sufficiently small, we can ensure that all solutions of (1.1) starting in Σ leave Σ and that $\tau(\Sigma; x_0) > T$. As a consequence, each trajectory starting at Γ_+ leaves $\overline{\Sigma}$ and returns to Γ_-, if it ever returns, at a time greater than T.

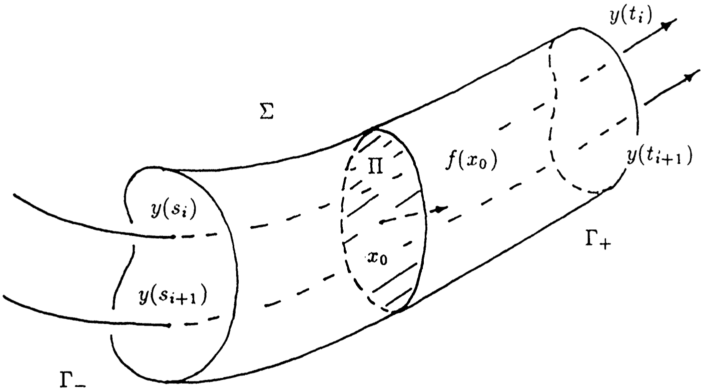


FIG. 1.

Let g be a C^1 local ϵ-perturbation of f at x_0 such that $\text{supp}(f-g) \subset \Sigma$. Consider the differential equation (2.1). K is also absorbing for (2.1) if Σ is sufficiently small since f and g agree on $D \backslash \Sigma$. Let B^f and B^g denote the matrix B defined in (3.2) and $\overline{q}_2(f)$ and $\overline{q}_2(g)$ the quantity $\overline{q}_2$ in (3.1) for f and g, respectively. If the trajectory of a solution $y(t, y_0)$ to (2.1) does not intersect Σ after a certain time, then it coincides with the trajectory of a solution to (1.1) for sufficiently large t. There exists $\bar{t} > 0$ such that no solution of (1.1) and (2.1) remains in $\overline{\Sigma}$ for a time interval greater than $\bar{t}$. For such a solution, it follows from (3.9) that

$$\frac{1}{t}\int_0^t \mu\left(B^g(y(s, y_0))\right) ds \leq -\frac{\delta}{4}.$$

Suppose the trajectory of $y(t, y_0)$ intersects Σ infinitely often. We may assume that $y_0 \in \Gamma_+ \cap K$. Let $t_0 = 0$ and

$$T < s_1 < t_1 < s_2 < t_2 < \cdots < s_n < t_n < s_{n+1} < \cdots$$

be a sequence such that

(i) s_i and t_i are the successive time $y(t, y_0)$ intersects Γ_- and Γ_+, respectively, when it returns to Σ,

(ii) $y(t, y_0) \in \Sigma$, $s_i \leq t \leq t_i$, for each $i \geq 1$,

(iii) $y(t, y_0) \notin \Sigma$, $t_i < t < s_{i+1}$ for each $i \geq 0$.

Then we have

(iv) $t_i - s_i \leq \bar{t}$ for each $i \geq 1$,

(v) $s_{i+1} - t_i > T$ for each $i \geq 0$,

(vi) $y(t, y_0)$ coincides with the solution $x(t, y_i)$ of (1.1) for $t_i < t < s_{i+1}$, where $y_i = y(t_i, y_0)$ for each $i \geq 0$ (see Figure 1).

Since $|f-g|_{C^1} < \epsilon$, we may choose ϵ sufficiently small that

$$|\mu\left(B^f(x)\right) - \mu\left(B^g(y)\right)| < \frac{\delta}{4\bar{t}}$$

for x, y in Σ. Therefore, for each $i \geq 0$,

(3.10)

$$\int_{t_i}^{t_{i+1}} \mu\left(B^g(y(s,y_0))\right) ds = \int_{t_i}^{t_{i+1}} \mu\left(B^f(x(s,y_i))\right) ds + $$

$$\int_{t_i}^{t_{i+1}} \left[\mu\left(B^f(x(s,y_i))\right) - \mu\left(B^g(y(s,y_0))\right)\right] ds$$

$$\leq -\frac{\delta}{2}\,(t_{i+1}-t_i) + (t_{i+1}-s_{i+1})\,\frac{\delta}{4\bar{t}}$$

$$\leq -\frac{\delta}{2}\,(t_{i+1}-t_i) + \frac{\delta}{4} \leq -\frac{\delta}{4}\,(t_{i+1}-t_i)$$

since $t_{i+1}-t_i \geq T > 1$. Thus for all sufficiently large t, $t_n < t \leq t_{n+1}$ for some n, and

$$\frac{1}{t}\int_0^t \mu\left(B^g(y(s,y_0))\right) ds = \frac{1}{t}\int_0^{t_n} \mu(B^g) + \frac{1}{t}\int_{t_n}^{t} \mu(B^g)$$

$$= \frac{1}{t}\sum_{i=0}^{n-1}\int_{t_i}^{t_{i+1}} \mu(B^g) + \frac{1}{t}\int_{t_n}^{t} \mu(B^g)$$

$$\leq -\frac{\delta}{4}\frac{1}{t}\sum_{i=0}^{n-1}(t_{i+1}-t_i) + \frac{1}{t}\int_{t_n}^{t} \mu(B^g)$$

$$\leq -\frac{\delta}{4}\frac{t_n}{t} + \frac{1}{t}\int_{t_n}^{t} \mu\left(B^g\right).$$

If $t-t_n > T$, then, as in (3.10), $\frac{1}{t}\int_{t_n}^{t} \mu(B^g) < -\frac{\delta}{4}\frac{t-t_n}{t}$. Therefore, in this case,

$$\frac{1}{t}\int_0^t \mu(B^g) \leq -\frac{\delta}{4}.$$

If $t-t_n \leq T$, then $\frac{t-t_n}{t} \leq \frac{T}{t}$ and thus $\frac{t_n}{t} \geq 1-\frac{T}{t} > \frac{1}{2}$ when t is sufficiently large. Hence, in this case,

$$\frac{1}{t}\int_0^t \mu(B^g) < -\frac{\delta}{4}\frac{t_n}{t} + \frac{t-t_n}{t}\max_{x\in K}\mu(B^g(x)) < -\frac{\delta}{16}.$$

Therefore, for sufficiently large t and for $y_0 \in K$,

$$\frac{1}{t}\int_0^t \mu\left(B^g(y(s,y_0))\right) ds < -\frac{\delta}{16},$$

which leads to $\bar{q}_2(g) < 0$, completing the proof of the lemma. □

By Theorem 2.4, the results established in Theorem 3.1 and Proposition 3.3 imply that the global stability of the unique equilibrium $\bar{x}$ is equivalent to its local stability under condition (3.7). The following result is in the spirit of Proposition 2.4 in [12]; it deals with the asymptotic behaviour of solutions to (1.1) near an equilibrium under condition (3.7) when multiple equilibria are allowed.

PROPOSITION 3.4. *Under assumptions* (H_1) *and* (H_2), *if* $\bar{q}_2 < 0$, *then the dimension of the stable manifold of any equilibrium is at least* $(n-1)$; *if an equilibrium is not isolated, then its stable manifold has dimension* $(n-1)$ *and it has a centre manifold of dimension one which contains all nearby equilibria.*

Proof. Observe that at an equilibrium x_1, $\bar{q}_2 < 0$ implies

$$\mu\left(A\frac{\partial f^{[2]}}{\partial x}A^{-1}\right) < 0$$

since $f(x_1) = 0$ implies $A_{f(x_1)}(x_1) = 0$. This is inequality (8) in [12], and the rest of the proof is the same as the corresponding part in the proof of Proposition 2.4 in [12]. □

THEOREM 3.5. *Under assumptions* (H_1), (H_2), *and* (H_3), *the unique equilibrium* $\bar{x}$ *is globally stable in* D *if* $\bar{q}_2 < 0$.

Proof. From Theorems 2.4 and 3.1 and Proposition 3.3, it remains to prove the local stability of $\bar{x}$. Assume the contrary. Then $\bar{x}$ is both the omega limit point and alpha limit point of a homoclinic trajectory, which gives rise to a simple closed rectifiable curve γ whose existence is precluded by $\bar{q}_2 < 0$ from Theorem 3.1. The proof for the rectifiability of γ is the same as that given in the proof of Corollary 2.6 in [12]. The key to that proof was the local structure of solutions to (1.1) near equilibria established in Proposition 3.4. □

Remark. In the presence of multiple equilibria, it was proved in [12] and [22] that every nonempty alpha and and omega limit set is a single equilibrium under any of the Bendixson criteria (2.4), (2.6), and (2.7). Results of this type are called *autonomous convergence theorems.* The main ingredients in the proof given in [12] are the C^1 robustness of the Bendixson criterion, a result like Proposition 3.4, and the center manifold theorem (see [4]). Therefore, the same result holds under assumptions (H_1) and (H_2) and our weaker condition $\bar{q}_2 < 0$.

4. Example: An epidemiological model.

Let S, E, I, and R denote the susceptible, exposed, infectious, and recovered fractions in a population. A one-population *SEIRS* model for the spread of an infectious disease in the population is described by the following system of differential equations:

$$\begin{aligned} S' &= -\lambda SI + \nu - \nu S + \delta R, \\ E' &= \lambda IS - (\epsilon + \nu)E, \\ I' &= \epsilon E - (\gamma + \nu)I, \\ R' &= \gamma I - (\delta + \nu)R. \end{aligned} \tag{4.1}$$

Individuals are susceptible, then exposed (in the latent period), then infectious, then recovered with temporary immunity, and then susceptible again when the immunity is lost. The parameter δ describes the rate that the recovered population loses immunity, ϵ is the rate that the exposed population becomes infectious, and γ denotes the rate that the infectious population becomes recovered. There is no disease-related death. The natural death rate and birth rate are assumed to be equal (denoted by ν), and thus $S + E + I + R = 1$ for all time. All parameters are nonnegative. The case $\nu = 0$ corresponds to no death and no birth. If $\delta = 0$, infected individuals recover with permanent immunity; that is, once recovered, they will not become susceptible again. Note that ϵ is assumed to be positive since we consider an infectious disease. We also assume that $\nu + \delta > 0$; otherwise, the model (4.1) is not interesting in that all the population will eventually be recovered and there will be no susceptibles.

The model (4.1) has been extensively studied in the literature; see [14] and its references. It is known that the qualitative behaviour of (4.1) is determined by the contact number $\sigma = \lambda\epsilon/(\epsilon+\nu)(\gamma+\nu)$, which satisfies a threshold condition. If $\sigma \leq 1$, the disease-free equilibrium $P_0 = (1,0,0,0)$ is the only equilibrium and is globally asymptotically stable in the feasible region $\Gamma = \{(S,E,I,R) \in \mathbf{R}^4_+ \ : \ S+E+I+R = 1\}$; namely, the disease dies out. If $\sigma > 1$, then P_0 loses its stability and a unique endemic equilibrium P^* emerges in the interior of Γ and is locally asymptotically stable. It has been conjectured (see [14]) that P^* is globally asymptotically stable in the interior of Γ when $\sigma > 1$ such that the disease remains endemic and approaches a unique endemic equilibrium for all initial configurations.

This conjecture was proved to be true for the case $\delta = 0$ in [13]. A crucial part of the proof is that, when $\delta = 0$, (4.1) can be reduced to a three-dimensional competitive system. Since this property of (4.1) may not be preserved for $\delta > 0$, the method in [13] does not apply to the case $\delta > 0$. In this section, we apply the theory developed in previous sections to show that this conjecture is also true for small δ.

Using $R = 1 - S - E - I$, we can reduce (4.1) to the following three-dimensional system:

$$\begin{aligned} S' &= -\lambda SI + \nu - \nu S + \delta(1 - S - E - I), \\ E' &= \lambda SI - (\epsilon+\nu)E, \\ I' &= \epsilon E - (\gamma+\nu)I, \end{aligned} \tag{4.2}$$

and transform the simplex $\Gamma \subset \mathbf{R}^4_+$ to the following convex region in $\mathbf{R}^3_+$:

$$T = \{(S,E,I) \in \mathbf{R}^3_+ \ : \ 0 \leq S+E+I \leq 1\}.$$

The disease-free equilibrium P_0 becomes $(1,0,0)$ and the endemic equilibrium P^* becomes an interior equilibrium in T. For simplicity of notation, we will denote these two equilibria of (4.2) by P_0 and P^*. The following result can be proved in the same way as in §3 of [14].

PROPOSITION 4.1. *If $\sigma \leq 1$, P_0 is globally asymptotically stable in T. If $\sigma > 1$, P_0 is unstable and the trajectories sufficiently close to P_0 leave P_0 except those on the S-axis which approach P_0 along this axis.*

In the following, we apply Theorem 3.5 to (4.2) to show that P^* is globally stable in the interior of T when $\sigma > 1$.

To show the existence of a compact set which is absorbing in the interior of T is equivalent to proving that (4.2) is uniformly persistent, which means that there exists $c > 0$ such that every solution $(S(t), E(t), I(t))$ of (4.2) with $(S(0), E(0), I(0))$ in the interior of T satisfies

$$\liminf_{t\to\infty} |(S(t), E(t), I(t))| \geq c$$

(cf. [2]). Uniform persistence may also be defined for discrete dynamical systems or iterated maps (see [9]). It can be proved that (4.2) is uniformly persistent if and only if the time-one map associated with (4.2) is uniformly persisitent in the sense of [9].

A compact invariant set $F \subset T$ of (4.2) is said to be *isolated* if there is a neighbourhood $N \subset T$ of F such that F is the maximal invariant subset of N. The *stable set*

F^s of F is the set of $P \in T$ such that the omega limit set $\omega(P) \subset F$. These concepts can be similarly defined for the time-one map associated with (4.2) (see [9]). Using Theorem 4.1 of [9], we can prove the following result.

PROPOSITION 4.2. *System* (4.2) *is uniformly persistent in* T *when* $\sigma > 1$.

Proof. We show that, when $\sigma > 1$, the time-one map associated with (4.2) satisfies the conditions of Theorem 4.1 of [9], namely, (i) the maximal compact invariant set M in the boundary of T is isolated and (ii) the stable set M^s of M is contained in the boundary of T. It can be shown that the time-one map of (4.2) satisfies (i) and (ii) if (4.2) does. Since $M = \{P_0\}$, Proposition 4.1 implies that, when $\sigma > 1$, M^s is contained in the S-axis and thus in the boundary of T. It also implies that M^s is isolated in T. Therefore, the proposition follows from Theorem 4.1 of [9]. □

THEOREM 4.3. *Assume that* $\sigma > 1$. *Then there exists* $\bar{\delta} > 0$ *such that the unique interior equilibrium* P^* *is globally stable in the interior of* T *when* $\delta \leq \bar{\delta}$.

Proof. By Proposition 4.2, when $\sigma > 1$, there exists a compact set K in the interior of T which is absorbing for (4.2). The proof of the theorem consists of choosing a suitable vector norm $|\cdot|$ in $\mathbf{R}^3$ and a 3×3 matrix-valued function $A(x)$ so that the quantity $\bar{q}_2$ defined in (3.1) is negative. We set A as the following diagonal matrix:

$$A(S,E,I) = \operatorname{diag}\left(1,\ \frac{E}{I},\ \frac{E}{I}\right). \tag{4.3}$$

Then A is C^1 and nonsigular in the interior of T. Let f denote the vector field of (4.2). Then

$$A_f A^{-1} = \operatorname{diag}\left(0,\ \frac{I}{E}\left(\frac{E}{I}\right)_f,\ \frac{I}{E}\left(\frac{E}{I}\right)_f\right).$$

The second compound matrix $J^{[2]}$ of the Jacobian matrix $J = \frac{\partial f}{\partial x}$ can be calculated as follows:

$$J^{[2]} = \begin{bmatrix} -\lambda I - \delta - \epsilon - 2\nu & \lambda S & \lambda S + \delta \\ \epsilon & -\lambda I - \delta - \gamma - 2\nu & -\delta \\ 0 & \lambda I & -\epsilon - \gamma - 2\nu \end{bmatrix}$$

(see the appendix of [12]). Therefore, the matrix $B = A_f A^{-1} + A J^{[2]} A^{-1}$ can be written in the following block form:

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \tag{4.4}$$

with

$$B_{11} = -\lambda I - \delta - \epsilon - 2\nu, \quad B_{12} = \left(\frac{\lambda S I}{E},\ \frac{(\lambda S + \delta) I}{E}\right), \quad B_{21} = \left(\frac{\epsilon E}{I},\ 0\right)^*,$$

$$B_{22} = \begin{bmatrix} \frac{I}{E}\left(\frac{E}{I}\right)_f - \lambda I - \delta - \gamma - 2\nu & -\delta \\ \lambda I & \frac{I}{E}\left(\frac{E}{I}\right)_f - \epsilon - \gamma - 2\nu \end{bmatrix}.$$

The vector norm $|\cdot|$ in $\mathbf{R}^3 \cong \mathbf{R}^{\binom{3}{2}}$ is chosen as

$$|(u,v,w)| = \sup\{|u|,\ |v| + |w|\}. \tag{4.5}$$

The Lozinskiĭ measure $\mu(B)$ with respect to $|\cdot|$ can be estimated as follows (see [16] or [18]):

$$\mu(B) \le \sup\{g_1, g_2\}, \tag{4.6}$$

where

$$g_1 = B_{11} + |B_{12}| = -\lambda I - \delta - \epsilon - 2\nu + \frac{(\lambda S + \delta)I}{E}, \tag{4.7}$$

$$g_2 = \mu_1(B_{22}) + |B_{21}| \le \frac{I}{E}\left(\frac{E}{I}\right)_f - \delta - \gamma - 2\nu + \frac{\epsilon E}{I} \tag{4.8}$$

if $\delta \le \epsilon/2$. Note that $\mu_1(B_{22})$ is the Lozinskiĭ measure of the 2×2 matrix B_{22} with respect to the l_1 norm in $\mathbf{R}^2$, $|B_{12}|$ and $|B_{21}|$ are the operator norms of B_{12} and B_{21} when they are regarded as mappings from $\mathbf{R}^2$ to $\mathbf{R}$ and from $\mathbf{R}$ to $\mathbf{R}^2$, respectively, and $\mathbf{R}^2$ is endowed with the l_1 norm. Also note that since B_{11} is a scalar, its Lozinskiĭ measure with respect to any vector norm in $\mathbf{R}^1$ is equal to B_{11}. A solution $(S(t), E(t), I(t))$ to (4.2) with $(S(0), E(0), I(0))$ in the absorbing set K exists for all $t > 0$. From the equations in (4.2), we find

$$\frac{I}{E}\left(\frac{E}{I}\right)_f = \frac{E'}{E} - \frac{I'}{I} \tag{4.9}$$

and

$$\frac{\lambda S I}{E} = \frac{E'}{E} + \epsilon + \nu, \tag{4.10}$$

$$\frac{\epsilon E}{I} = \frac{I'}{I} + \gamma + \nu. \tag{4.11}$$

Relations (4.6)–(4.11) imply

$$\mu(B) \le \frac{E'}{E} - \delta - \nu + \sup\left\{\frac{\delta}{E} - \lambda I,\ 0\right\}.$$

Since (4.2) is uniformly persistent when $\sigma > 1$, there exist $c > 0$ and $T > 0$ such that $t > T$ implies

$$E(t) \ge c, \qquad I(t) \ge c, \qquad \text{and} \qquad \frac{1}{t}\log E(t) < \frac{\delta + \nu}{2}$$

for all $(S(0), E(0), I(0)) \in K$. Set $\bar{\delta} = \min\{\epsilon/2,\ \lambda c^2\}$. Then $t > T$ and $\delta < \bar{\delta}$ imply $\delta/E - \lambda I \le 0$, and thus

$$\frac{1}{t}\int_0^t \mu(B)\,dt < \log E(t) - (\delta + \nu) < -\frac{1}{2}(\delta + \nu)$$

for all $(S(0), E(0), I(0)) \in K$, which in turn implies that $\bar{q}_2 < 0$, completing the proof of Theorem 4.3. □

Acknowledgments. The final draft of this paper was completed when one of the authors (M. Y. Li) was visiting the Centre de Recherches Mathématiques (CRM) at the Université de Montréal as a NSERC postdoctoral fellow. He is grateful for the hospitality of the CRM and the support of the NSERC Postdoctoral Fellowship. The authors would like to thank the referees for their valuable suggestions that helped to improve the original manuscript.

REFERENCES

[1] G. J. Butler, S. B. Hsu, and P. Waltman, *Coexistence of competing predators in a chemostat*, J. Math. Biol., 17 (1983), pp. 133–151.

[2] G. J. Butler and P. Waltman, *Persistence in dynamical systems*, Proc. Amer. Math. Soc., 96 (1986), pp. 425–430.

[3] W. A. Coppel, *Stability and Asymptotic Behavior of Differential Equations*, Heath, Boston, 1965.

[4] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Appl. Math. Sci. 42, Springer-Verlag, New York, 1983.

[5] J. K. Hale, *Ordinary Differential Equations*, John Wiley, New York, 1969.

[6] P. Hartman, *Ordinary Differential Equations*, John Wiley, New York, 1965.

[7] M. W. Hirsch, *Systems of differential equations that are competitive or cooperative* VI: *A local C^r closing lemma for* 3*-dimensional systems*, Ergodic Theory Dynamical Systems, 11 (1991), pp. 443–454.

[8] ———, *Systems of differential equations which are competitive or cooperative* I: *Limit sets*, SIAM J. Math. Anal., 13 (1982), pp. 167–179.

[9] J. Hofbauer and J. W.-H. So, *Uniform persistence and repellors for maps*, Proc. Amer. Math. Soc., 107 (1989), pp. 1137–1142.

[10] J. P. LaSalle, *The Stability of Dynamical Systems*, Regional Conference Series in Applied Mathematics 25, Society for Industrial and Applied Mathematics, Philadephia, 1976.

[11] Y. Li and J. S. Muldowney, *On Bendixson's criterion*, J. Differential Equations, 106 (1994), pp. 27–39.

[12] Y. M. Li and J. S. Muldowney, *On R. A. Smith's autonomous convergence theorem*, Rocky Mountain J. Math., 25 (1995), pp. 365–379.

[13] ———, *Global Stability for the SEIR model in epidimiology*, Math. Biosci., 125 (1995), pp. 155–164.

[14] W-M. Liu, H. W. Hethcote, and S. A. Levin, *Dynamical behavior of epidemiological models with nonlinear incidence rates*, J. Math. Biol., 25 (1987), pp. 359–380.

[15] D. London, *On derivations arising in differential equations*, Linear and Multilinear Algebra, 4 (1976), pp. 179–189.

[16] R. H. Martin, Jr., *Logarithmic norms and projections applied to linear differential systems*, J. Math. Anal. Appl., 45 (1974), pp. 432–454.

[17] J. S. Muldowney, *Compound matrices and ordinary differential equations*, Rocky Mountain J. Math., 20 (1990), pp. 857–872.

[18] ———, *Dichotomies and asymptotic behaviour for linear differential systems*, Trans. Amer. Math. Soc., 283 (1984), pp. 465–484.

[19] C. C. Pugh, *An improved closing lemma and the general density theorem*, Amer. J. Math., 89 (1967), pp. 1010–1021.

[20] C. C. Pugh and C. Robinson, *The C^1 closing lemma including Hamiltonians*, Ergodic Theory Dynamical Systems, 3 (1983), pp. 261–313.

[21] H. L. Smith, *Systems of ordinary differential equations which generate an order preserving flow*, SIAM Rev., 30 (1988), pp. 87–113.

[22] R. A. Smith, *Some applications of Hausdorff dimension inequalities for ordinary differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 104 (1986), pp. 235–259.

[23] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci. 68, Springer-Verlag, New York, 1988.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 1084–1094, July 1996

013

BIFURCATION OF FIXED POINTS IN COUPLED JOSEPHSON JUNCTIONS*

M. ST. VINCENT[†]

Abstract. The fixed points of the equations describing a pair of coupled Josephson junctions are investigated and are shown to undergo an infinite number of bifurcations as parameters are varied. The bifurcation curves are then analyzed in a region of parameter space where interesting dynamical behavior is also known to occur. It is shown that the bifurcation curves fall into four one-parameter families, and the curves in each family are described. The paper concludes with a discussion of a conjectured mechanism by which simultaneous bifurcations could give rise to interesting dynamical solutions.

Key words. bifurcation, fixed point, Josephson junctions, coupled oscillators, coupled pendulums

AMS subject classification. 34C23

1. Introduction.

In this paper, we investigate the fixed points of the system

$$\ddot{\varphi}_1 + \gamma\dot{\varphi}_1 + \sin\varphi_1 + k(\varphi_1 - \varphi_2 + H) = I, \tag{1.1a}$$

$$\ddot{\varphi}_2 + \gamma\dot{\varphi}_2 + \sin\varphi_2 - k(\varphi_1 - \varphi_2 + H) = 0, \tag{1.1b}$$

which has been used to describe the behavior of a pair of coupled Josephson junctions in the current-driven case [1–3] ((1.1) also describes the behavior of a pair of damped nonlinear pendulums connected along a common axis of rotation by a linear spring, with one of them driven by a constant torque I, as stated in [2, 4]). Here γ, H, I, and k are constants, with the damping γ and coupling parameter k both positive. Following Imry and Schulman [1], we observe that the case of $I < 0$ can be transformed to $I > 0$ by replacing solutions φ_1, φ_2 by $-\varphi_1$, $-\varphi_2$, and so will also require I to be positive.

Previous work on (1.1) has focused on its dynamic behavior. Some discussion of fixed points was presented by Imry and Schulman [1], and a more comprehensive discussion of equilibria in a related system that contains the case $H = 0$ was done by Henderson, Levi, and Odeh in [6]. The results of this paper provide a thorough analysis of the fixed points of (1.1) and their bifurcations, especially in the region of weak coupling (small k), and provide support for a conjectured mechanism by which bifurcating fixed points may give rise to dynamical behavior. The interesting dynamical behavior investigated by Levi [2, 4] and Imry and Schulman [1] occurs in that region also.

In the work done here, γ and H are fixed, and we investigate the behavior of the fixed points as the parameters I and k vary. As will be seen, fixed points can occur only for $|I| < 2$, and occur for any such I whenever k is small enough. Furthermore, since the flow $(\varphi_1, \dot{\varphi}_1, \varphi_2, \dot{\varphi}_2)$ induced on $\mathbb{R}^4$ by (1.1) commutes with translation by $(2\pi, 0, 2\pi, 0)$, each fixed point has an infinite number of equivalent copies. When points of $\mathbb{R}^4$ that differ by such a translation are identified, the flow on the resulting phase space $S^1 \times \mathbb{R}^3$ has only a finite number of fixed points, and it will be shown that the number of fixed points is $O(1/k)$ as $k \to \infty$.

From the dependence of the number of fixed points on I and k, it is clear that the fixed points of (1.1) undergo a large number of bifurcations as these parameters are

* Received by the editors October 8, 1992; accepted for publication (in revised form) December 6, 1994.

[†] Department of Mathematics and Computer Science, Merrimack College, North Andover, MA 01845.

varied. As was shown by Imry and Schulman [1], these must all be static bifurcations in which a pair of fixed points merge and disappear. It will be shown here that these bifurcations are of two types: (1) those in which a sink merges with a saddle having one eigenvalue with positive real part, and (2) those in which a saddle having one eigenvalue with positive real part merges with one having two. Conditions for the occurrence of these bifurcations will be analyzed for (I,k) in the set

$$S=\left\{(I,k):0<I<2, 0<k<\min\left(\frac{1}{2}\sqrt{1-(I-1)^2},\ \frac{1}{2}\sqrt{1-I^2/4}\right)\right\},$$

allowing us to obtain bifurcation diagrams there (Figs. 2a, 2b).

The bifurcation diagrams in S make clear the existence of a countable number of points at which simultaneous bifurcations occur. This has suggested a possible connection with some of the *running periodic* solutions of (1.1). These are solutions for which $\varphi_i(t+T)=\varphi_i(t)+2m\pi$, $i=1,2$, where T is positive and m is an integer. It turns out that m must also be positive (this follows from a slight modification of an argument in [2]). These solutions are periodic in the cylindrical phase space $S^1\times\mathbb{R}^3$ but not in $\mathbb{R}^4$. Among the most interesting of these are the so-called *beating* solutions (Imry and Schulman [1] and Levi [2, 4]). These are running periodic solutions in which first one oscillator is nearly constant while the other runs through a number of oscillations, and then they exchange roles for the remainder of the period. In the two transitions the solution is for a short while nearly static. This led Kopell [5] to speculate that such solutions might arise from a pair of simultaneous sink-saddle bifurcations in which the closure of the unstable manifolds of the saddles forms a closed curve in the $S^1\times\mathbb{R}^3$ phase space, with each unstable manifold connecting the sinks. This closed curve would become the orbit of the beating solution when the fixed points merge and disappear (see Fig. 3). Support for this conjecture will be provided by demonstrating that there are a countable number of points in S at which simultaneous sink-saddle bifurcations occur, although no attempt will be made to show that the unstable manifolds ever connect the sinks. Furthermore, it will be seen that some of these bifurcations occur in the small region of S where Imry and Schulman found beating solutions and involve fixed points near the nearly static transitions in those beating solutions. It should be noted that beating solutions also occur for the systems in [6–8].

The remainder of the paper is organized as follows. Section 2 contains general results on the existence and stability of fixed points that are valid for all parameter values. In particular, it is shown that fixed points correspond to the intersection of certain horizontal lines with a closed curve C that depends on I and k. Further progress is made in §3 by restricting attention to values of (I,k) in S. This allows estimates to be obtained that make it possible to determine the shape of C (see Fig. 1), and that will be needed later. It is then shown how the character of a fixed point is determined by the location of the corresponding point on C. Section 4 contains an analysis of the bifurcation curves in S. It is shown that these bifurcation curves fall into four one-parameter families, and descriptions of the curves in each family are obtained. This is the main result of the paper and is contained in Theorem 4.2. The paper ends with a demonstration that the results obtained are consistent with the conjectures on beating solutions.

2. Existence and stability of fixed points. As previously mentioned, the analysis of (1.1) can be simplified by restricting our attention to the $S^1\times\mathbb{R}^3$ phase space. This will be done here by representing each equivalence class of fixed points by its representative $(\varphi_1,0,\varphi_2,0)$ with $\varphi_1\in[-\pi/2,3\pi/2)$. From (1.1) we see immediately

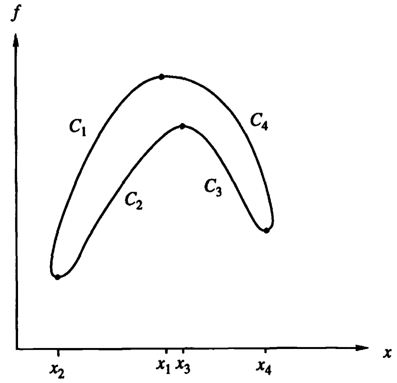


FIG. 1. *A sketch of the curve $C = C_1 \cup C_2 \cup C_3 \cup C_4$ for values of (I, k) in S with the locations of the critical points $x_1, \ldots, x_4$ indicated. Each intersection of C by a horizontal line of height $I - k(2L\pi + h)$ corresponds to an equivalence class of fixed points, each having φ_1 equal to the x-coordinate of the point on C* (mod 2π)*. Bifurcations occur when these correspond to one of the critical points.* (*Note: certain features of C have been exaggerated for purposes of illustration.*)

that $(\varphi_1, 0, \varphi_2, 0)$ will be a fixed point if and only if the constants φ_1, φ_2 satisfy

$$\sin\varphi_1 + k(\varphi_1 - \varphi_2 + H) = I,$$

$$\sin\varphi_2 - k(\varphi_1 - \varphi_2 + H) = 0,$$

which is equivalent to

$$\sin\varphi_1 + k(\varphi_1 - \varphi_2 + H) = I, \tag{2.1a}$$

$$\sin\varphi_1 + \sin\varphi_2 = I. \tag{2.1b}$$

From (2.1b), it follows that $|I| \le 2$ is necessary for the existence of fixed points, and that φ_2 must be given by either

$$\varphi_2 = \sin^{-1}(I - \sin\varphi_1) - 2L\pi \tag{2.2.1}$$

or

$$\varphi_2 = \pi - \sin^{-1}(I - \sin\varphi_1) - 2L\pi \tag{2.2.2}$$

for some integer L. Together with (2.1a), this shows that if $0 < I < 2$ and $\varphi_1 \in [-\pi/2, 3\pi/2)$, then $(\varphi_1, 0, \varphi_2, 0)$ will be a fixed point if and only if there is an integer L such that φ_1 satisfies

$$f_i(\varphi_1) = I - k(2L\pi + H) \tag{2.3.i}$$

for either $i = 1$ or $i = 2$, and φ_2 is given by the corresponding equation (2.2.i). The functions $f_i : J_I \to \mathbb{R}$ $(i = 1, 2)$ are given by

$$f_1(x) = \sin x + k[x - \sin^{-1}(I - \sin x)],$$

$$f_2(x) = \sin x + k[x + \sin^{-1}(I - \sin x) - \pi],$$

with domain $J_I = [\sin^{-1}(I-1),\ \pi - \sin^{-1}(I-1)]$. To emphasize their dependence on parameters, $f_i(x; I, k)$ will sometimes be written for $f_i(x)$.

As is readily verified, the functions $f_i(x)$ are continuous on J_I and smooth on its interior (with respect to I and k as well as x) and agree at its endpoints. Furthermore, $f_1(x) > f_2(x)$ on the interior of J_I, with $f_1'(x) \to +\infty$, $f_2' \to -\infty$ as x approaches the left endpoint and $f_1'(x) \to -\infty$, $f_2'(x) \to +\infty$ as x approaches the right endpoint. Consequently, the graphs of $f_i(x)$ together form a smooth closed curve C. As shown by (2.3), we now have the following proposition.

PROPOSITION 2.1. *Each equivalence class of fixed points corresponds to an intersection of C by a horizontal line of height $I - k(2L\pi + H)$.*

Since the horizontal lines are separated by a distance $2\pi k$ and the functions f_i have O(1) variation across J_I, (1.1) must have O($1/k$) nonequivalent fixed points as $k \to 0$, whenever $0 < I < 2$. Further results will be obtained from the shape of C, which the restriction to $(I, k) \in S$, made in the next section, will make it possible to determine (see Fig. 1).

To consider the stability of fixed points, let $\nu_i = \dot{\varphi}_i$ and rewrite (1.1) as the first-order system

$$\frac{d}{dt}\begin{pmatrix} \varphi_1 \\ \nu_1 \\ \varphi_2 \\ \nu_2 \end{pmatrix} = \begin{pmatrix} \nu_1 \\ I - k(\varphi_1 - \varphi_2 + H) - \sin\varphi_1 - \gamma\nu_1 \\ \nu_2 \\ k(\varphi_1 - \varphi_2 + H) - \sin\varphi_2 - \gamma\nu_2 \end{pmatrix}.$$

The linearization of this system about a fixed point $(\varphi_1, 0, \varphi_2, 0)$ has coefficient matrix

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ -(k+\cos\varphi_1) & -\gamma & k & 0 \\ 0 & 0 & 0 & 1 \\ k & 0 & -(k+\cos\varphi_2) & -\gamma \end{pmatrix},$$

which has eigenvalues

$$\frac{1}{2}\left(-\gamma \pm \sqrt{\gamma^2 - 4a_+}\right) \text{ and } \frac{1}{2}\left(-\gamma \pm \sqrt{\gamma^2 - 4a_-}\right),$$

where

$$a_{\pm} = k + \frac{\cos\varphi_1 + \cos\varphi_2}{2} \pm \left[\left(\frac{\cos\varphi_1 - \cos\varphi_2}{2}\right)^2 + k^2\right]^{1/2}.$$

Thus each fixed point has at least two eigenvalues with a negative real part, with the remaining eigenvalues having real parts whose signs are determined by the relative magnitudes of

$$\left|k + \frac{\cos\varphi_1 + \cos\varphi_2}{2}\right| \text{ and } \left[\left(\frac{\cos\varphi_1 - \cos\varphi_2}{2}\right)^2 + k^2\right]^{1/2}$$

and the sign of $k + (\cos\varphi_1 + \cos\varphi_2)/2$.

Now define

$$g(\varphi_1, \varphi_2) = \cos\varphi_1\cos\varphi_2 + k(\cos\varphi_1 + \cos\varphi_2)$$

as in [1], and let

$$h(\varphi_1, \varphi_2) = k + (\cos\varphi_1 + \cos\varphi_2)/2.$$

It is easily verified that $|h(\varphi_1, \varphi_2)|$ has the same relation $(>, =, <)$ to $[((\cos\varphi_1 - \cos\varphi_2)/2)^2 + k^2]^{1/2}$ as $g(\varphi_1, \varphi_2)$ has to zero, so we now have the following proposition.

PROPOSITION 2.2. *An equilibrium solution* $(\varphi_1, 0, \varphi_2, 0)$ *will be*

(i) *a sink if* $g(\varphi_1, \varphi_2) > 0$ *and* $h(\varphi_1, \varphi_2) > 0$,

(ii) *a saddle having a one-dimensional unstable manifold if* $g(\varphi_1, \varphi_2) < 0$,

(iii) *a saddle having a two-dimensional unstable manifold if* $g(\varphi_1, \varphi_2) > 0$ *and* $h(\varphi_1, \varphi_2) < 0$.

Since the equilibrium solution is structurally stable in each of the above cases, and since h cannot be zero if g is positive, bifurcations can occur only when $g(\varphi_1, \varphi_2) = 0$. In that case there will be exactly one eigenvalue with real part zero. This rules out the possibility of Hopf bifurcations, which require the existence of a fixed point having two eigenvalues on the imaginary axis. We note also that the types of equilibria described above are the only types that occur for the related systems in [6–8].

We now show that the sign of g at a fixed point can be determined by the location of the corresponding point where the curve C is intersected by a horizontal line of height $I - k(2L\pi + H)$. Since φ_2 is given by (2.2.i) when this point lies on the part of C given by the graph of f_i, we define functions $g_i : J_I \to \mathbb{R}$ $(i = 1, 2)$ by $g_1(x) = g(x, \sin^{-1}(I - \sin x))$ and $g_2(x) = g(x, \pi - \sin^{-1}(I - \sin x))$. Thus

$$g_1(x) = (\cos x)\sqrt{1 - (I - \sin x)^2} + k(\cos x + \sqrt{1 - (I - \sin x)^2}),$$

$$g_2(x) = -(\cos x)\sqrt{1 - (I - \sin x)^2} + k(\cos x - \sqrt{1 - (I - \sin x)^2}).$$

Comparison of $g_i(x)$ with $f_i'(x)$ shows that

$$(2.4) \qquad g_1(x) = f_1'(x)\sqrt{1 - (I - \sin x)^2} \text{ and } g_2(x) = -f_2'(x)\sqrt{1 - (I - \sin x)^2}$$

on the interior of J_I. Since the term under the radical is zero only at the endpoints of J_I, the sign of g at an equilibrium solution is determined by the slope at the corresponding point of C and the segment (graph of f_1 or f_2) it lies on. It is also easily verified that $g_i \neq 0$ at the endpoints of J_I for $0 < I < 2$, so we now have the next proposition.

PROPOSITION 2.3. *Bifurcations can only occur at fixed points that correspond to a point at which the curve C has a horizontal tangent.*

3. Restriction to parameter region S.

From now on we will only consider $(I, k) \in S$. This will allow us to determine the shape of C (Fig. 1). In particular, we will see that C has a horizontal tangent at four points. Bifurcations will occur only when a horizontal line of height $I - k(2L\pi + H)$ passes through one of these critical points, and the stability of nonbifurcating equilibria will be determined by which of the four segments bounded by the critical points the corresponding point of C lies on.

To begin, observe that the shape of C is determined by the way in which g_1 and g_2 change sign across J_I, as shown by (2.4). In order to understand the behavior of g_1 and g_2, we need some observations about $\cos x$ and $[1 - (I - \sin x)^2]^{1/2}$ on J_I. As a notational convenience, define $\sigma : J_I \to \mathbb{R}$ by $\sigma(x) = [1 - (I - \sin x)^2]^{1/2}$. We then have Lemma 3.1.

LEMMA 3.1. *If $(I, k) \in S$, then there are points a_i, b_i in $(\sin^{-1}(I - 1), \pi/2) \subset J_I$ and their reflections $a_i' = \pi - a_i$, $b_i' = \pi - b_i$ in $(\pi/2, \pi - \sin^{-1}(I - 1)) \subset J_I (i = 1, 2, 3)$ with $a_1 < a_2 < a_3 < b_1 < b_2 < b_3$ and $b_1 > 0$ such that*

(i) $\sigma(a_1) = k/2$, $\sigma(a_2) = k$, $\sigma(a_3) = 2k$, *and* $\sigma(a_i') = \sigma(a_i)(i = 1, 2, 3)$;

(ii) $\cos b_1 = 2k$, $\cos b_2 = k$, $\cos b_3 = k/2$, *and* $\cos b_i' = -\cos b_i (i = 1, 2, 3)$;

(iii) σ *is increasing on* $[\sin^{-1}(I-1), a_3]$ *and decreasing on* $[a_3', \pi - \sin^{-1}(I-1)]$ *with* $\sigma > 2k$ *on* (a_3, a_3');

(iv) $\cos x$ *is decreasing on* $[b_1, b_1']$ *with* $\cos x > 2k$ *on* $[\sin^{-1}(I-1), b_1)$ *and* $\cos x < -2k$ *on* $(b_1', \pi - \sin^{-1}(I-1)]$.

Proof. First, observe that σ has a critical point at $\pi/2$ and that $\sigma(\pi/2) = [1 - (I-1)^2]^{1/2}$ is greater than $2k$ since $(I, k) \in S$. Furthermore, σ is zero only at the endpoints of J_I and is symmetrical about $\pi/2$. If $1 \leq I < 2$, then σ is increasing on $[\sin^{-1}(I-1), \pi/2]$, while if $0 < I < 1$, then σ increases until it reaches its maximum value at a point of $(\sin^{-1}(I-1), \pi/2)$ and then decreases to its value at $\pi/2$. In either case, (i) and (ii) follow immediately, with $a_1 < a_2 < a_3$ all in $(\sin^{-1}(I-1), \pi/2)$.

Now observe that $2k < 1$ since $k < \sqrt{2}/3$ in S, so we can define $b_1 = \cos^{-1}(2k)$, etc. Also, $\cos x$ is decreasing on $[b_1, b_1']$ since $[b_1, b_1'] \subset (0, \pi)$, and $b_1 < b_2 < b_3$ are in $(\sin^{-1}(I-1), \pi/2)$ since $\cos(\sin^{-1}(I-1)) = [1-(I-1)^2]^{1/2}$ is greater than $2k$ in S. The inequalities in (iv) now follow easily, so (ii) and (iv) are proved.

All that remains is to show that $a_3 < b_1$. But we must have $a_3 < \sin^{-1}(I/2) < b_1$, since $\sigma(\sin^{-1}(I/2)) = \cos(\sin^{-1}(I/2)) = (1 - I^2/4)^{1/2}$ is greater than $2k$ for $(I, k) \in S$. □

We can now obtain enough information about g_1 and g_2 to determine the shape of C when $(I, k) \in S$. The results are stated in Proposition 3.2.

PROPOSITION 3.2. *If* $(I, k) \in S$, *then*

(i) f_1 *has a local maximum at a point* x_1 *in* (b_3', b_2') *and*

(ii) f_2 *has a local maximum at a point* x_3 *in* (b_2', b_1') *and a pair of local minima at* x_2 *in* (a_1, a_2) *and* x_4 *in* (a_3', a_2').

Furthermore, f_1 *and* f_2 *do not have any other critical points.*

Proof. In view of (2.4), it is sufficient to show that g_1 is positive on $[\sin^{-1}(I-1), b_3']$ and negative on $[b_2', \pi - \sin^{-1}(I-1)]$, with $g_1' < 0$ on $[b_3', b_2']$, and, similarly, that g_2 is positive on $[\sin^{-1}(I-1), a_1] \cup [b_1', a_3']$ and negative on $[a_2, b_2'] \cup [a_2', \pi - \sin^{-1}(I-1)]$, with $g_2' < 0$ on $[a_1, a_2] \cup [a_3', a_2']$ and $g_2' > 0$ on $[b_2', b_1']$.

To show that g_2 is positive on $[b_1', a_3']$, observe that $\cos x < -2k$ on $(b_1', a_3']$ and $\sigma(x) > 2k$ on $[b_1, a_3')$ from Lemma 3.1, so that $g_2(x) = (k - \frac{1}{2}\sigma(x))\cos x - (k + \frac{1}{2}\cos x)\sigma(x)$ is clearly positive on $[b_1', a_3']$. Similarly, to show that $g_2' < 0$ on $[a_1, a_2]$, observe that $g_2' = -[(k + \cos x)(I - \sin x)\cos x]/\sigma - (k - \sigma)\sin x$. Also, on $[a_1, a_2]$ we have $I - \sin x > 0$ (since σ is increasing), $k/2 \leq \sigma \leq k$, and $\cos x > 2k$. Then $I - \sin x \geq (1 - k^2)^{1/2}$ since $\sigma \leq k$, and $|\sin x| < (1 - 4k^2)^{1/2}$ since $\cos x > 2k$, so $g_2' < -[3k(1-k^2)^{1/2} 2k]/k + (k/2)(1-4k^2)^{1/2} = (k/2)(1-4k^2)^{1/2} - 6k(1-k^2)^{1/2} < 0$.

The other proofs are similar and are thus omitted. □

It can also be shown that $f_2(x_2) < f_2(x_4)$, so that f_2 takes its global minimum only at x_2. Thus the shape of C must be as indicated in Fig. 1.

The critical points divide C into four segments, which we label $C_1, \ldots, C_4$ as in Fig. 1. We now show that the stability of an equilibrium solution is completely determined by which of these segments the corresponding point of C lies on. To begin, observe that (2.4) shows that g will be positive at a fixed point if the corresponding point lies on C_1 or C_3 and will be negative if it lies on C_2 or C_4. By Proposition 2.2, this determines the nature of equilibria whose corresponding points lie on C_2 or C_4, but for the rest we also need to know the sign of h. To this end, we make use of (2.2) to define

$$h_1(x) = h(x, \sin^{-1}(I - \sin x)) = (2k + \sigma(x) + \cos x)/2$$

and

$$h_2(x) = h(x, \pi - \sin^{-1}(I - \sin x)) = (2k - \sigma(x) + \cos x)/2.$$

From Lemma 3.1 we see that h_1 is positive on $[\sin^{-1}(I-1), b_1']$ and that h_2 is positive on $[\sin^{-1}(I-1), a_3]$ and negative on $[\pi/2, \pi - \sin^{-1}(I-1)]$. Together with Proposition 2.2, this now yields Theorem 3.3.

THEOREM 3.3. *If (I,k) is in S, then an equilibrium solution $(\varphi_1, 0, \varphi_2, 0)$ will be*

(i) *a sink if the corresponding point lies on C_1,*

(ii) *a saddle having a one-dimensional unstable manifold if the corresponding point lies on C_2 or C_4,*

(iii) *a saddle having a two-dimensional unstable manifold if the corresponding point lies on C_3.*

Recalling Proposition 2.3, it is now clear that when $(I,k) \in S$, a bifurcation will occur at an equilibrium solution if and only if it corresponds to one of the critical points $x_1, \ldots, x_4$ of C.

4. Existence of bifurcation points in S. Now define functions F_i $(i = 1, 2, 3, 4)$ by

$$F_1 = f_1(x_1(I,k), I, k) - I + k(2L\pi + H),$$

$$F_i = f_2(x_i(I,k), I, k) - I + k(2L\pi + H) \quad (i = 2, 3, 4).$$

If (I,k) is in S, then a horizontal line of height $I - k(2L\pi + H)$ will pass through the critical point $x_i(I,k)$ of C if and only if $F_i = 0$ for those values of I, k, L, and H. Thus bifurcation points in S correspond to zeros of the functions F_i and, as will be seen, for any given $H \in \mathbb{R}$ each function F_i will determine a one-parameter family of bifurcation curves in S parameterized by the integer L. To analyze these curves, we need estimates for the functions F_i and their derivatives

$$\frac{\partial F_1}{\partial k} = x_1 - \sin^{-1}(I - \sin x_1) + 2L\pi + H,$$

$$\frac{\partial F_1}{\partial I} = -[k/\sigma(x_1)] - 1,$$

$$\frac{\partial F_i}{\partial k} = x_i + \sin^{-1}(I - \sin x_i) + (2L-1)\pi + H,$$

$$\frac{\partial F_i}{\partial I} = [k/\sigma(x_i)] - 1 \quad (i = 2, 3, 4).$$

The required estimates are easily obtained from Lemma 3.1 and are contained in the following lemma. The proof is elementary and so is omitted.

LEMMA 4.1. *Given any $H \in \mathbb{R}$ and integer L, the functions F_i satisfy the following for all (I,k) in S.*

$$\text{(i)} \quad \sqrt{1-k^2} - I + k(2L\pi + H) < F_1 < \sqrt{1 - k^2/4} - I + k[(2L+1)\pi + H],$$

$$-\sqrt{1-k^2/4} + k[(2L-1)\pi + H) < F_2 < -\sqrt{1-k^2} + k(2L\pi + H),$$

$$\sqrt{1-4k^2} - I + k[(2L-1)\pi + H] < F_3 < \sqrt{1-k^2} - I + k[(2L+1/2)\pi + H],$$

$$-\sqrt{1-k^2} + k[(2L-1/2)\pi + H] < F_4 < -\sqrt{1-4k^2} + k[(2L+1)\pi + H].$$

(ii)
$$2L\pi + H < \frac{\partial F_1}{\partial k} < (2L+1)\pi + H,$$

$$(2L-1)\pi + H < \frac{\partial F_2}{\partial k} < 2L\pi + H,$$

$$(2L-1)\pi + H < \frac{\partial F_i}{\partial k} < (2L+1/2)\pi + H, \quad (i = 3,4).$$

(iii)
$$-\frac{3}{2} < \frac{\partial F_1}{\partial I} < -1, 0 < \frac{\partial F_2}{\partial I} < 1,$$

$$-1 < \frac{\partial F_3}{\partial I} < -\frac{1}{2}, \quad -\frac{1}{2} < \frac{\partial F_4}{\partial I} < 0.$$

Lemma 4.1 contains most of what we will need to analyze the bifurcation curves in S. The results are contained in Theorem 4.2.

THEOREM 4.2. *Let H be given. For each of the functions F_i, the zeros in S form a family of smooth curves, each associated with a value of the integer L. Furthermore, we have the following.*

(i) *For each L for which $2L\pi + H \notin [-\pi, 0]$, the zeros of F_1 in S form a smooth curve having one endpoint at $(I,k) = (1,0)$ and the other on the upper boundary of S. Curves in the subfamily for which $2L\pi + H > 0$ have positive slope $k'(I)$, while those for which $2L\pi + H < -\pi$ have negative slope. Each curve lies above those of its subfamily that corresponds to larger values of $|L|$.*

(ii) *For each L for which $2L\pi + H \geq \pi + 4\sqrt{7}/3$, the zeros of F_2 in S form a smooth curve with endpoints that are on the upper boundary of S and opposite sides of the line $I = 2/3$. Each of these curves has negative slope and lies entirely above those that correspond to larger values of L. F_2 does not have any zeros in S for those values of L for which $2L\pi + H \leq 3/\sqrt{2}$.*

(iii) *For each L for which $2L\pi + H \notin [-\pi/2, \pi]$, the zeros of F_3 in S form a smooth curve having one endpoint at $(I,k) = (1,0)$, and the other on the upper boundary of S. Curves in the subfamily for which $2L\pi + H > \pi$ have positive slope, while those for which $2L\pi + H < -\pi/2$ have negative slope. Each curve lies above those of its subfamily that correspond to larger values of $|L|$.*

(iv) *For each L for which $2L\pi + H \geq \pi + \sqrt{3}$, the zeros of F_4 in S form a smooth curve with endpoints that are on the upper boundary of S and opposite sides of the line $I = 2/3$. Each of these curves has positive slope and lies entirely above those that correspond to larger values of L. F_4 does not have any zeros in S for those values of L for which $2L\pi + H \leq 3/\sqrt{2} - \pi/2$.*

In each case, the curves accumulate on the I-axis as $L \longrightarrow +\infty$, and as $L \longrightarrow -\infty$ in cases (i) *and* (iii).

Proof. That the zeros of F_i in S form smooth curves follows from the fact that $\partial F_i/\partial I$ is never zero in S by (iii) of Lemma 4.1. Proofs of (i) and (ii) of Theorem 4.2 follow. The proofs of (iii) and (iv) are done in the same way.

(i) First, observe that $F_1 = 1 - I$ on the lower boundary ($k = 0$) of S, by (i) of Lemma 4.1. Now consider the case when $2L\pi + H > 0$. Then $\partial F_1/\partial k$ is bounded above zero in S from Lemma 4.1, so F_1 will have a unique zero in S for each value of I in (1, 2) that is close enough to 1 but will not have any for $I \leq 1$. The curve formed by these zeros must have positive slope since $\partial F_1/\partial I < 0$ from Lemma 4.1. Consequently, the other endpoint must lie on the upper boundary of S. Furthermore, it is clear from $\partial F_1/\partial I < 0$ that F_1 cannot have any other zeros in S for this value of

L. That each curve in this subfamily lies above those corresponding to larger values of L follows easily from (ii) of Lemma 4.1, since the lower bound for $\partial F_1/\partial k$ when $L = j+1$ is larger than the upper bound when $L = j$ for each integer j.

The case when $2L\pi + H < -\pi$ is similar but with $\partial F_1/\partial k$ bounded below zero.

(ii) Observe that $F_2 = -1$ on the lower boundary ($k = 0$) of S, and that the largest value of k to occur in the closure of S is $\sqrt{2}/3$ and occurs only at the point $(I,k) = (2/3, \sqrt{2}/3)$ on the upper boundary of S. To see that F_2 cannot have zeros in S when $2L\pi + H < 3/\sqrt{2}$, notice that then $\partial F_2/\partial k < 3/\sqrt{2}$ from Lemma 4.1. But then the maximum value of F_2 in the closure of S must be less than $-1+(\sqrt{2}/3)(3/\sqrt{2}) = 0$.

Now let L be such that $2L\pi + H \geq \pi + 4\sqrt{7}/3$. Then $\partial F_2/\partial k \geq 4\sqrt{7}/3$ in S from Lemma 4.1. At any point of S for which $k \geq \sqrt{7}/8$, we have $F_2 > -1 + (\sqrt{7}/8)(4\sqrt{7}/3) = 1/6$. Since such points occur for all $I \in (1/4, 3/2)$, it follows that F_2 has zeros in S for all such I. The curve formed by these zeros has negative slope since $\partial F_2/\partial I$ and $\partial F_2/\partial k$ are both positive and clearly must have endpoints on the upper boundary of S. That F_2 cannot have any other zeros in S for this value of L follows from the fact that they would form curves that have endpoints on the upper boundary of S and must either lie to the left of $I = 1/4$ or to the right of $I = 3/2$. But F_2 cannot have more than one zero on the upper boundary of S to the left of $I = 1/4$ since $dF_2/dI > 0$ there. Similarly, F_2 cannot have more than one zero to the right of $I = 3/2$ on the upper boundary of S since $dF_2/dI < 0$ there.

All that remains is to show that each curve of this family lies entirely above those that correspond to larger values of L. But this follows from (ii) of Lemma 4.1 in the same way as the corresponding claim in (i). □

Theorem 4.2 is the main result of this paper. Although it does not describe the bifurcation curves outside of S, it provides a nearly complete description of them in S. As is clear from the theorem, the bifurcation curves in case (iii) appear similar to those in case (i). Similarly, those in case (iv) appear like those in case (ii), though with slopes of opposite signs. The curves in cases (i) and (ii) are sketched in Figs. 2a and b.

Remark. We now have enough information about the bifurcations in S to determine whether or not the conjecture by Kopell [5] that beating solutions could arise from simultaneous sink-saddle bifurcations is consistent with what is known about the beating solutions from numerical experiments (see Fig. 3). Let L_1 and L_2 be integers such that the line of height $I - k(2L_i\pi + H)$ passes through the critical point of C at x_i ($i = 1, 2$). If the conjecture is correct, then we would expect the beating solutions to be such that transitions when φ_2 becomes nearly stationary occur for (φ_1, φ_2) near points equivalent to $(x_1, \sin^{-1}(I - \sin x_1) - 2L_1\pi)$ (mod 2π in each component), and that the other transitions occur near points equivalent to $(x_2, \pi - \sin^{-1}(I - \sin x_2) - 2L_2\pi)$. Furthermore, when φ_1 is nearly constant we would expect it to slowly move from near a value equivalent to x_2 (mod 2π) to near one equivalent to x_1. Similarly, φ_2 would be expected to slowly vary from near a value equivalent to $\sin^{-1}(I - \sin x_1)$ to near one equivalent to $\pi - \sin^{-1}(I - \sin x_2)$ when it is nearly constant. Appropriate values of L_1 and L_2 can be found by observing that $2L_1\pi$ would approximate the minimum and $2L_2\pi$ the maximum value of $\varphi_1 - \varphi_2$ in the beating solution.

For $\gamma = 0.5$, $H = 0$, and $k = 0.001$, Imry and Schulman [1] report finding beating solutions only for $|I - 1.6| < 0.05$, and that the extreme values of $\varphi_1 - \varphi_2$ for these are close to 188π and 320π. Since L_1 corresponds to a curve in family (i), and L_2 to a curve in family (ii), a first test of the conjecture is to determine if the bifurcation curve with $L = 94$ in the family from F_1 intersects the one with $L = 160$ in the family from F_2 at a point of S near $(I,k) = (1.6, 0.001)$.

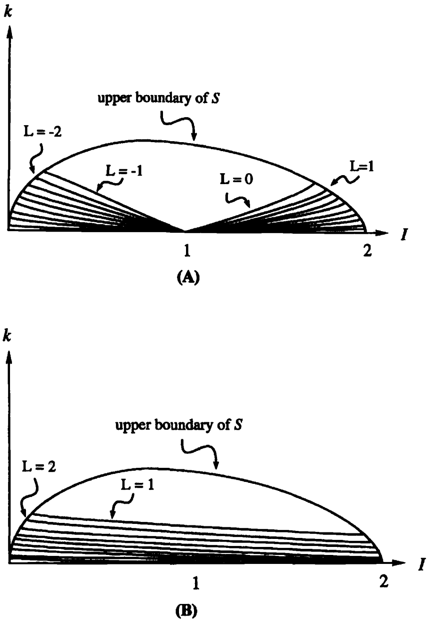


FIG. 2. (a) *A sketch of the bifurcation curves in family* (i) *for* $H = \frac{1}{2}$. *Bifurcation curves in the subfamily having positive slope correspond successively to* $L = 0, 1, 2, \ldots$, *while those in the subfamily having negative slope correspond successively to* $L = -1, -2, -3, \ldots$. *In each case, the curves accumulate on the* I*-axis as* $|L| \to \infty$. (b) *A sketch of the bifurcation curves in family* (ii) *for* $H = \frac{1}{2}$. *These curves have negative slope, correspond successively to* $L = 1, 2, 3, \ldots$, *and accumulate on the* I*-axis as* $L \to \infty$.

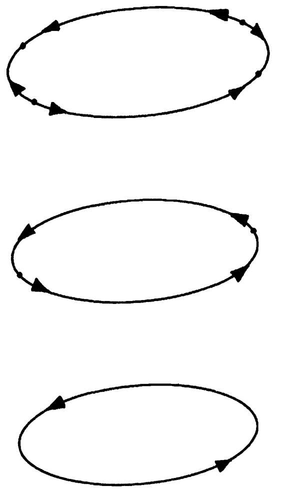


FIG. 3. *A highly schematic illustration of the idea of the conjecture. The fixed points disappear in a pair of simultaneous sink-saddle bifurcations, leaving a beating solution in their wake.*

From Theorem 4.2, we see that these two curves intersect in at most one point, since they have slopes of opposite sign. Using $L = 160$ in the estimate for F_2 in Lemma 4.1 shows that $0.000994 < k < 0.000998$ along the corresponding bifurcation curve. Similarly, the estimate for F_1 with $L = 94$ shows that $k < 0.000994$ at $I = 1.587$, and $k > 0.000998$ at $I = 1.593$ on the $L_1 = 94$ bifurcation curve. Consequently, these curves must intersect at a point very close to $(I,k) = (1.59, 0.000996)$, which is in good agreement with the conjectured intersection near $(I,k) = (1.6, 0.001)$.

To further test the conjecture, it should be mentioned that numerical approximation of the beating solutions shows that during the time when they are nearly stationary φ_1 and φ_2 slowly vary from being near values equivalent to $\pi/5 \pmod{2\pi}$ to being near values equivalent to $\pi/2$. On the other hand, the estimates in Lemma 3.1 show that when k is small (e.g., near 0.001), x_1 and x_2 are very close to $\pi/2$ and $\sin^{-1}(I-1)$, respectively. Since I is near 1.6, we see that x_2 is near $\pi/5$. It now follows that $\sin^{-1}(I - \sin x_1)$ is near $\pi/5$, and that $\pi - \sin^{-1}(I - \sin x_2)$ is near $\pi/2$. Thus, the conjectured behavior of φ_1 and φ_2 when they are nearly constant is also in good agreement with the numerical results.

In light of the conjecture, the similarity of the bifurcation curves in families (iii) and (iv) to those in families (i) and (ii) suggests the possibility that there might be another class of beating solutions. In this regard, we mention that the existence of *unstable* beating solutions has been proved by Levi [2, 4]. Such solutions could perhaps arise by a mechanism similar to that of the conjecture but involving simultaneous bifurcations corresponding to an intersection of curves from families (iii) and (iv). In that case, the role of the sinks in the conjecture would be played by saddles having a one-dimensional unstable manifold, and the role of the saddles would be played by saddles having a two-dimensional unstable manifold. The work done by Levi in [2] and [4] does not appear to have an impact on these conjectures, since it excludes the region of parameter space where bifurcations of the beating solutions occur.

Acknowledgments. It is with pleasure that I thank Nancy Kopell and Mark Levi for many interesting discussions of Josephson junctions.

REFERENCES

[1] Y. Imry and L. Schulman, *Qualitative theory of the nonlinear behavior of coupled Josephson junctions*, J. Appl. Phys., 49 (1978), pp. 749–758.

[2] M. Levi, *Beating Modes in the Josephson Junction*, in Chaos in Nonlinear Dynamical Systems, J. Chandra, ed., Society for Industrial and Applied Mathmatics, Philadelphia, PA, 1984, pp. 56–73.

[3] M. Levi, F. C. Hoppensteadt, and W. L. Miranker, *Dynamics of the Josephson junction*, Quart. Appl. Math., 36 (1978), pp. 167–198.

[4] M. Levi, *Caterpillar solutions in coupled pendula*, Ergodic Theory Dynamical Systems, 8* (1988), pp. 153–174.

[5] N. Kopell, private communication.

[6] M. E. Henderson, M. Levi, and F. Odeh, *The geometry and computation of the dynamics of coupled pendula*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 1 (1991), pp. 27–50.

[7] E. J. Doedel, D. G. Aronson, and H. G. Othmer, *The dynamics of coupled current-biased Josephson junctions:* I, IEEE Trans. Circuits Systems, 35 (1988), pp. 810–817.

[8] D. G. Aronson, E. J. Doedel, and H. G. Othmer, *The dynamics of coupled current-biased Josephson junctions* II, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 1 (1991), pp. 55–66.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 1095–1101, July 1996

014

MINIMAL PERIODS FOR SOLUTIONS OF SOME CLASSICAL FIELD EQUATIONS*

D. STUART[†]

Abstract. Time-periodic solutions of a class of coupled semilinear wave equations in any space dimension are considered. It is shown that there is a lower bound on the period of solutions which have finite energy and decay at a certain rate at spatial infinity. This generalises results of Coron and Levine which obtained the same bound for, respectively, scalar equations in one space dimension and solutions with spherical symmetry on exterior domains in higher space dimensions. The methods used are adapted from work of Kato on the absence of positive eigenvalues for Schrödinger operators.

Key words. breathers, minimal period, nonlinear wave equations, period solutions

AMS subject classification. 35B

1. Introduction.

In this note we shall extend results of Coron [2] and Levine [7] on the minimal period of (time-) periodic solutions of nonlinear wave equations to certain systems of semilinear wave equations in several space dimensions. The proof is an adaptation of Kato's method in Schrödinger operators [6]. We shall start by recalling some known time-periodic solutions, and then we will explain the lower bound on the period of such solutions. In one dimension, the sine-Gordon equation

$$u_{tt} - u_{xx} + \sin u = 0,$$

where u is real valued, has periodic solutions called breathers with period larger than 2π (see, for example, [8]). For single equations in one space dimension, this example is expected to be essentially unique; see [1, 3]. Spherically symmetric solutions on $\mathbf{R}^3$ which are not bounded at the origin are discussed in [9]. However, for systems, there are plenty of periodic solutions given by the following method. Consider the *complex* field equation for $\phi : \mathbf{R}^{1+n} \to \mathbf{C}$ of the form

$$\phi_{tt} - \Delta\phi + a^2\phi = f(|\phi|)\phi,$$

where f is real and $f(0) = 0, f'(0) = 0$. This has periodic solutions for a large class of functions f (see, for example, [4]). These solutions are of the form $\phi(t,x) = e^{i\omega t}u(x)$, where u satisfies an elliptic equation and $\omega^2 \le a^2$. Therefore, the period of these solutions is always bigger than or equal to $\frac{2\pi}{a}$.

In [2], the equation

$$u_{tt} - u_{xx} + g(u) = 0$$

with $g \in C^2(\mathbf{R})$, $g(0) = 0$, was considered. It was proved that if $u(t,x)$ is a twice differentiable real-valued function such that $u(t+T,x) = u(t,x)$ and

$$\int_{-\infty}^{+\infty}\int_0^T |u(t,x)|dtdx < \infty,$$

$$\lim_{|x|\to+\infty}\int_0^T (u_t^2 + u_x^2)dt = 0,$$

* Received by the editors September 12, 1994; accepted for publication (in revised form) December 6, 1994. This research was supported by National Science Foundation grant DMS-9214067.

[†] Department of Mathematics, University of California at Davis, Davis, CA 95616 (dmstuart@aztec.ucdavis.edu).

$$\lim_{|x|\to+\infty}\max_{0\le t\le T}|u(t,x)|=0,$$

then either u is independent of time or $(\frac{2\pi}{T})^2 \le g'(0)$. This gives a lower bound for the period. In [7], this was generalised to real scalar equations in several space dimensions under the assumption of spherical symmetry. (In this paper, it was only assumed that the solution existed in an exterior domain; thus regularity at the origin was not required). In the present paper, we will generalise this lower bound on the period to classes of systems in several space dimensions *without assuming the solution to be spherically symmetric.* We will do this by by obtaining lower bounds at infinity on a solution which violates this condition. We then show that these bounds imply it has infinite energy. To do this, it is necessary to make some assumptions about the decay of the solution at spatial infinity which are somewhat stronger than those required for Coron's result. We will state some consequences of the main theorem before stating the result in generality. In what follows, a T-periodic function u is a function of $(t,x)\in\mathbf{R}^{1+n}$ such that $u(t+T,x)=u(t,x)$.

THEOREM 1. *Consider a T-periodic solution u of the following system of coupled wave equations in* $\mathbf{R}^{1+3}$,

$$u^\alpha_{tt}-\Delta u^\alpha+u^\alpha+f^\alpha(u)=0,\qquad \alpha=1,\dots,m,$$

which is not independent of time, which satisfies the condition in equation (3) *below, and which is such that* $u=O(\frac{1}{|x|})$ *as* $|x|\to\infty$. *Then if for all* α, $f^\alpha(u)$ *is a homogeneous polynomial of third degree, then the period* $T\ge 2\pi$.

THEOREM 2. *Consider a time-periodic solution u of the system of nonlinear wave equations in* $\mathbf{R}^{1+n}n\ge 3$

$$u^\alpha_{tt}-\Delta u^\alpha+f^\alpha(u)=0,\qquad \alpha=1,\dots,m,$$

which satisfies the condition in equation (3) *below and which is such that* $u=O(|x|^{-\frac{(n-1)}{2}})$ *as* $|x|\to\infty$. *Then if* $n\ge 4$ *and if for every* α, $f^\alpha(u)$ *is a homogeneous polynomial of second degree, then u is independent of time. If* $n=3$ *and if for every* α, $f^\alpha(u)$ *is a homogeneous polynomial of third degree, then u is independent of time. In the case* $n=3$, *if* $u=o(\frac{1}{|x|})$ *as* $|x|\to\infty$ *and if for every* α, $f^\alpha(u)$ *is a homogeneous polynomial of second degree, then u is independent of time.*

These theorems are consequences of the main theorem, which we give in the next section. Another type of restriction on the possible existence of time-periodic solutions can be obtained by arguments based on Pohazaev-type identities [5].

2. The main theorem.

Consider the system of equations

$$u^\alpha_{tt}-\Delta u^\alpha+k_\alpha^2u^\alpha+f^\alpha(u)=0 \tag{1}$$

for a vector-valued function $u=\{u^\alpha\}_{\alpha=1}^m$ of $(t,x)\in\mathbf{R}^{1+n}$. We do *not* use the summation convention. We will assume that $f:\mathbf{R}^m\to\mathbf{R}^m$ is a twice continuously differentiable function with $f(0)=f'(0)=0$ and that u is a twice continuously differentiable solution which is T-periodic in time and *not independent of time.* We can then write

$$u(t,x)=\overline{u}(x)+\hat{u}(t,x),$$

where $\overline{u} = \frac{1}{T}\int_0^T u(t,x)dt$ and $\hat{u} \neq 0$ and satisfies

$$\int_0^T |\hat{u}_t|^2 dt \geq \left(\frac{2\pi}{T}\right)^2 \int_0^T |\hat{u}|^2 dt.$$

It turns out to be useful to think of the equation as an ordinary differential equation in r with values in the Hilbert space H which consists of T-periodic functions on the unit sphere with norm

$$\|u(r)\|^2 \equiv \int_0^T \int_{S^{n-1}} \sum_\alpha |u^\alpha|^2 dSdt, \tag{2}$$

where dS is the usual measure on the unit sphere. The corresponding inner product will be written $(\cdot,\cdot)$. Since f is C^2, it follows from the fundamental theorem of calculus that there exist continuous functions $h^\alpha_{\beta\gamma}(\overline{u},\hat{u})$ such that

$$f^\alpha(u) = f^\alpha(\overline{u}) + f^\alpha(\hat{u}) + \sum_{\beta,\gamma} \overline{u}^\beta \hat{u}^\gamma h^\alpha_{\beta\gamma}(\overline{u},\hat{u}).$$

Introduce the following function:

$$q[u](R) \equiv \max_{|x|=R} \max_{0\leq t\leq T} \left\{ \frac{|f(\hat{u}(t,x))|}{\|\hat{u}(R)\|} + \left| \sum_\beta \overline{u}_\beta(t,x) h^\alpha_{\beta\gamma}(\overline{u}(t,x),\hat{u}(t,x)) \right| \right\}.$$

Notation. Here the norm on the linear transformation $\sum_\beta \overline{u}_\beta h^\alpha_{\beta\gamma}(\overline{u},\hat{u}) : \mathbf{R}^m \to \mathbf{R}^m$ is that induced from the Euclidean norm on $\mathbf{R}^m$. We will often omit the argument $[u]$ from q. Define $k = \max_\alpha |k_\alpha|$.

We will prove the following theorem.

MAIN THEOREM. *Assume that $u \in C^2(\mathbf{R}^{1+n};\mathbf{R}^m)$ is a T-periodic solution of equation* (1) *which is not independent of time and such that the following conditions are satisfied:*

$$\sum_{\alpha=1}^m \int_0^T \int_{\mathbf{R}^n} (|u^\alpha_t|^2 + |\nabla u^\alpha|^2) dxdt < \infty, \tag{3}$$

$$\lim_{R\to\infty} Rq[u](R) = 0. \tag{4}$$

Then $T \geq \frac{2\pi}{k}$.

Remark. To prove this theorem, we will assume we have a solution for which

$$\left(\frac{2\pi}{T}\right)^2 > k^2 \tag{5}$$

and use a sucession of lemmas which place lower bounds on the behaviour of certain spherical averages as $|x| = r \to \infty$ which contradict the first assumption.

The technical difficulties which arise from the lack of spherical symmetry are due to the fact that the operator $-\Delta$ contains a component $-r^{-2}\Delta_S$, where Δ_S is the spherical Laplacian. This term has the opposite sign to the term $\partial_t^2 + k_\alpha^2$. It seems not to be possible to eliminate this difficulty by averaging over the sphere due to the presence of nonlinear terms.

Let K be a diagonal $m \times m$ matrix such that $(Ku)^\alpha = k_\alpha u^\alpha$ and $(K^2 u)^\alpha = k_\alpha^2 u^\alpha$. It will be clear from the proof that we could generalise this to consider a system of equations in which the matrix K is positive definite but not diagonal. We will work with $\hat{u}$, which satisfies

$$\hat{u}_{tt}^\alpha - \Delta \hat{u}^\alpha + k_\alpha^2 \hat{u}^\alpha + f^\alpha(\overline{u}) + f^\alpha(\hat{u}) + \sum_{\beta,\gamma} \overline{u}^\beta \hat{u}^\gamma h_{\beta\gamma}^\alpha(\overline{u}, \hat{u}) - \Delta \overline{u}^\alpha + k_\alpha^2 \overline{u}^\alpha = 0.$$

It is convenient to rescale as follows: define

$$w \equiv r^{\frac{n-1}{2}} \hat{u} \quad \text{and} \quad w_m \equiv r^m w. \tag{6}$$

We then calculate that these satisfy the following equations:

$$\begin{aligned}
&-(w^\alpha)'' + \left(k_\alpha^2 + \frac{\mu}{r^2}\right) w^\alpha + \partial_t^2 w^\alpha - \frac{1}{r^2} \Delta_s w^\alpha \\
&\qquad + r^{\frac{n-1}{2}} \left(f^\alpha(\overline{u}) + f^\alpha(\hat{u}) + \sum_{\beta,\gamma} \overline{u}^\beta \hat{u}^\gamma h_{\beta\gamma}^\alpha(\overline{u}, \hat{u}) - \Delta \overline{u}^\alpha + k_\alpha^2 \overline{u}^\alpha \right) = 0, \\
&-(w_m^\alpha)'' + \frac{2m}{r} (w_m^\alpha)' + \left(k_\alpha^2 + \frac{\mu - m(m+1)}{r^2}\right) w_m^\alpha + \partial_t^2 w_m^\alpha - \frac{1}{r^2} \Delta_s w_m^\alpha \\
&\qquad + r^{m + \frac{n-1}{2}} \left(f^\alpha(\overline{u}) + f^\alpha(\hat{u}) + \sum_{\beta,\gamma} \overline{u}^\beta \hat{u}^\gamma h_{\beta\gamma}^\alpha(\overline{u}, \hat{u}) - \Delta \overline{u}^\alpha + k_\alpha^2 \overline{u}^\alpha \right) = 0,
\end{aligned}$$

where $\mu = \frac{1}{4}(n-1)(n-3)$, Δ_s is the spherical Laplacian, and $'$ means $\frac{d}{dr}$. To understand the solutions, we introduce the following quantity:

$$Q(r) \equiv \|w'\|^2 + \|\dot{w}\|^2 - \|Kw\|^2 - \frac{\mu}{r^2} \|w\|^2 + \frac{1}{r^2} (w, \Delta_s w).$$

This is useful because of the following result.

LEMMA. *There exists a number R_1 such that for $r \geq R_1$,*

$$\frac{\partial}{\partial r} \Big(rQ(r) \Big) \geq 0.$$

Proof. First of all, using $\int_0^T w dt = 0$, we calculate that

$$\begin{aligned}
(rQ)' &\geq \|w'\|^2 + \left(\left(\frac{2\pi}{T} \right)^2 - k^2 + \frac{\mu}{r^2} \right) \|w\|^2 - \frac{1}{r^2} (w, \Delta_s w) \\
&\qquad + 2 \left(r(w^\alpha)', r^{\frac{n-1}{2}} f^\alpha(\hat{u}) + \sum_{\beta,\gamma} \overline{u}^\beta w^\gamma h_{\beta\gamma}^\alpha(\overline{u}, \hat{u}) \right) \\
&\geq \|w'\|^2 + \left(\left(\frac{2\pi}{T} \right)^2 - k^2 + \frac{\mu}{r^2} \right) \|w\|^2 - 2|rq(r)| \|w'\| \|w\|,
\end{aligned}$$

where $k = \max_\alpha |k_\alpha|$. It is clear that this is a nonnegative quadratic form for large r, on account of equations (4) and (5).

The next stage is to show that there are arbitrarily large values of r for which

$Q > 0$. To do this, we introduce, following Kato, a subsidiary function:

$$\begin{aligned} &Q(r; m, s) \\ &\equiv \|w'_m\|^2 + \|\dot{w}_m\|^2 - \|Kw_m\|^2 - \left(\frac{\mu}{r^2} + \frac{2s}{r} - \frac{m(m+1)}{r^2}\right)\|w_m\|^2 + \frac{1}{r^2}(w_m, \Delta_s w_m). \end{aligned}$$

LEMMA. *For every $R_2 > 0$, there exists a number $s_1 > 0$ such that for $s < s_1$, there exists a number M_1 such that for $m \geq M_1$ and $r \geq R_2$, we have*

$$(r^2 Q(r; m, s))' \geq 0.$$

Proof. Using $\int_0^T w_m dt = 0$, we calculate that

$$\begin{aligned} (r^2 Q(r; m, s))' = 2r \Bigg[&(2m+1)\|w'_m\|^2 + \|\dot{w}_m\|^2 - \frac{s}{r}\|w_m\|^2 - \|Kw_m\|^2 \\ &+ \left(r(w^\alpha_m)', r^{m+\frac{n-1}{2}} f^\alpha(\hat{u}) + \sum_{\beta,\gamma} \overline{u}^\beta w^\gamma_m h^\alpha_{\beta\gamma}(\overline{u}, \hat{u}) - \frac{2s}{r} w^\alpha_m \right) \Bigg] \\ \geq 2r \Bigg[&(2m+1)\|w'_m\|^2 + \left(\left(\frac{2\pi}{T}\right)^2 - k^2 - \frac{s}{r} \right) \|w_m\|^2 \\ &-2|(rq(r) + 2s)| \|w'_m\| \|w_m\| \Bigg]. \end{aligned}$$

The condition for this form to be definite is therefore satisfied for sufficiently large m as long as we choose s such that

$$\left(\left(\frac{2\pi}{T}\right)^2 - k^2 - \frac{s}{r} \right) \geq 0.$$

So we let s_1 be given by

$$s_1 = R_2 \left(\left(\frac{2\pi}{T}\right)^2 - k^2 \right).$$

Next, for $s < s_1$, the condition for positivity of the quadratic form is

$$(rq + 2s)^2 \leq (2m+1)\left(\left(\frac{2\pi}{T}\right)^2 - k^2 - \frac{s}{r} \right),$$

and since $rq(r)$ is bounded, this is satisfied for $r \geq R_2, s < s_1$ if m larger than some number $M_1(s)$.

Before going on, we will record the following formula for $Q(r; m, s)$:

$$\begin{aligned} Q(r; m, s) = r^{2m} \Bigg(&\|w'\|^2 + \frac{2m}{r}(w, w') + \|\dot{w}\|^2 - \|Kw\|^2 \\ &- \left(\frac{\mu}{r^2} + \frac{2s}{r} - \frac{m(2m+1)}{r^2} \right) \|w\|^2 + \frac{1}{r^2}(w, \Delta_s w) \Bigg) \end{aligned}$$

$$= r^{2m} \left(Q(r) + \frac{2m}{r}(w, w') - \left(\frac{2s}{r} - \frac{m(2m+1)}{r^2} \right) \|w\|^2 \right). \tag{7}$$

COROLLARY. *There exist arbitrarily large values of r for which $Q(r) > 0$.*

Proof. Since we are assuming that $\int u_t^2 dx = \int_0^\infty \|w\|^2 dr < \infty$, we may assume that there is an unbounded subset S of the positive real axis such that for $r \in S$, $(w, w')(r) \leq 0$. At such a point, we know from equation (7) that

$$Q(r; m, s) \leq r^{2m} \left(Q(r) - \left(\frac{2s}{r} - \frac{m(2m+1)}{r^2} \right) \|w\|^2 \right) \quad \text{for } r \in S.$$

Now let R_3 be a value of r for which $\|w\| \neq 0$. Such a number exists because we assumed that u is not independent of time. Then, first of all, we can find a number M_2 such that for $m \geq M_2$, $Q(R_3; m, s) > 0$ because $Q(R_3; m, s)$ is quadratic in m from equation (7). Then we apply the previous lemma to deduce that for sufficiently small s and large m, $Q(r; m, s) > 0$ for all $r \geq R_3$. Choosing $r \in S$ sufficently large that $\frac{2s}{r} - \frac{m(2m+1)}{r^2} > 0$, it then follows from the preceeding formula that $Q(r) > 0$ as required on an unbounded set.

Proof of Main Theorem. Finally, to prove the theorem, we choose a number $R_4 > R_1$ for which $Q(R_4) > 0$. Then by the first lemma, we know that for $r > R_4$,

$$Q(r) \geq \frac{R_4 Q(R_4)}{r},$$

from which it follows that $Q(r)$ is not integrable. Now to complete the proof, notice that

$$\begin{aligned}
r^{n-1}(\|\nabla u\|^2 + \|u_t\|^2) &\geq r^{n-1}(\|\nabla \hat{u}\|^2 + \|\hat{u}_t\|^2) \\
&\geq r^{n-1}\left(\|\hat{u}'\|^2 + \frac{2\pi^2}{T^2}\|\hat{u}\|^2 \right) + \frac{1}{2}\|w_t\|^2 \quad \text{from equation (2)} \\
&\geq \|w'\|^2 - \frac{(n-1)}{r}(w, w') + \frac{(n-1)^2}{4r^2}\|w\|^2 \\
&\qquad + \frac{2\pi^2}{T^2}\|w\|^2 + \frac{1}{2}\|w_t\|^2 \quad \text{from equation (6)} \\
&\geq \frac{1}{2}(\|w'\|^2 + \|w_t\|^2) + \left(\frac{2\pi^2}{T^2} - \frac{(n-1)^2}{4r^2} \right)\|w\|^2 \\
&\geq \frac{1}{2}Q(r) \quad \text{for sufficently large r} \geq R_5.
\end{aligned}$$

It follows that $r^{n-1}(\|\nabla u\|^2 + \|u_t\|^2)$ is not integrable on $[R_5, \infty)$ and hence that the finite-energy assumption

$$\int_0^T \int_{\mathbf{R}^n} (u_t^2 + |\nabla u|^2) dx dt = \int_0^\infty r^{n-1}(\|\nabla u\|^2 + \|u_t\|^2) dr < \infty$$

is violated. This completes the proof.

Acknowledgments. I would like to thank Richard Montgomery for discussions related to this work and the referee for helpful comments.

REFERENCES

[1] B. BIRNIR, H. MCKEAN, AND A. WEINSTEIN, *Nonexistence of breathers*, Comm. Pure Appl. Math., 47 (1994), pp. 1043–1053.

[2] J-M. CORON, *Periode minimale pour une corde vibrante de longuer infinie*, Comptes Rendues Acad. Sci. Paris, A294 (1982), pp. 127–129.

[3] J. DENZLER, *Nonpersistence of breather families for the perturbed sine-Gordon equation*, Comm. Math. Phys., 158 (1993), pp. 397–430.

[4] H. BERESTYCKI AND P-L. LIONS, *Nonlinear scalar field equations, parts one and two*, Arch. Rational Mech. Anal., 82 (1983), pp. 313–375.

[5] L. V. KAPITANSKII, *Absence of time-periodic solutions for certain multidimensional nonlinear wave equations*, J. Soviet Math., 40 (1988), pp. 622–629.

[6] T. KATO, *Growth properties of solutions to the reduced wave equation with variable coefficients*, Comm. Pure Appl. Math., 12 (1959), pp. 403–425.

[7] H. LEVINE, *Minimal periods for solutions of semilinear wave equations in exterior domains and for solutions of the equations of nonlinear elasticity*, J. Math. Anal. Appl., 135 (1988), pp. 297–308.

[8] S. NOVIKOV, S. V. MANAKOV, L. P. PITAEVSKII, AND V. E. ZAKHAROV, *Theory of Solitons*, Consultants Bureau, New York, 1984.

[9] M. SMILEY, *Breathers and forced oscillations of nonlinear wave equations on* $\mathbf{R}^3$, J. Reine Angew. Math., 398 (1989), pp. 25–35.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 1102–1128, July 1996

ASYMPTOTIC AND NUMERICAL APPROXIMATIONS OF THE ZEROS OF FOURIER INTEGRALS*

DAVID SENOUF[†]

To the memory of Midge Bennahum.

Abstract. The asymptotic behavior as $y \to +\infty$ of Fourier integrals of the form

$$f_n(y) = \int_{-\infty}^{\infty} e^{-t^{2n}+iyt}dt, \quad n \in \mathbb{N}, \quad n \geq 2,$$

is derived via the method of steepest descents. A general formula is found for the coefficients of the expansion of $f_n(y)$ in the sector $|\arg y| < \frac{2n-1}{2n}\frac{\pi}{2}$ centered about the anti-Stokes line $y \in \mathbb{R}$. A high-order asymptotic approximation of the real zeros of $f_n(y)$ is also obtained. A simple numerical method designed to compute the zeros of $f_n(y)$ is described. For $n = 2$ and $n = 3$, the asymptotic estimates of the zeros are compared to numerically computed values.

Key words. asymptotic expansions, steepest descents, Fourier integrals, Pearcey integral, zeros

AMS subject classifications. 30E15, 33B10, 41A60

1. Introduction.

In [20], Pólya showed that functions of the form

$$\int_{-\infty}^{\infty} e^{-at^{4n}+bt^{2n}+ct^2+iyt}dt, \qquad n \in \mathbb{N}, \quad n \geq 1, \quad a > 0, \quad b \in \mathbb{R}, \quad c \geq 0, \tag{1}$$

have only real zeros. Similar results are found in [19] concerning functions of the form

$$f_n(y) = \int_{-\infty}^{\infty} e^{-t^{2n}+iyt}dt, \qquad n \in \mathbb{N}, \quad n \geq 2. \tag{2}$$

In [7], de Bruijn generalized Pólya's results to a larger class of functions whose zeros are real. Recently, Paris analyzed in [16] a generalized form of the Pearcey integral of which the "Pólya" functions given by (1) are particular cases. He studied the asymptotic behavior of

$$P_n'(X,Y) = \int_{-\infty}^{\infty} e^{i(u^{2n}+Xu^n+Yu)}du, \qquad n \in \mathbb{N}, \quad n \geq 2, \tag{3}$$

as $|X| \to \infty$ or $|Y| \to \infty$. By rotation of the path of integration ($u = te^{\frac{\pi i}{4n}}$) and use of Jordan's lemma, it can be expressed as

$$P_n'(X,Y) = P_n(x,y) = e^{\frac{\pi i}{4n}} \int_{-\infty}^{\infty} e^{-t^{2n}-xt^n+iyt}dt \tag{4}$$

with $x = Xe^{-\frac{\pi i}{4}}$ and $y = Ye^{\frac{\pi i}{4n}}$. For n even, $x > 0$, $P_n(x,y)$ falls into the category of functions (1) considered by Pólya. For $x = 0$, f_n and P_n are related by $f_n(y) = e^{-\frac{\pi i}{4n}}P_n(0,y)$. The Pearcey integral $P_2(x,y)$ has been studied by Kaminski [14], Paris [17], and references therein. The advantage of the method described by Paris is that it

[*] Received by the editors August 1, 1994; accepted for publication (in revised form) January 9, 1995. This research was partially supported by NSF grant DMS-9306720.

[†] Department of Mathematics, University of California at Los Angeles, Los Angeles, CA 90095-1555. Current address: 4 rue Jean Ferrandi, Paris 75006, France (senouf@calvanet.calvacom.fr).

avoids the complicated, sometimes impossible task of finding a closed-form expression for the saddle points.

Although a large portion of this work is devoted to the derivation of the asymptotic expansion of $f_n(y)$ as $y \to +\infty$, our main objective is the derivation of asymptotic approximations for the zeros of $f_n(y)$. The order of this real analytic even function, which is defined as the positive number λ_n for which $\max_{|y|=r} |f_n(y)| \leq \exp(r^{\lambda_n+\epsilon})$ $\forall \epsilon > 0$ as soon as r is sufficiently large, is the rational number $1 < \lambda_n = \frac{2n}{2n-1} < 2$. The order being fractional, it is known that $f_n(y)$ has infinitely many zeros [2, 4], which, from the results of Pólya, must be real; thus we are mainly interested in the behavior of $P_n(0, y)$ for large real values of y. The expansion we find is an expression on the anti-Stokes line $\arg y = 0$ where two saddle points have equal contributions. Hence we first consider the real-valued function $f_n(y) : \mathbb{R} \to \mathbb{R}$, which, by the change of variable $t \to z\left(\frac{y}{2n}\right)^{1/(2n-1)}$ for $|\arg y| < \frac{2n-1}{2n}\pi/2$, can be expressed in terms of another function $\mathcal{F}_n(\mu)$ defined as follows:

$$f_n(y) = \left(\frac{y}{2n}\right)^{\frac{1}{2n-1}} \mathcal{F}_n\left(\left(\frac{y}{2n}\right)^{\frac{2n}{2n-1}}\right), \tag{5}$$

where for $|\arg \mu| < \pi/2$,

$$\mathcal{F}_n(\mu) = \int_{-\infty}^{\infty} e^{\mu\,(2niz - z^{2n})} dz = 2\int_0^{+\infty} \cos(2n\mu z) e^{-\mu z^{2n}} dz. \tag{6}$$

The advantage of introducing the function $\mathcal{F}_n(\mu)$ is that its saddle points are fixed to the unit disk, contrary to those of $f_n(y)$, which depend on the large variable y.

The expansion of functions of the type of $f_n(y)$ can be found as early as 1916 in the work of Brillouin in [5] and then in 1924 in the work of Burwell [8], who obtained first-order asymptotics for the location of the zeros of such functions (see also [3]). Recently, Christ also characterized the zeros of similar functions in [9, Lem. 2.1]. In [18, Chap. 3], Paris and Wood investigate the asymptotic properties of high-order differential equations whose solutions have integral representations closely related to (2). In their work, they derive recurrence relations to determine the coefficients of the asymptotic expansions (see, for example, [18, Eq. 3.4.16]). We provide a different approach than Paris and Wood's using the classical method of steepest descents, and we derive the full (generalized) asymptotic expansion of $\mathcal{F}_n(\mu)$ as $\mu \to +\infty$ valid in the sector $|\arg \mu| < \pi/2$ together with high-order asymptotic approximations of its zeros. We provide a systematic way of calculating every coefficient of the expansion of $\mathcal{F}_n(\mu)$ via series reversion in terms of multinomial coefficients. This formulation can be compared to the results of Paris and Wood in [18] in which the coefficients are described by a $2n$-term recursion relation which is derived from ordinary differential equation (ODE) methods.

In the last section, we describe a simple numerical algorithm which computes the zeros of the function $\mathcal{F}_n(\mu)$. This algorithm is efficient for small values of the zeros, and for $n = 2$ and $n = 3$, it is implemented to gauge the accuracy of the asymptotic estimates. An application of the high-order asymptotic approximations of the zeros of $\mathcal{F}_2(\mu)$ derived in this article can be found in [22]–[24].

We introduce the following definitions and notations.

Definition 1.1. Compound asymptotic expansion (c.a.e.) *of $f(z)$ with respect* *to the asymptotic sequences* $\{\phi_n^1(z)\}$ *and* $\{\phi_n^2(z)\}$: *we write*

$$f(z) \overset{z\to z_0}{\sim} g_1(z)\left[\sum_{n=0}^{\infty} f_n^1(z);\ \{\phi_n^1(z)\}\right] + g_2(z)\left[\sum_{n=0}^{\infty} f_n^2(z);\ \{\phi_n^2(z)\}\right],$$

where it is understood that

$$f(z) \sim g_1(z)\left[\sum_{n=0}^{\infty} f_n^1(z) + o(\phi_n^1(z))\right] + g_2(z)\left[\sum_{n=0}^{\infty} f_n^2(z) + o(\phi_n^2(z))\right] \quad \text{as } z \to z_0.$$

As an alternative to a c.a.e., it may be possible to express the asymptotic expansion of $f(z)$ as a generalized asymptotic expansion.

DEFINITION 1.2. Generalized asymptotic expansion (g.a.e.) *of* $f(z)$ *with respect to the asymptotic sequence* $\{\phi_n(z)\}$: *let* $\{\phi_n(z)\}$ *be an asymptotic sequence as* $z \to z_0 \in R$, *where* R *is a region in the complex plane and* $f(z)$ *and* $f_n(z)$, $n = 0, 1, \ldots$, *are functions such that for each positive integer* m,

$$f(z) = \sum_{n=0}^{m-1} f_n(z) + \mathcal{O}\left(\phi_m(z)\right) \quad (z \to z_0 \in R).$$

Then we say that $\sum_n f_n(z)$ *is a g.a.e. with respect to the asymptotic sequence* $\{\phi_n(z)\}$ *and write*

$$f(z) \sim \sum_{n=0}^{\infty} f_n(z); \quad \{\phi_n(z)\} \text{ as } z \to z_0 \in R.$$

For convenience, we also write it as

$$f(z) \underset{\{\phi_n(z)\}}{\overset{z\to z_0}{\sim}} \sum_{n=0}^{\infty} f_n(z).$$

We prove the following.

THEOREM 1.1. *Let* $n \in \mathbb{N}, n \geq 2$, *and for* $|\arg \mu| < \pi/2$, *let*

$$\mathcal{F}_n(\mu) = \int_{-\infty}^{\infty} e^{\mu\,(2niz - z^{2n})} dz.$$

The g.a.e. of $\mathcal{F}_n(\mu)$ *as* $\mu \to +\infty$ *with respect to the asymptotic sequence* $\{\phi_j(\mu) = \mu^{-j}\}$, *valid in the sector* $|\arg \mu| < \pi/2$, *is*

$$\mathcal{F}_n(\mu) \underset{\{\mu^{-j}\}}{\overset{\mu\to+\infty}{\sim}} \sqrt{\frac{4\pi}{n(2n-1)\mu}} \exp\left\{-\mu\,(2n-1)\sin\left(\frac{\pi}{4n-2}\right)\right\} \mathcal{H}_n(\mu),$$

where

$$\mathcal{H}_n(\mu) = \sum_{j=0}^{\infty} \frac{\alpha_{n,j}}{\mu^j} \cos\left(\mu\,(2n-1)\cos\left(\frac{\pi}{4n-2}\right) + \frac{\pi}{4n-2}(1 - n(1+2j))\right)$$

and the coefficients $\alpha_{n,j}$ *are normalized rational numbers* $(\alpha_{n,0} = 1)$ *given by*

$$\alpha_{n,j} = \frac{\Gamma(j+1/2)}{\sqrt{\pi}\big(n(2n-1)\big)^j} \cdot \sum_{m=0}^{2j} \left\{ \frac{(1/2-j-m)_m}{\big(n(2n-1)\big)^m} \cdot {\sum_{\sigma}}' \prod_{k=1}^{2n-2} \frac{1}{\sigma_k!} \binom{2n}{k+2}^{\sigma_k} \right\}.$$

The summation ${\sum'}_{\sigma}$ *is to take place over all possible* $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{2n-2}) \in \mathbb{N}^{2n-2}$ *such that* $\sigma_1 + \sigma_2 + \cdots + \sigma_{2n-2} = m$ *and* $\sigma_1 + 2\sigma_2 + \cdots + (2n-2)\sigma_{2n-2} = 2j$. *Moreover, the first-order approximation of the kth ordered positive zero of* $\mathcal{F}_n(\mu)$ *is given by* (*for* $k \geq 1$)

$$\mu_{k,n}^{(0)} = \frac{\pi}{4n-2} \sec\left(\frac{\pi}{4n-2}\right) \left(\frac{n-1}{2n-1} - 1 + 2k\right) + \mathcal{O}\left(\frac{1}{k}\right) \quad \text{as } k \to +\infty.$$

Let

$$\mathcal{G}_n(\mu) = \mu + \frac{\sec(\frac{\pi}{4n-2})}{(2n-1)\,\mu} \left\{ \alpha_{n,1} \sin\left(\frac{n\pi}{2n-1}\right) - \frac{\alpha_{n,1}^2 - 2\alpha_{n,2}}{2\mu} \sin\left(\frac{2n\pi}{2n-1}\right) \right.$$
$$\left. + \frac{\alpha_{n,1}^3 - 3\alpha_{n,1}\alpha_{n,2} + 3\alpha_{n,3}}{3\mu^2} \sin\left(\frac{3n\pi}{2n-1}\right) - \frac{\sec(\frac{\pi}{4n-2})}{(2n-1)} \frac{\alpha_{n,1}^2}{\mu^2} \sin^2\left(\frac{n\pi}{2n-1}\right) \right\}.$$

Then the fourth-order approximation is given by

$$\mu_{k,n} = \mathcal{G}_n\left(\mu_{k,n}^{(0)}\right) + \mathcal{O}\left(\frac{1}{k^4}\right) \quad as\, k \to +\infty.$$

The corresponding kth ordered zero $y_{k,n}$ of $f_n(y)$ is given by

$$y_{k,n} = \pm 2n \left(\mu_{k,n}\right)^{\frac{2n-1}{2n}}. \tag{7}$$

Similarly, the corresponding expansion for the function $f_n(y)$ is obtained from that of $\mathcal{F}_n(\mu)$ by relation (5). Note finally that the behavior of $f_n(y)$ is exponentially small for large y in the sector $|\arg y| < \pi/(4n)$, and the behavior of $\mathcal{F}_n(\mu)$ is exponentially small for large μ in the sector $|\arg \mu| < \pi/(4n-2)$. In the respective complements of these sectors within $|\arg y| < \frac{2n-1}{2n}\frac{\pi}{2}$ (resp. $|\arg \mu| < \pi/2$), the behavior of $f_n(y)$ (resp. $\mathcal{F}_n(\mu)$) is exponentially large for large y (resp. large μ) (cf. [17, §5]).

2. Asymptotic expansion of $F(\mu) = \int_{-\infty}^{\infty} e^{\mu(4iz - z^4)} dz$ as $\mu \to +\infty$. We first describe the procedure for $n = 2$ corresponding to a special case of the Pearcey integral, which we generalize in the following section to arbitrary $n \in \mathbb{N}$. The result in this section has been derived by Paris and Wood in [18, pp. 64–72] using differential equation methods and will serve as comparison. The coefficients they derive, corresponding to the coefficients $\alpha_{2,j}$ in Corollary 2.1, are given in terms of a recurrence relation, whereas we offer a different approach and a different formulation in terms of elementary functions and combinatorial coefficients. This section serves as exposition for the general case $n \in \mathbb{N}$ and as such contains more details.

We are interested in deriving an asymptotic expansion of $F(\mu)$ as $\mu \to +\infty$, where

$$F(\mu) = \int_{-\infty}^{\infty} e^{\mu w(z)} dz, \quad w(z) = 4iz - z^4.$$

The method we use to do so is a standard method for asymptotic expansions of integrals depending on a parameter (cf. [6, 11, 25]). This method, known as Debye's method of steepest descents, is based on deforming the original path of integration through the local extrema of the integrand. The new path is chosen in such a way that along it, the integrand does not oscillate; i.e., the imaginary part of $w(z)$ remains constant. If there are several extrema z_s, only those for which $\Re w(z_s)$ is greatest are taken into consideration. Those that qualify are called the contributing saddle points. In our analysis, we expect two such extrema which must satisfy the condition $\Re w(z_0) = \Re w(z_1)$. These two equally relevant saddle points allow for the cancellation which generates the zeros of $F(\mu)$.

2.1. Saddle points, steepest paths, and contour deformation. We first locate the zeros of $w'(z)$, which we denote $z_s = \xi_s + i\,\eta_s$ and refer to as the saddle points of the integrand. For the quartic polynomial $w(z) = 4iz - z^4$, there are three saddle points:

$$0 = w'(z_s) = 4i - 4z_s^3 \Longrightarrow \{z_0, z_1, z_2\} = \left\{e^{\frac{\pi i}{6}}, e^{\frac{5\pi i}{6}}, e^{-\frac{\pi i}{2}}\right\}. \tag{8}$$

To determine which saddle points have a dominant contribution, we find $\Re w(z_s)$ for $s = 0, 1, 2$. Since $0 = \frac{z_s}{4}\, w'(z_s) = iz_s - z_s^4$, we find $w(z_s) = 4iz_s - z_s^4 = 3iz_s$ and therefore

$$\begin{aligned}\{w(z_0), w(z_1), w(z_2)\} &= \left\{3e^{\frac{2\pi i}{3}}, 3e^{\frac{4\pi i}{3}}, 3\right\}\\ &= \left\{-3/2 + i\,3\sqrt{3}/2, -3/2 - i\,3\sqrt{3}/2, 3\right\}.\end{aligned}$$

It would therefore seem that the dominant contribution comes from $z_2 = -i$. However, we will see that it is not possible to deform the original integration path through z_2. It is also apparent that z_0 and z_1 are equally valid candidates, for they have the same contribution:

$$\Re w(z_0) = \Re w(z_1) = -3/2.$$

It is, in fact, this symmetry which allows for the cancellation of the two asymptotic expansions generated by z_0 and z_1, which in turn will permit the determination of the asymptotic zeros of $F(\mu)$ with as much precision as necessary. Note that subsequently we often use the subscript s to state a property that is valid for both relevant saddle-points indexed by $s = 0, 1$. The deformed path of integration must satisfy the following conditions:

(i) The new path must go through a zero z_s of $w'(z)$.
(ii) $\Im w(z) = \Im w(z_s)$ on the new path.
(iii) $\Re w(z) \le \Re w(z_s)$ on the new path.

The next step consists of analyzing the hills, valleys, and paths of steepest descent and ascent of these saddle points. The level curves separating the hills and valleys of the saddle points z_s and the steepest paths emerging from them are given by

(i) steepest paths: $\Im\{\, w(z) - w(z_s)\,\} = 0$,
(ii) level curves: $\Re\{\, w(z) - w(z_s)\,\} = 0$,

where

$$\begin{aligned}w(z) = w(\xi + i\eta) &= 4i(\xi + i\eta) - (\xi + i\eta)^4\\ &= -4\eta - \xi^4 + 6\eta^2\xi^2 - \eta^4 + 4i(\xi - \xi^3\eta + \eta^3\xi).\end{aligned}$$

The level curves that separate the hills and valleys above and below the saddle points are determined by the real branches of the following equations:

$$-4\eta - \xi^4 + 6\eta^2\xi^2 - \eta^4 = \Re w(z_s),$$
$$\{\Re w(z_0), \Re w(z_1), \Re w(z_2)\} = \{-3/2, -3/2, 1\}.$$

Solving the biquadratic equation in ξ for ξ as a function of η wherever it is permitted (both ξ and η are real variables), the asymptotic behavior of these curves as $\eta \to \pm\infty$ is given by $\xi(\eta) \sim \pm\sqrt{3 \pm 2\sqrt{2}}\,\eta$; that is, they all end at $\infty \exp(\frac{2k+1}{8}\pi i)$ for some $k \in \mathbb{N}$.

The steepest paths out of each of the saddle points are determined by the real branches of the following cubic equations:

$$\xi - \xi^3\eta + \xi\eta^3 = \Im w(z_s)/4,$$
$$\{\Im w(z_0), \Im w(z_1), \Im w(z_2)\} = \left\{3\sqrt{3}/2, -3\sqrt{3}/2, 0\right\}.$$

It can be shown that the (steepest) descent paths emerging from the saddle points go from $\infty e^{i\pi/2} \leftarrow z_0 \to +\infty$ and from $-\infty \leftarrow z_1 \to \infty e^{i\pi/2}$; the (steepest) ascent paths go from $\infty e^{-i\pi/4} \leftarrow z_0 \to \infty e^{i\pi/4}$ and from $\infty e^{i3\pi/4} \leftarrow z_1 \to \infty e^{i5\pi/4}$. The steepest descent path through $z_2 = -i$ is the imaginary axis ($\xi = 0$), and as such, we may not deform the original path through it. Therefore, this saddle point does not contribute to the asymptotic expansion of $F(\mu)$. The convergence of the integral is preserved because the new paths always remain in the valleys of z_0 and z_1, and $w(e^{i\frac{\pi}{2}}z) = \mathcal{O}(-(iz)^4) = \mathcal{O}(-z^4)$ as $z \to +\infty$. The path deformation through z_0 and z_1 displayed in Fig. 1 is justified by a simple application of Cauchy's theorem. The solid lines represent the steepest paths and the dotted ones represent the level curves separating the hills and valleys above and below the saddle points z_0, z_1, z_2. The complete topography of the surface $u(\xi, \eta)$ is shown on Fig. 1.

Although we have just seen that it is possible to carry out the full (global) analysis of the steepest descent paths, it is not necessary do so. From a local analysis of the steepest directions at the saddle points, one can choose a simple path that will have the desired properties. Let α_s be the steepest descent direction at the saddle point z_s (also known as the axis of the saddle point z_s). It is determined by the inequality

$$(z - z_s)^2 \frac{w''(z_s)}{2!} \leq 0 \Longrightarrow \arg\{z - z_s\} = \pm\pi/2 - \arg\{w''(z_s)\}$$

(cf. [6, Chap. 5]). Since $\alpha_s = \lim_{z\to z_s} \arg\{z - z_s\}$ along the steepest paths, where the correct choice of α_s is determined by the direction in which the saddle point is crossed, we find that $\alpha_s = -(-1)^s\pi/6$. The path we consider is a combination of half-lines and line segments in the complex plane: $\gamma = \left(-\infty, -\sqrt{3}\right] \cup \mathcal{L}_1 \cup \mathcal{L}_0 \cup \left[\sqrt{3}, +\infty\right)$, where $\mathcal{L}_0$ and $\mathcal{L}_1$ are given by

$$\begin{cases} \mathcal{L}_0: & z(t) = e^{i\pi/6} + e^{-i\pi/6}t \quad -1 \leq t \leq 1, \\ \mathcal{L}_1: & z(t) = e^{i5\pi/6} + e^{i\pi/6}t \quad -1 \leq t \leq 1 \end{cases}$$

(see Fig. 2). Note that $\mathcal{L}_s$ is the parametrized line interval $z(t) = z_s + e^{i\alpha_s}t$ for $-1 \leq t \leq 1$, on which $z(0) = z_s$, $z(-1) = -\sqrt{3}$, $z(1) = i$ on $\mathcal{L}_1$ and $z(-1) = i$, $z(1) = \sqrt{3}$ on $\mathcal{L}_0$. Let $h_s(t) = \Re\{w(z(t)) - w(z_s)\} : [-1,1] \to \mathbb{R}$. Since $h_s(t) = -6t^2 - (-1)^s 2t^3 + t^4/2$, its only maximum for $t \in [-1,1]$ is located at $t = 0$, which

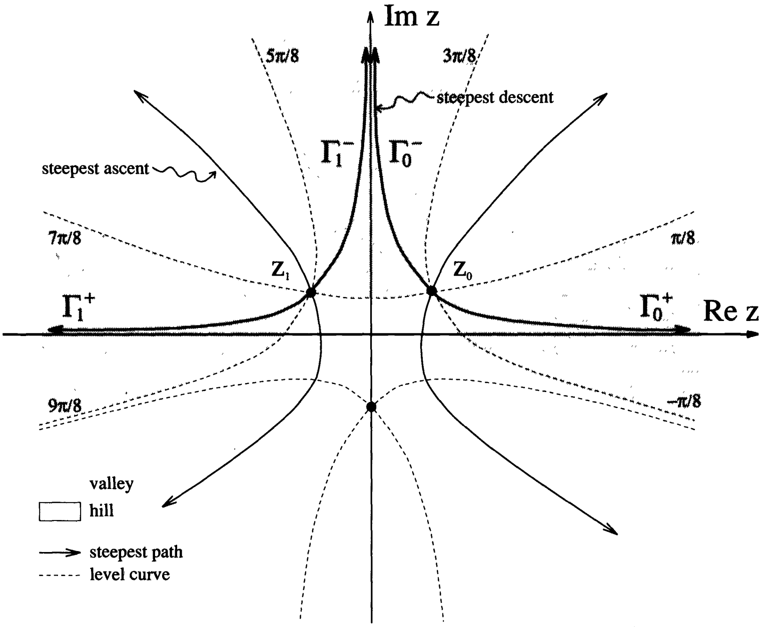


FIG. 1. *Hills, valleys, level curves, and steepest paths of the saddle points* $z_0 = e^{i\pi/6}$, $z_1 = e^{i5\pi/6}$ *relevant to the expansion of* $F(\mu) = \int_{-\infty}^{\infty} e^{\mu(4iz-z^4)}dz$ *as* $\mu \to +\infty$.

corresponds to $z = z_s$. Hence the maximal contribution on the paths $\mathcal{L}_0$, $\mathcal{L}_1$ occurs at z_0, z_1. Moreover, the contributions on the real intervals $(-\infty, -\sqrt{3}]$ and $[\sqrt{3}, +\infty)$ are negligible since $\Re w(z) = -z^4$. Thus the new path γ is admissible and we can still apply the series-reversion process that follows. Indeed, we would only need to verify that the steepest descent path runs from $\pm\infty$ to $i\infty$, and we would infer that the descent path γ is asymptotically equivalent to the steepest descent path Γ.

Now that the path deformation is justified, we can proceed with the expansion regardless of whether we use the exact steepest descent path Γ or just an approximate descent path γ. Since on the steepest descent paths $\Im\{w(z) - w(z_s)\} = 0$ for $s = 0, 1$, we have $w(z) - w(z_s) = -\tau$, $\tau \in \mathbb{R}^+$. Therefore,

(i) as $z \to e^{i\frac{k\pi}{2}}\infty$ for $k \in \mathbb{N}$, $\tau = w(z_s) - w(z) \to +\infty$,

(ii) as $z \to z_s$, $\tau = w(z_s) - w(z) \to 0$.

Deforming the path of integration from the real axis to

$$\Gamma = \Gamma_0 \cup \Gamma_1 = \Gamma_1^- - \Gamma_1^+ + \Gamma_0^+ - \Gamma_0^-,$$

$F(\mu)$ can be written as

$$F(\mu) = \int_\Gamma e^{\mu w(z)}\,dz = \int_{\Gamma_1^- - \Gamma_1^+} e^{\mu w(z)}\,dz + \int_{\Gamma_0^+ - \Gamma_0^-} e^{\mu w(z)}\,dz.$$

Here $\Gamma_s^\pm$ are the respective steepest descent paths emerging from the saddle points z_s, where the ± signs refer to the corresponding branches $z_s^\pm(\tau)$ of the solution to the equation $\tau = w(z_s) - w(z)$ on Γ_s. The assignment of the correct branches $z_s^\pm(\tau)$ to the two steepest descent paths emerging from the saddle points z_s will follow once we have the series expansion for $z_s^\pm(\tau)$ about $\tau = 0$. Proceeding with the path deformation,

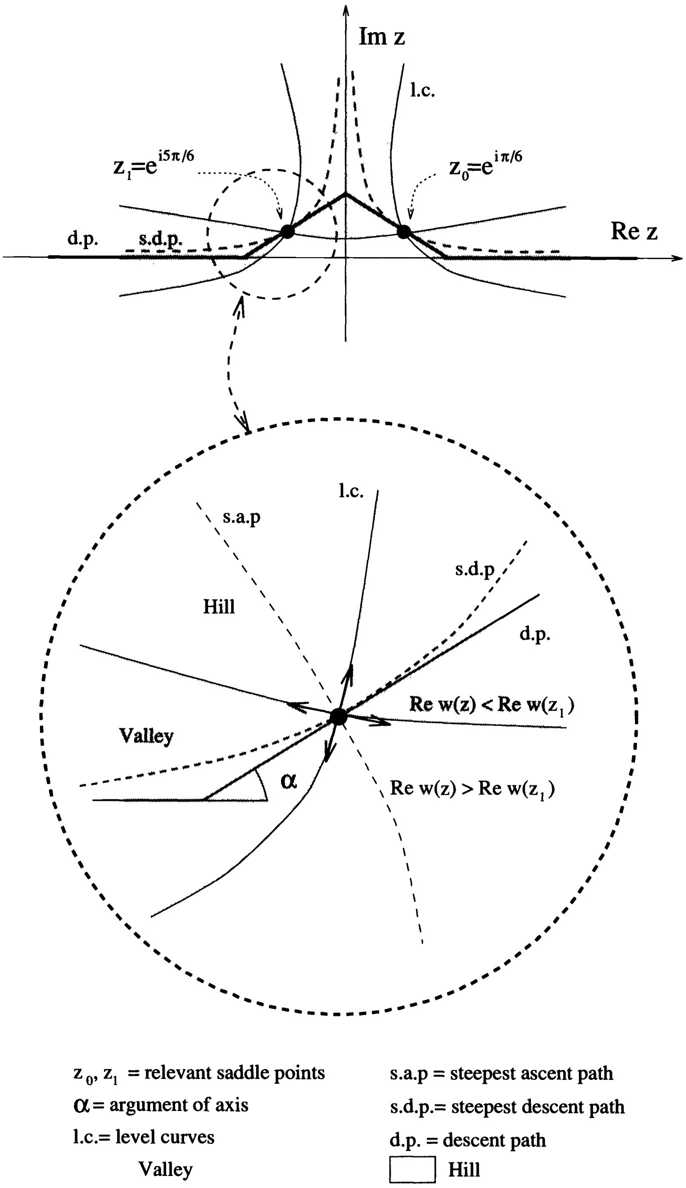


FIG. 2. *Comparison of the local/global analysis of the hills, valleys, level curves, and steepest paths of the saddle points* $z_0 = e^{i\pi/6}$, $z_1 = e^{i5\pi/6}$ *relevant to the expansion of* $F(\mu) = \int_{-\infty}^{\infty} e^{\mu(4iz-z^4)}dz$ *as* $\mu \to +\infty$.

we have

$$
\begin{aligned}
F(\mu) &= \int_{\Gamma_1^- - \Gamma_1^+} e^{\mu(w(z_1)-\tau)}\,dz + \int_{\Gamma_0^+ - \Gamma_0^-} e^{\mu(w(z_0)-\tau)}\,dz \\
&= \int_0^{+\infty} e^{\mu(w(z_1)-\tau)} \left(\frac{dz_1^-}{d\tau} - \frac{dz_1^+}{d\tau}\right)(\tau)\,d\tau \\
&\quad + \int_0^{+\infty} e^{\mu(w(z_0)-\tau)} \left(\frac{dz_0^+}{d\tau} - \frac{dz_0^-}{d\tau}\right)(\tau)\,d\tau \\
&= \sum_{s=0,1} (-1)^s e^{\mu w(z_s)} \int_0^{+\infty} \Phi_s(\tau)\, e^{-\mu\tau}\,d\tau,
\end{aligned}
$$

where $\Phi_s(\tau) = \left(\frac{dz_s^+}{d\tau} - \frac{dz_s^-}{d\tau}\right)(\tau)$.

2.2. Series reversion. We have transformed the original integral into a sum of Laplace integrals. It is now necessary to find a series expansion in the sense of Watson (cf. [11, §22]) for $\Phi_s(\tau)$ and justify the use of Watson's lemma in order to obtain the asymptotic expansion of $F(\mu)$ by term-by-term integration. From the Lagrange formula for the reversion of series (see [12]), we can invert the equation $\tau = w(z_s) - w(z)$ for z as a function of τ for τ near 0. We find that the two branches $z_s^\pm(\tau)$ corresponding to the steepest descent paths starting at $z = z_s^\pm(0) = z_s$ are given by convergent series in powers of $\sqrt{\tau}$ in a neighborhood of $\tau = 0$:

$$z_s^\pm(\tau) = z_s + \sum_{n=1}^{\infty} c_n(z_s)\,(\pm\sqrt{\tau})^n, \tag{9}$$

where

$$c_n(z_s) = \frac{1}{n!}\lim_{z\to z_s}\frac{d^{n-1}}{dz^{n-1}}\left\{f(z)^{-n/2}\right\}, \tag{10}$$

and $f(z) = (w(z_s) - w(z))/(z - z_s)^2$ is defined by

$$\tau = w(z_s) - w(z) = (z - z_s)^2 f(z) = (z - z_s)^2\left(6z_s^2 + 4z_s(z - z_s) + (z - z_s)^2\right).$$

In the definition of the coefficients $c_n(z_s)$, we are taking the principal value of the square root of $f(z)$ for which $\sqrt{f(z_s)} = \sqrt{6}z_s$. We thus have

$$\frac{dz_s^\pm}{d\tau}(\tau) = \sum_{n=1}^{\infty}\frac{n}{2}c_n(z_s)\,(\pm1)^n\tau^{n/2-1} \tag{11}$$

so that

$$\Phi_s(\tau) = \frac{dz_s^+}{d\tau} - \frac{dz_s^-}{d\tau} = \sum_{n=1}^{\infty} nc_n(z_s)\left(\frac{1-(-1)^n}{2}\right)\tau^{n/2-1}.$$

We may now look into assigning different branches of $z_s^\pm(\tau)$ to the different paths $\Gamma_s^\pm$. The motion of $z_s^\pm(\tau)$ along the paths $\Gamma_s^\pm$ as τ increases from $\tau = 0$ to some $0 < \tau \ll 1$ is determined by

$$z_s^\pm(\tau) = z_s \pm c_1(z_s)\sqrt{\tau} + \mathcal{O}(\tau) \qquad \text{as } \tau \to 0^+. \tag{12}$$

Since $c_1(z_s) = \lim_{z\to z_s} f(z)^{-1/2} = (\sqrt{6}z_s)^{-1}$, we have $\arg(c_1(z_s)) = -\arg(z_s)$. Since $\arg(z_0) = \pi/6$ and $\arg(z_1) = 5\pi/6$, we have $\cos(\arg(z_1)) = \cos(5\pi/6) < 0$ and $\cos(\arg(z_0)) = \cos(\pi/6) > 0$. Hence $z_0^+(\tau)$ has increasing real part for τ increasing, and therefore $z_0^+(\tau)$ is the branch that goes from z_0 to $+\infty$ and conversely $z_0^-(\tau)$ is the branch that goes from z_0 to $e^{i\frac{\pi}{2}}\infty$. We name the branches, respectively, Γ_0^+ and Γ_0^-. Similarly, we find that $z_1^-(\tau)$ is the branch that goes from z_1 to $e^{i\frac{\pi}{2}}\infty$ and $z_1^+(\tau)$ is the branch that goes from z_1 to $-\infty$. We name the branches, respectively, Γ_1^- and Γ_1^+.

2.3. Watson's lemma and asymptotic development. In order to apply Watson's lemma, we need to verify that

$$|\Phi_s(\tau)| = \left| \frac{dz_s^+}{d\tau} - \frac{dz_s^-}{d\tau} \right| < Ke^{b\tau}$$

for some positive K and b independent of τ when $\tau \geq \tau_0 > 0$ (see [11]):

$$\begin{aligned} &\tau = w(z_s) - w(z), \qquad \text{for } z \in \Gamma_s \\ \Longrightarrow &\frac{d\tau}{dz} = -w'(z) = 4(z^3 - i) = 4(z^3 - z_s^3) \\ \Longrightarrow &\Phi_s(\tau) = \frac{dz_s^+}{d\tau}(\tau) - \frac{dz_s^-}{d\tau}(\tau) = \frac{1}{4}\left(\frac{1}{z_s^+(\tau)^3 - z_s^3} - \frac{1}{z_s^-(\tau)^3 - z_s^3} \right). \end{aligned}$$

Since $|\Phi_s(\tau)| = \mathcal{O}(1)$ as $\tau \to +\infty$, using Watson's lemma, we may substitute the series expansion of $\Phi_s(\tau)$ (in the sense of Watson) and integrate term by term to obtain a compound asymptotic expansion with respect to the asymptotic sequence $\{\phi_n(\mu) = \mu^{-n}\}$ (see Definition 1.1):

$$\begin{aligned} F(\mu) &\overset{\mu\to+\infty}{\sim} \sum_{s=0,1} (-1)^s e^{\mu w(z_s)} \sum_{n=1}^{\infty} n c_n(z_s) \left(\frac{1-(-1)^n}{2} \right) \Gamma\left(\frac{n}{2}\right) \mu^{-n/2} \\ &= \sum_{s=0,1} (-1)^s e^{\mu w(z_s)} \sum_{n=0}^{\infty} (2n+1) c_{2n+1}(z_s) \Gamma\left(n + \frac{1}{2}\right) \mu^{-(n+1/2)}. \end{aligned}$$

We let

$$a_n(z_s) = (2n+1)c_{2n+1}(z_s) = \frac{1}{(2n)!} \lim_{z\to z_s} \frac{d^{2n}}{dz^{2n}} \left\{ f(z)^{-(n+1/2)} \right\} \tag{13}$$

so that

$$F(\mu) \overset{\mu\to+\infty}{\sim} \frac{1}{\sqrt{\mu}} \sum_{s=0,1} (-1)^s e^{\mu w(z_s)} \sum_{n=0}^{\infty} a_n(z_s) \Gamma\left(n + \frac{1}{2}\right) \mu^{-n}. \tag{14}$$

We proceed with the explicit determination of the coefficients $a_n(z_s)$ in the expansion of F, where $a_n(z_s)$ is the $2n$th coefficient in the Taylor expansion of $f(z)^{-(n+\frac{1}{2})}$ about $z = z_s$. Using the binomial theorem twice, we find

$$\begin{aligned} \left(\frac{f(z)}{6z_s^2} \right)^{-(n+1/2)} &= \left\{ 1 + \frac{2}{3z_s}(z - z_s) \left(1 + \frac{1}{4z_s}(z - z_s) \right) \right\}^{-(n+1/2)} \\ &= \sum_{k=0}^{\infty} \sum_{j=0}^{k} \binom{-n-\frac{1}{2}}{k} \binom{k}{j} \left(\frac{8}{3}\right)^k (4z_s)^{j-2k} (z - z_s)^{2k-j}. \end{aligned}$$

In order to take into account every term that contributes to $a_n(z_s)$, we find the range of k over which we sum by setting $2k - j = 2n$, that is $j = 2(k - n)$. Since j ranges from 0 to k, k ranges from n to $2n$. Thus summing over k from n to $2n$, we find

$$\begin{aligned} a_n(z_s) &= \frac{6^{-(n+1/2)}}{4^{2n}} z_s^{-(4n+1)} \sum_{k=n}^{2n} \binom{-n-\frac{1}{2}}{k} \binom{k}{2(k-n)} \left(\frac{8}{3}\right)^k \\ &= 6^{-(2n+1/2)} z_s^{-(4n+1)} \sum_{k=0}^{n} \binom{-n-\frac{1}{2}}{k+n} \binom{k+n}{2k} \left(\frac{8}{3}\right)^k. \end{aligned}$$

We introduce normalized coefficients ($\alpha_0 = 1$) which do not depend on z_s: let

$$\alpha_n = \Gamma(n+1/2)\sqrt{\frac{6}{\pi}}\, a_n(z_s)\, z_s^{4n+1};$$

then the compound asymptotic expansion of $F(\mu)$ with respect to the asymptotic sequence $\{\mu^{-n}\}$ as $\mu \to +\infty$ is

$$F(\mu) \overset{\mu\to+\infty}{\sim} \sqrt{\frac{\pi}{6\mu}} \sum_{s=0,1} (-1)^s e^{\mu w(z_s)} \sum_{n=0}^{\infty} \alpha_n z_s^{-(4n+1)} \mu^{-n},$$

where the rational coefficients α_n are given by

$$\alpha_n = \frac{\Gamma(\frac{1}{2}+n)}{6^{2n}\sqrt{\pi}} \sum_{k=0}^{n} \binom{-n-\frac{1}{2}}{k+n} \binom{k+n}{2k} \left(\frac{8}{3}\right)^k. \tag{15}$$

The first five values of α_n can easily be computed using a computer algebra system such as Mathematica [21] either directly from (15) or using the Mathematica code provided in Appendix C (see also Table 3):

$$\alpha_0 = 1,\ \alpha_1 = \frac{7}{144},\ \alpha_2 = \frac{385}{41472},\ \alpha_3 = \frac{39655}{17915904},\ \alpha_4 = \frac{665665}{10319560704}. \tag{16}$$

One can compare the coefficients α_n to the coefficients c_n introduced by Paris and Wood in [17, p. 397] and [18, p. 72, Eq. 3.4.16], which are found via a 4-term recurrence relation. It is easy to see that they are related by $\alpha_n = c_n/3^n$. Since

$$z_0 = e^{i\frac{\pi}{6}},\quad z_0^{4n} = e^{\frac{2n\pi}{3}i},\quad w(z_0) = w(e^{i\frac{\pi}{6}}) = -\frac{3}{2} + i\frac{3\sqrt{3}}{2},$$

$$z_1 = e^{i\frac{5\pi}{6}},\quad z_1^{4n} = e^{-\frac{2n\pi}{3}i},\quad w(z_1) = w(e^{i\frac{5\pi}{6}}) = -\frac{3}{2} - i\frac{3\sqrt{3}}{2},$$

we have

$$F(\mu) \overset{\mu\to+\infty}{\sim} \sqrt{\frac{\pi}{6\mu}} e^{-\frac{3}{2}\mu} \left\{ e^{i(3\frac{\sqrt{3}}{2}\mu-\frac{\pi}{6})} \sum_{n=0}^{\infty} \alpha_n e^{-i\frac{2n\pi}{3}} \mu^{-n} \right.$$
$$\left. + e^{-i(3\frac{\sqrt{3}}{2}\mu-\frac{\pi}{6})} \sum_{n=0}^{\infty} \alpha_n e^{i\frac{2n\pi}{3}} \mu^{-n} \right\}.$$

We can formulate this as a generalized asymptotic expansion with respect to the asymptotic sequence $\{\mu^{-n}\}$ (see Definition 1.2):

$$F(\mu) \underset{\{\mu^{-n}\}}{\overset{\mu\to+\infty}{\sim}} \sqrt{\frac{2\pi}{3\mu}} e^{-\frac{3}{2}\mu} \sum_{n=0}^{\infty} \alpha_n \cos\left(3\frac{\sqrt{3}}{2}\mu - \frac{\pi}{6} - \frac{2n\pi}{3}\right) \mu^{-n}. \tag{17}$$

2.4. Asymptotic zeros of $F(\mu)$. To determine the asymptotic zeros of $F(\mu)$, we make use of the compound asymptotic relation

$$h(\mu) = \sqrt{\frac{3\mu}{2\pi}} e^{\frac{3}{2}\mu} F(\mu) \overset{\mu\to+\infty}{\sim} \cos\left(3\frac{\sqrt{3}}{2}\mu - \frac{\pi}{6}\right) \sum_{n=0}^{\infty} \alpha_n \cos\left(\frac{2n\pi}{3}\right) \mu^{-n}$$
$$+ \sin\left(3\frac{\sqrt{3}}{2}\mu - \frac{\pi}{6}\right) \sum_{n=1}^{\infty} \alpha_n \sin\left(\frac{2n\pi}{3}\right) \mu^{-n}. \tag{18}$$

Let $h_m(\mu)$ be the partial sum

$$h_m(\mu) = \cos\left(3\frac{\sqrt{3}}{2}\mu - \frac{\pi}{6}\right)\sum_{n=0}^{m} \alpha_n \cos\left(\frac{2n\pi}{3}\right)\mu^{-n} + \sin\left(3\frac{\sqrt{3}}{2}\mu - \frac{\pi}{6}\right)\sum_{n=1}^{m} \alpha_n \sin\left(\frac{2n\pi}{3}\right)\mu^{-n}. \tag{19}$$

Then solving the equation $h_m(\mu) = 0$ is equivalent to solving

$$\tan\left(3\frac{\sqrt{3}}{2}\mu - \frac{\pi}{6}\right) = -\frac{\sum_{n=0}^{m} \alpha_n \cos\left(\frac{2n\pi}{3}\right)\mu^{-n}}{\sum_{n=1}^{m} \alpha_n \sin\left(\frac{2n\pi}{3}\right)\mu^{-n}}; \tag{20}$$

that is,

$$3\frac{\sqrt{3}}{2}\mu - \frac{\pi}{6} = k\pi - \tan^{-1}\left(\frac{\sum_{n=0}^{m} \alpha_n \cos\left(\frac{2n\pi}{3}\right)\mu^{-n}}{\sum_{n=1}^{m} \alpha_n \sin\left(\frac{2n\pi}{3}\right)\mu^{-n}}\right), \tag{21}$$

where $|\tan^{-1} x| < \pi/2$. For m sufficiently large, we have, as $\mu \to +\infty$,

$$\tan^{-1}\left(\frac{\sum_{n=0}^{m} \alpha_n \cos\left(\frac{2n\pi}{3}\right)\mu^{-n}}{\sum_{n=1}^{m} \alpha_n \sin\left(\frac{2n\pi}{3}\right)\mu^{-n}}\right) = \frac{\pi}{2} - \frac{7}{96\sqrt{3}\,\mu} + \frac{7}{576\sqrt{3}\,\mu^2} + \frac{49}{497664\sqrt{3}\,\mu^4} - \frac{379351}{268738560\sqrt{3}\,\mu^5} + \mathcal{O}\left(\frac{1}{\mu^6}\right). \tag{22}$$

Thus we have, as $\mu \to +\infty$,

$$3\frac{\sqrt{3}}{2}\mu - \frac{\pi}{6} = k\pi - \left\{\frac{\pi}{2} - \frac{7}{96\sqrt{3}\,\mu} + \frac{7}{576\sqrt{3}\,\mu^2} + \frac{49}{497664\sqrt{3}\,\mu^4} - \frac{379351}{268738560\sqrt{3}\,\mu^5}\right\} + \mathcal{O}\left(\frac{1}{\mu^6}\right). \tag{23}$$

Letting

$$\mu_k^{(0)} = \frac{2\pi}{3\sqrt{3}}(k - 1/3), \qquad k \geq 1, \tag{24}$$

we have

$$\mu = \mu_k^{(0)} + \frac{7}{432\,\mu}\left\{1 - \frac{1}{6\,\mu} - \frac{7}{5184\,\mu^3} + \frac{54193}{2799360\,\mu^4}\right\} + \mathcal{O}\left(\frac{1}{\mu^6}\right). \tag{25}$$

We now state a lemma which enables us to improve the first-order asymptotic estimate to a higher-order one. Its proof is based on the reversion of (asymptotic) series by either Lagrange's formula (see [15, p. 21, §8.4]) or using the method of successive resubstitution (see Appendix A).

Lemma 2.1. *If* $\mu^{(0)} = \mathcal{O}(k)$ *as* $k \to +\infty$ *and*

$$\mu = \mu^{(0)} + \frac{a_1}{\mu}\left(a_2 + \frac{a_3}{\mu} + \frac{a_4}{\mu^2} + \frac{a_5}{\mu^3} + \frac{a_6}{\mu^4}\right) + \mathcal{O}\left(\frac{1}{\mu^6}\right)$$

is an asymptotic relation which holds as $\mu \to +\infty$, *then*

$$\mu = \mu^{(0)} + \frac{a_1}{\mu^{(0)}} \left(a_2 + \frac{a_3}{\mu^{(0)}} + \frac{a_4 - a_1 a_2^2}{{\mu^{(0)}}^2} + \frac{a_5 - 3a_1a_2a_3}{{\mu^{(0)}}^3} \right.$$
$$\left. + \frac{a_6 - 2a_1a_3^2 + 2a_1^2a_2^3 - 4a_1a_2a_4}{{\mu^{(0)}}^4} \right) + \mathcal{O}\left(\frac{1}{k^6}\right) \qquad \textit{as } k \to +\infty.$$

Combining (15), (24), (25), and Lemma 2.1, we have proven the following.

COROLLARY 2.1. *For* $n = 2$, *the rational coefficients* $\alpha_{2,j}$ *in the expansion of* $\mathcal{F}_2(\mu) = \int_{-\infty}^{\infty} e^{\mu(4ix - x^4)}dx$ *are*

$$\alpha_{2,j} = \frac{\Gamma(\frac{1}{2} + j)}{6^{2j}\sqrt{\pi}} \sum_{k=0}^{j} \binom{-j - \frac{1}{2}}{k + j} \binom{k + j}{2k} \left(\frac{8}{3}\right)^k .$$

For $k \geq 1$, *the approximation of the* k*th ordered positive zero* $\mu_{k,2}$ *of* $\mathcal{F}_2(\mu)$ *is given by*

$$\mu_{k,2}^{(0)} = \frac{2\pi}{3\sqrt{3}} \left(k - \frac{1}{3}\right) + \mathcal{O}\left(\frac{1}{k}\right) \qquad \textit{as } k \to +\infty.$$

The sixth-order approximation is

$$\mu_{k,2} = \mathcal{G}_2\left(\mu_{k,2}^{(0)}\right) + \mathcal{O}\left(\frac{1}{k^6}\right) \qquad \textit{as } k \to +\infty,$$

$$\mathcal{G}_2(\mu) = \mu + \frac{7}{432\mu}\left(1 - \frac{1}{6\mu}\left(1 + \frac{7}{72\mu}\left(1 - \frac{5}{12\mu}\left(1 + \frac{53143}{18900\mu}\right)\right)\right)\right).$$

The fact that this is actually the expansion of the kth ordered zero of $F(\mu)$ can be proved by the argument principle (see [12, 15]).

3. Asymptotic expansion of $\mathcal{F}_n(\mu)$ as $\mu \to +\infty$. For $n \geq 2$ and $|\arg \mu| < \pi/2$, we consider the function introduced in (6):

$$\mathcal{F}_n(\mu) = \int_{-\infty}^{\infty} e^{\mu\,(2niz - z^{2n})} dz.$$

The saddle points z_s of the integrand and their contributions are given by

$$\left\{ \begin{array}{l} w_n(z) = 2niz - z^{2n}, \\ w_n'(z_s) = 2niz_s - 2nz_s^{2n-1} = 0, \\ 0 = \frac{z_s}{2n} w_n'(z_s) = iz_s - z_s^{2n}, \end{array} \right. \implies \left\{ \begin{array}{l} z_s = \exp(\frac{i\pi}{4n-2}(1 + 4k)), \\ w_n(z_s) = (2n - 1)iz_s, \\ \Re w_n(z_s) = -(2n - 1)\Im z_s. \end{array} \right.$$

Thus

$$\Re w_n(z_s) = -(2n - 1)\sin\left(\frac{\pi}{4n - 2}(1 + 4k)\right), \qquad k = 0, 1, \ldots, 2n - 2.$$

On the two steepest descent paths emerging from relevant saddle points, we have

$$w_n(z) - w_n(z_s) = -(z - z_s)^2 f_n(z) = -\tau \leq 0,$$
$$f_n(z) = -\sum_{k=0}^{2n-2} \frac{w_n^{(k+2)}(z_s)}{(k + 2)!}(z - z_s)^k.$$

As before, we expect to have contributions from two equally relevant saddle points which come in symmetric pairs satisfying the relation $z_0 = -\overline{z_1}$ (see §3.2). The assignment of the branches $z_s^{\pm}(\tau)$ to the steepest descent paths must be dealt with carefully. From Lagrange's formula, we express $z_s^{\pm}(\tau)$ as convergent series in powers of $\sqrt{\tau}$ in a neighborhood of $\tau = 0$:

$$z_s^{\pm}(\tau) = z_s + \sum_{j=1}^{\infty} \mathfrak{c}_{n,j}(z_s)(\pm\sqrt{\tau})^j$$

with

$$\mathfrak{c}_{n,j}(z_s) = \frac{1}{j!} \lim_{z \to z_s} \frac{d^{j-1}}{dz^{j-1}} \left\{ f_n(z)^{-j/2} \right\}.$$

Since $\mathfrak{c}_{n,1}(z_s) = f_n(z_s)^{-1/2} = \binom{2n}{2}^{-1/2} z_s^{1-n}$, the behavior of $z_s^{\pm}(\tau)$ in a neighborhood of $\tau = 0$ is determined by

$$z_s^{\pm}(\tau) = z_s \pm \frac{z_s^{1-n}}{\sqrt{\binom{2n}{2}}} \sqrt{\tau} + \mathcal{O}(\tau) \qquad \text{as } \tau \to 0^+.$$

In the definition of $\mathfrak{c}_{n,j}(z_s)$, we have taken the principal value of $\sqrt{f_n(z)}$ for which $\sqrt{f_n(z_s)} = \sqrt{\binom{2n}{2}} z_s^{n-1}$. In what follows, we assume that the pair of relevant saddle points $\{z_0, z_1 = -\overline{z_0}\}$ is the first pair with smallest positive imaginary part, that is, $z_0 = \exp(\frac{i\pi}{4n-2})$ and $z_1 = -\exp(-\frac{i\pi}{4n-2})$. This is so because it is the only pair whose steepest descent paths are admissible in the sense that the original path of integration cannot be deformed through any of the steepest descent paths emerging from the other saddle points with negative imaginary part. Indeed, such saddle points would yield an incorrect increasing exponential behavior since we would then have $\Re w_n(z_s) = -(2n-1)\,\Im z_s > 0$. None of the other saddle points with positive imaginary part (all saddle points come in symmetric pair $\{z_s, -\overline{z_s}\}$ except those for which $\Re z_s = 0$) have admissible steepest descent paths. Even if it were possible to deform the path of integration through another pair, their contribution would be exponentially smaller than that of the pair $\{z_0, -\overline{z_0}\}$. On the steepest descent paths $\Gamma_1^{\pm}$ corresponding to the equation $\tau = w(z_1) - w(z)$, we have

$$z_1^{\pm}(\tau) = z_1 \mp (-1)^n \frac{e^{i\frac{n-1}{4n-2}\pi}}{\sqrt{\binom{2n}{2}}} \sqrt{\tau} + \mathcal{O}(\tau) \qquad \text{as } \tau \to 0^+.$$

Following the motion of $z_1^{\pm}(\tau)$ along $\Gamma_1^{\pm}$ for increasing $\tau > 0$ as in (12), one can correctly choose the branches $z_1^{\pm}(\tau)$. Hence we see that the assignment of the branches $z_1^{\pm}(\tau)$ changes from the upper branch to the lower one (as shown in Fig. 3) depending on the parity of the index n. Note that this feature is not present in the case of $z_0^{\pm}(\tau)$. In the case $n = 3$, the paths of steepest descent labeled $\Gamma_0^{\pm}$ go from $e^{\frac{\pi i}{3}}\infty \leftarrow z_0 = e^{\frac{\pi i}{10}} \to +\infty$; the ones labeled $\Gamma_1^{\pm}$ go from $-\infty \leftarrow z_1 = -\overline{z_0} = e^{\frac{9\pi i}{10}} \to e^{\frac{2\pi i}{3}}\infty$. There is a third path labeled Γ_2 which connects Γ_1^{+} and Γ_0^{-}. This third path remains in the common valley of the saddle points z_0 and $-\overline{z_0}$ and is subdominant with respect to the other paths. In other words, its contribution is exponentially small compared to the contributions of Γ_0 and Γ_1. In the general case, the topography remains similar. We

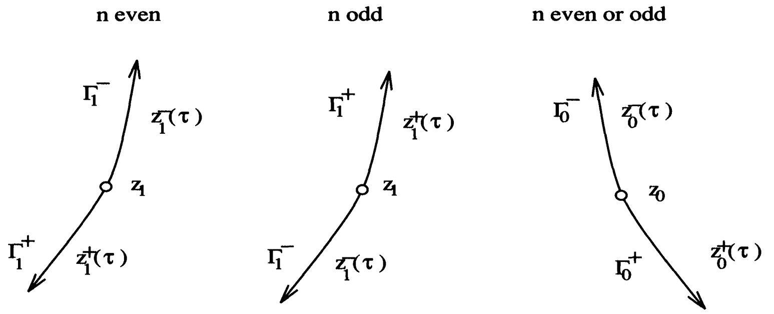


FIG. 3. *Interchange of the branch assignment $z_1^{\pm}(\tau)$ according to the parity of the index n.*

expect the paths of steepest descent emerging from the saddle points at $z_0 = e^{\frac{\pi i}{4n-2}}$ and $z_1 = -\overline{z_0}$ to end in respective valleys. Let

$$\Gamma_n = \Gamma_0^+ - \Gamma_0^- + \Gamma_1^{(-)^{n+1}} - \Gamma_1^{(-)^n} + \Gamma_2$$

denote the new path of integration, where

$$\Gamma_1^{(-)^n} = \begin{cases} \Gamma_1^+ & \text{if } n \text{ even,} \\ \Gamma_1^- & \text{if } n \text{ odd} \end{cases}$$

and reciprocally for $\Gamma_1^{(-)^{n+1}}$. The asymptotic behavior of these paths is as follows:

$$\begin{aligned} &\Gamma_0^+ : z_0 \to +\infty, \qquad \Gamma_0^- : z_0 \to \infty\, e^{i\pi/n} \\ &\Gamma_1^{(-)^{n+1}} : z_1 \to \infty\, e^{i(\pi-\pi/n)}, \qquad \Gamma_1^{(-)^n} : z_1 \to -\infty, \\ &\Gamma_2 : \infty\, e^{i(\pi-\pi/n)} \to \infty\, e^{i\pi/n}. \end{aligned}$$

One can also choose a simple descent path which is the straight line $\gamma_n : \Im z = \Im z_0 = \sin(\frac{\pi}{4n-2})$ in the complex plane going through both saddle points z_0 and $-\overline{z_0}$ parallel to the real axis depicted in Fig. 4 as a dashed path (see the argument in [9, Lem. 2.1]). We now deform the original contour of integration along the path Γ_n or γ_n as in Fig. 4, and we take into account the interchange of the branches $z_1^{\pm}(\tau)$ based on the parity of the index n by including a factor $(-1)^{s\cdot n}$, $s = 0, 1$, in (14). We notice that $|\Phi_s(\tau)| = |dz_s^+/d\tau - dz_s^-/d\tau| = \mathcal{O}(1)$ as $\tau \to +\infty$, so we can appeal to Watson's lemma to find a compound asymptotic expansion for $\mathcal{F}_n(\mu)$ with respect to the asymptotic sequence $\{\phi_j(\mu) = \mu^{-j}\}$:

$$\mathcal{F}_n(\mu) \overset{\mu\to+\infty}{\sim} \frac{1}{\sqrt{\mu}} \sum_{s=0,1} (-1)^{s(1+n)} e^{\mu w_n(z_s)} \sum_{j=0}^{\infty} \mathfrak{a}_{n,j}(z_s)\, \Gamma\Big(j + \frac{1}{2}\Big) \mu^{-j}, \tag{26}$$

$$\mathfrak{a}_{n,j}(z_s) = (2j+1)\, \mathfrak{c}_{n,2j+1}(z_s) = \frac{1}{(2j)!} \lim_{z\to z_s} \frac{d^{2j}}{dz^{2j}} \left\{ f_n(z)^{-(j+1/2)} \right\}. \tag{27}$$

3.1. Coefficients of the expansion.

Let

$$\mathfrak{a}_{n,j}(z_s) = \frac{1}{(2j)!} \lim_{z\to z_s} \frac{d^{2j}}{dz^{2j}} \left\{ g_j(f_n(z)) \right\},$$

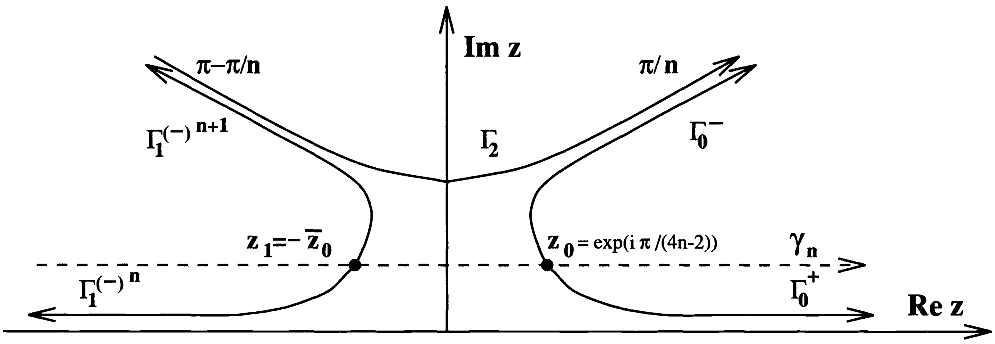


FIG. 4. *Steepest descent path (solid)* $\Gamma_n = \Gamma_0^+ - \Gamma_0^- + \Gamma_1^{(-)^{n+1}} - \Gamma_1^{(-)^n} + \Gamma_2$ *and alternate path (dashed)* $\gamma_n : \Im z = \Im z_0 = \sin(\frac{\pi}{4n-2})$ *in the expansion of* $\mathcal{F}_n(\mu) = \int_{-\infty}^{\infty} e^{\mu(2niz-z^{2n})}\,dz$ *as* $\mu \to +\infty$ *for* $n \geq 3$.

where

$$g_j(z) = z^{-j-1/2}, \quad f_n(z) = \sum_{k=0}^{2n-2} \binom{2n}{k+2} z_s^{2n-2-k}(z-z_s)^k.$$

According to the definition of Pochhammer's symbol $(z)_n$, we let $(1/2-j-m)_m = \Gamma(1/2-j)/\Gamma(1/2-j-m) = (-j-1/2)\cdot(-j-3/2)\cdots(1/2-j-m)$ for $m \geq 1$ and let $(1/2-j)_0 = 1$ so that we may write

$$g_j^{(m)}(f_n(z_s)) = \left(n(2n-1)\,z_s^{2n-2}\right)^{-j-1/2-m}(1/2-j-m)_m, \tag{28}$$

$$\frac{f_n^{(k)}(z_s)}{k!} = \begin{cases} \binom{2n}{k+2} z_s^{2n-2-k}, & 0 \leq k \leq 2n-2, \\ 0, & k \geq 2n-1. \end{cases} \tag{29}$$

Using (28) and (29) in Faà di Bruno's formula (see (50)), we find

$$\begin{aligned} \mathfrak{a}_{n,j}(z_s) = \sum_{m=0}^{2j} \Bigg\{ & \left(n(2n-1)\,z_s^{2n-2}\right)^{-j-\frac{1}{2}-m}(1/2-j-m)_m \\ & \times \sum_{\boldsymbol{\sigma}}{}' \prod_{k=1}^{2j} \frac{1}{\sigma_k!}\left(\binom{2n}{k+2} z_s^{2n-2-k}\right)^{\sigma_k} \Bigg\}. \end{aligned} \tag{30}$$

The summation $\sum'_{\boldsymbol{\sigma}}$ is taken over all $2j$-vectors $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{2j}) \in \mathbb{N}^{2j}$ such that the following conditions are simultaneously satisfied:

$$\left\{ \begin{array}{c} \sigma_1 + \sigma_2 + \cdots + \sigma_{2j} = m, \\ \sigma_1 + 2\sigma_2 + \cdots + 2j\sigma_{2j} = 2j, \\ \sigma_k = 0 \quad \forall k \geq 2n-1 \end{array} \right\}. \tag{31}$$

The last condition $\sigma_k = 0 \forall k \geq 2n-1$ arises from the fact that $f_n(z)$ is a polynomial of order $2n-2$ and therefore any derivative of order $k \geq 2n-1$ of $f_n(z)$ is zero. In order for the product to be nontrivial, we must set the corresponding powers σ_k to zero when $k \geq 2n-1$ (see (50)). This amounts to using the truncated $(2n-2)$-vector $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{2n-2}) \in \mathbb{N}^{2n-2}$ in the last product, whereby we can reduce conditions (31) to a new set of conditions

$$\left\{ \begin{array}{c} \sigma_1 + \sigma_2 + \cdots + \sigma_{2n-2} = m, \\ \sigma_1 + 2\sigma_2 + \cdots + (2n-2)\sigma_{2n-2} = 2j, \\ \sigma_k = 0 \quad \forall k \geq 2n-1 \end{array} \right\}. \tag{32}$$

We express (30) as

$$\begin{aligned} \mathfrak{a}_{n,j}(z_s) = \sum_{m=0}^{2j} \Bigg\{ & \left(n(2n-1)\, z_s^{2n-2}\right)^{-j-\frac{1}{2}-m} (1/2-j-m)_m \\ & \times \sum_{\sigma}{}' \prod_{k=1}^{2n-2} \frac{1}{\sigma_k!} \left(\binom{2n}{k+2} z_s^{2n-2-k} \right)^{\sigma_k} \Bigg\}. \end{aligned} \tag{33}$$

Using the first and second condition in (32), we notice that

$$\prod_{k=1}^{2n-2} z_s^{(2n-2-k)\cdot\sigma_k} = z_s^{(2n-2)\sum_{k=1}^{2n-2}\sigma_k - \sum_{k=1}^{2n-2} k\sigma_k} = z_s^{(2n-2)m-2j}.$$

We can therefore extract the z_s dependency from the summation signs in (33):

$$\begin{aligned} \mathfrak{a}_{n,j}(z_s) = \frac{z_s^{1-n(1+2j)}}{\left(n(2n-1)\right)^{j+1/2}} \sum_{m=0}^{2j} \Bigg\{ & \frac{(1/2-j-m)_m}{\left(n(2n-1)\right)^m} \\ & \times \sum_{\sigma}{}' \prod_{k=1}^{2n-2} \frac{1}{\sigma_k!} \binom{2n}{k+2}^{\sigma_k} \Bigg\}. \end{aligned} \tag{34}$$

We introduce normalized coefficients ($\alpha_{n,0} = 1$) which do not depend on the saddle points z_s:

$$\alpha_{n,j} = \Gamma(j+1/2) \sqrt{\frac{n(2n-1)}{\pi}} z_s^{n(1+2j)-1} \, \mathfrak{a}_{n,j}(z_s). \tag{35}$$

The jth coefficient for $j \geq 0$ is then a rational number given by

$$\begin{aligned} \alpha_{n,j} = \frac{\Gamma(j+1/2)}{\sqrt{\pi}\left(n(2n-1)\right)^j} \sum_{m=0}^{2j} \Bigg\{ & \frac{(1/2-j-m)_m}{\left(n(2n-1)\right)^m} \\ & \times \sum_{\sigma}{}' \prod_{k=1}^{2n-2} \frac{1}{\sigma_k!} \binom{2n}{k+2}^{\sigma_k} \Bigg\}, \end{aligned} \tag{36}$$

where the summation $\sum'_{\boldsymbol{\sigma}}$ is taken over all possible $\boldsymbol{\sigma} \in \mathbb{N}^{2n-2}$ such that

$$\left\{ \begin{array}{c} \sigma_1 + \sigma_2 + \cdots + \sigma_{2n-2} = m, \\ \sigma_1 + 2\sigma_2 + \cdots + (2n-2)\sigma_{2n-2} = 2j \end{array} \right\}. \tag{37}$$

A Mathematica code is provided for the reader's convenience in Appendix C to compute the coefficients $\alpha_{n,j}$ from (36) and (37) (see also Table 3).

3.2. Asymptotic expansion of $\mathcal{F}_n(\mu)$ as $\mu \to +\infty$. From equation (26), we find

$$\mathcal{F}_n(\mu) \overset{\mu\to+\infty}{\sim} \sqrt{\frac{\pi}{n(2n-1)\mu}} \sum_{s=0,1} (-1)^{s(1+n)} e^{\mu w_n(z_s)} \sum_{j=0}^{\infty} \alpha_{n,j}\, z_s^{1-n(1+2j)} \mu^{-j}.$$

Since $z_0 = -\overline{z_1} \in \Delta = \{z \in \mathbb{C} : |z| = 1\}$ and $w_n(z_0) = \overline{w_n(z_1)}$, we find

$$\mathcal{F}_n(\mu) \underset{\{\mu^{-j}\}}{\overset{\mu\to+\infty}{\sim}} \sqrt{\frac{4\pi}{n(2n-1)\mu}}\, e^{\mu \Re w_n(z_0)}\, \mathcal{H}_n(\mu), \tag{38}$$

where $\mathcal{H}_n(\mu)$ should be interpreted as a generalized asymptotic expansion with respect to the asymptotic sequence $\{\mu^{-j}\}$:

$$\mathcal{H}_n(\mu) = \sum_{j=0}^{\infty} \frac{\alpha_{n,j}}{\mu^j} \cos\left(\mu\, \Im w_n(z_0) + (1 - n(1+2j)) \arg(z_0)\right). \tag{39}$$

Since $w_n(z_0) = (2n-1) i z_0$ and $z_0 = e^{\frac{\pi i}{4n-2}}$,

$$\Re w_n(z_0) = -(2n-1)\Im z_0 = -(2n-1)\sin\left(\frac{\pi}{4n-2}\right),$$

$$\Im w_n(z_0) = (2n-1)\Re z_0 = (2n-1)\cos\left(\frac{\pi}{4n-2}\right).$$

Thus (38) is

$$\mathcal{F}_n(\mu) \underset{\{\mu^{-j}\}}{\overset{\mu\to+\infty}{\sim}} \sqrt{\frac{4\pi}{n(2n-1)\mu}} \exp\left\{-\mu\,(2n-1)\sin\left(\frac{\pi}{4n-2}\right)\right\} \mathcal{H}_n(\mu) \tag{40}$$

and (39) becomes

$$\mathcal{H}_n(\mu) = \sum_{j=0}^{\infty} \frac{\alpha_{n,j}}{\mu^j} \cos\left(\mu\,(2n-1)\cos\left(\frac{\pi}{4n-2}\right) + \pi\,\frac{1-n(1+2j)}{4n-2}\right). \tag{41}$$

3.3. Asymptotic zeros of $\mathcal{F}_n(\mu)$. $\mathcal{H}_n(\mu)$ is the component of the expansion of $\mathcal{F}_n(\mu)$ that determines its zeros, and its mth partial sum $\mathcal{H}_{n,m}(\mu)$ can also be expressed as a compound asymptotic expansion (see Definition 1.1):

$$\begin{aligned}\mathcal{H}_{n,m}(\mu) &= \cos\left(\mu\,(2n-1)\cos\left(\frac{\pi}{4n-2}\right) - \pi\,\frac{n-1}{4n-2}\right)\sum_{j=0}^{m} \frac{\alpha_{n,j}}{\mu^j}\cos\left(\frac{nj\pi}{2n-1}\right)\\ &+ \sin\left(\mu\,(2n-1)\cos\left(\frac{\pi}{4n-2}\right) - \pi\,\frac{n-1}{4n-2}\right)\sum_{j=1}^{m} \frac{\alpha_{n,j}}{\mu^j}\sin\left(\frac{nj\pi}{2n-1}\right).\end{aligned} \tag{42}$$

The first-order approximation for the zeros of $\mathcal{H}_{n,m}(\mu)$ is found immediately by setting $\cos(\mu\,(2n-1)\cos(\frac{\pi}{4n-2}) - \pi\,\frac{n-1}{4n-2}) = 0$. Thus we find that the kth ordered positive zero $\mu_{k,n}^{(0)}$ of $\mathcal{F}_n(\mu)$ is given (for $k \geq 1$ so that $\mu_{k,n}^{(0)} > 0$) by

$$\mu_{k,n}^{(0)} = \frac{\pi}{4n-2}\sec\left(\frac{\pi}{4n-2}\right)\left(\frac{n-1}{2n-1} - 1 + 2k\right) + \mathcal{O}\left(\frac{1}{k}\right) \tag{43}$$

as $k \to +\infty$. Solving the equation $\mathcal{H}_{n,m}(\mu) = 0$ yields

$$\begin{aligned}(44)\qquad &\mu(2n-1)\cos\left(\frac{\pi}{4n-2}\right) - \pi\frac{n-1}{4n-2}\\ &\qquad = k\pi - \tan^{-1}\left(\frac{\sum_{j=0}^{m}\frac{\alpha_{n,j}}{\mu^j}\cos\left(\frac{nj\pi}{2n-1}\right)}{\sum_{j=1}^{m}\frac{\alpha_{n,j}}{\mu^j}\sin\left(\frac{nj\pi}{2n-1}\right)}\right).\end{aligned}$$

We expand the $\tan^{-1}$ for large μ and sufficiently large m and combine it with (43) and (44) to find

$$\begin{aligned}\mu = \mu_{k,n}^{(0)} + \frac{\sec(\frac{\pi}{4n-2})}{(2n-1)\mu}\Bigg\{&\alpha_{n,1}\sin\left(\frac{n\pi}{2n-1}\right) - \left(\alpha_{n,1}^2 - 2\alpha_{n,2}\right)\sin\left(\frac{2n\pi}{2n-1}\right)\frac{1}{2\mu}\\ &+ \left(\alpha_{n,1}^3 - 3\alpha_{n,1}\alpha_{n,2} + 3\alpha_{n,3}\right)\sin\left(\frac{3n\pi}{2n-1}\right)\frac{1}{3\mu^2}\\ &- \left(\alpha_{n,1}^4 - 4\alpha_{n,1}^2\alpha_{n,2} + 2\alpha_{n,2}^2 + 4\alpha_{n,1}\alpha_{n,3} - 4\alpha_{n,4}\right)\sin\left(\frac{4n\pi}{2n-1}\right)\frac{1}{4\mu^3}\\ &+ \big(\alpha_{n,1}^5 - 5\alpha_{n,1}^3\alpha_{n,2} + 5\alpha_{n,1}\alpha_{n,2}^2 + 5\alpha_{n,1}^2\alpha_{n,3} - 5\alpha_{n,2}\alpha_{n,3}\\ &- 5\alpha_{n,1}\alpha_{n,4} + 5\alpha_{n,5}\big)\sin\left(\frac{5n\pi}{2n-1}\right)\frac{1}{5\mu^4}\Bigg\} + \mathcal{O}\left(\frac{1}{\mu^6}\right).\end{aligned}$$

Appealing to Lemma 2.1, we define

$$\begin{aligned}(45)\quad \mathcal{G}_n(\mu) = \mu + \frac{\sec(\frac{\pi}{4n-2})}{(2n-1)\mu}\Bigg\{&\alpha_{n,1}\sin\left(\frac{n\pi}{2n-1}\right) - \frac{\alpha_{n,1}^2 - 2\alpha_{n,2}}{2\mu}\sin\left(\frac{2n\pi}{2n-1}\right)\\ &+ \frac{\alpha_{n,1}^3 - 3\alpha_{n,1}\alpha_{n,2} + 3\alpha_{n,3}}{3\mu^2}\sin\left(\frac{3n\pi}{2n-1}\right) - \frac{\sec(\frac{\pi}{4n-2})}{(2n-1)}\frac{\alpha_{n,1}^2}{\mu^2}\sin^2\left(\frac{n\pi}{2n-1}\right)\Bigg\}.\end{aligned}$$

Let $\mu_{k,n}$ denote the kth ordered positive zero of $\mathcal{F}_n(\mu)$ so that the fourth-order approximation of $\mu_{k,n}$ is given by

$$\mu_{k,n} = \mathcal{G}_n\left(\mu_{k,n}^{(0)}\right) + \mathcal{O}\left(\frac{1}{k^4}\right) \quad \text{as } k \to +\infty.$$

Combining (36), (37), (40), (41), (43), and (45), Theorem 1.1 is proved. □

3.4. Coefficients $\alpha_{2,j}$. For $n = 2$, we verify the validity of formula (36) by finding the corresponding coefficients $\alpha_{2,j}$, which should match the coefficients α_j of §2. The conditions (32) on $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{2j})$ come out to be

$$\begin{cases}\displaystyle\sum_{k=1}^{2j}\sigma_k = m,\\ \displaystyle\sum_{k=1}^{2j}k\sigma_k = 2j,\\ \sigma_k = 0 \quad \forall k \geq 3\end{cases}$$

From the third condition, we have that the only nonzero coefficients are σ_1 and σ_2. From the first and second condition, they satisfy the 2×2 system

$$\begin{cases}\sigma_1 + \sigma_2 = m,\\ \sigma_1 + 2\sigma_2 = 2j,\end{cases}$$

whose unique solution is $\boldsymbol{\sigma} = \big(\sigma_1 = 2(m-j), \sigma_2 = 2j - m\big)$. Using $n = 2$, we have

$$\sum_{\boldsymbol{\sigma}}{}' \prod_{k=1}^{2n-2} \frac{1}{\sigma_k!}\binom{2n}{k+2}^{\sigma_k} = \frac{4^{2(m-j)}}{(2(m-j))!}\cdot\frac{1}{(2j-m)!}.$$

Equation (36) becomes

$$\alpha_{2,j} = \frac{\Gamma(j+1/2)}{\sqrt{\pi}\,6^j}\sum_{m=0}^{2j}\frac{(1/2-j-m)_m}{6^m}\cdot\frac{4^{2(m-j)}}{(2(m-j))!}\cdot\frac{1}{(2j-m)!}.$$

We discard all terms $m < j$; thus we let $m = k + j$ so that k ranges from 0 to j:

$$\begin{aligned}\alpha_{2,j} &= \frac{\Gamma(j+1/2)}{\sqrt{\pi}\,6^{2j}}\sum_{k=0}^{j}\frac{(k+j)!\binom{-j-1/2}{k+j}}{6^k}\cdot\frac{16^k}{(2k)!}\cdot\frac{1}{(j-k)!}\\ &= \frac{\Gamma(j+1/2)}{\sqrt{\pi}\,6^{2j}}\sum_{k=0}^{j}\binom{-j-1/2}{k+j}\binom{k+j}{2k}\left(\frac{8}{3}\right)^k\\ &= \alpha_j \qquad \text{(see (15))}.\end{aligned}$$

3.5. Asymptotic zeros of $\mathcal{F}_3(\mu)$. For $n = 3$, the first four coefficients $\alpha_{3,j}, j = 1, \dots, 4$ are given by

$$\alpha_{3,0} = 1, \quad \alpha_{3,1} = \frac{11}{180}, \quad \alpha_{3,2} = \frac{517}{64800}, \quad \alpha_{3,3} = \frac{-22253}{174960000} \tag{46}$$

(see Appendix C and Table 3). In order to describe the asymptotic approximations of the zeros of $\mathcal{F}_3(\mu)$, we need the following trigonometric expressions:

$$\begin{aligned}&\sin\left(\frac{\pi}{10}\right) = -\cos\left(\frac{3\pi}{5}\right) = \frac{\sqrt{5}-1}{4}, \quad \cos\left(\frac{6\pi}{5}\right) = -\frac{1+\sqrt{5}}{4},\\ &\sin\left(\frac{6\pi}{5}\right) = -\frac{1}{2}\sqrt{\frac{5-\sqrt{5}}{2}}, \quad \cos\left(\frac{\pi}{10}\right) = \sin\left(\frac{3\pi}{5}\right) = \frac{1}{2}\sqrt{\frac{5+\sqrt{5}}{2}},\\ &\sin\left(\frac{9\pi}{5}\right) = -\sin\left(\frac{\pi}{5}\right) = -2\sin\left(\frac{\pi}{10}\right)\cos\left(\frac{\pi}{10}\right).\end{aligned}$$

Using (36), the first-order approximation is

$$\mu_{k,3}^{(0)} = \frac{2\pi}{5}\sqrt{\frac{2}{5+\sqrt{5}}}\left(k - \frac{3}{10}\right), \qquad k \geq 1 \qquad (\mu > 0), \tag{47}$$

and combining (45) and (46), the fourth-order approximation is given in the following corollary.

COROLLARY 3.1. *For $n = 3$ and $k \geq 1$, the approximation of the kth ordered positive zero $\mu_{k,3}$ of $\mathcal{F}_3(\mu) = \int_{-\infty}^{\infty} e^{\mu(6iz - z^6)}dz$ is given by*

$$\mu_{k,3}^{(0)} = \frac{2\pi}{5}\sqrt{\frac{2}{5+\sqrt{5}}}\left(k - \frac{3}{10}\right) + \mathcal{O}\left(\frac{1}{k}\right) \quad \textit{as } k \to +\infty.$$

The fourth-order approximation is

$$\mu_{k,3} = \mathcal{G}_3\left(\mu_{k,3}^{(0)}\right) + \mathcal{O}\left(\frac{1}{k^4}\right) \quad \textit{as } k \to +\infty,$$

$$\mathcal{G}_3(\mu) = \mu + \frac{11}{900\,\mu}\left(1 - \frac{\sqrt{5}-1}{20\mu} - \frac{11}{900}\left(1 - \frac{119}{165}\frac{\sqrt{5}-1}{2}\right)\frac{1}{\mu^2}\right).$$

4. Numerical evaluation of the zeros of $\mathcal{F}_n(\mu)$.

In this section, a numerical method is designed to compute the zeros of the function $\mathcal{F}_n(\mu)$. The purpose of including such an analysis is to judge the accuracy of the asymptotic approximations. This numerical algorithm shows that the high accuracy of the asymptotic predictions is attained for moderatly large zeros, thereby confirming the strength of the asymptotics.

The function $\mathcal{F}_n(\mu)$ is approximated using Simpson's rule and extrapolation, to which we apply the secant method to locate the zeros. The asymptotic approximations of the zeros $\mu_{k,2}$ and $\mu_{k,3}$ of $\mathcal{F}_2(\mu)$ and $\mathcal{F}_3(\mu)$ derived in the previous sections are compared to their numerically calculated values. We also compare these estimates to the zeros of $h_{10}(\mu) = \mathcal{H}_{2,10}(\mu)$ and $\mathcal{H}_{3,10}(\mu)$ (see (19) and (42)), which are computed with the secant method.

4.1. Numerical approximation $\mathcal{F}_n^{m,l}(\mu)$ of $\mathcal{F}_n(\mu)$ by Simpson's rule.

The numerical evaluation of Pearcey-type integrals has been studied by Connor and Curtis in [10]. However, since we consider a special case of Pearcey integrals, we devise a simple algorithm to numerically evaluate $\mathcal{F}_n(\mu)$. Using the alternate expression for $\mathcal{F}_n(\mu)$ (see (6)) given by

$$\mathcal{F}_n(\mu) = 2\int_0^{+\infty} \cos(2n\mu y)\, e^{-\mu y^{2n}}\, dy,$$

we construct the approximation by dividing the range of integration into subintervals over which the integrand does not oscillate. Let

$$x_{-1} = 0, \quad x_k = x_k(\mu) = \frac{(k+1/2)\pi}{2n\mu} \quad \text{for } k \in \mathbb{N},$$

$$g_n(\mu, y) = 2\cos(2n\mu y)\exp(-\mu y^{2n}), \qquad \mathcal{I}_n^k(\mu) = (-1)^k \int_{x_{k-1}}^{x_k} |g_n(\mu, y)|\, dy$$

so that

$$\mathcal{F}_n(\mu) = \sum_{k=0}^{\infty} \mathcal{I}_n^k(\mu) = \overbrace{\sum_{k=0}^{m} \mathcal{I}_n^k(\mu)}^{\mathcal{Q}_n^m(\mu)} + \overbrace{\sum_{k=m+1}^{\infty} \mathcal{I}_n^k(\mu)}^{\mathcal{R}_n^m(\mu)},$$

where

$$\mathcal{Q}_n^m(\mu) = \int_0^{x_m} g_n(\mu, y)\, dy, \qquad \mathcal{R}_n^m(\mu) = \int_{x_m}^{+\infty} g_n(\mu, y)\, dy.$$

We first estimate the remainder $\mathcal{R}_n^m(\mu)$:

$$|\mathcal{R}_n^m(\mu)| \le 2\int_{x_m}^{+\infty} e^{-\mu y^{2n}}\, dy = \frac{\Gamma\left(\frac{1}{2n}, \mu x_m^{2n}\right)}{n\mu^{\frac{1}{2n}}},$$

where $\Gamma(a, x)$ is the incomplete gamma function defined for $\Re a > 0$. Since $\Gamma(a, x) \sim e^{-x} x^{a-1}$ as $x \to +\infty$ (see [1]), we find that

$$|\mathcal{R}_n^m(\mu)| \leq \frac{e^{-\mu x_m^{2n}}}{n\mu x_m^{2n-1}} \qquad \text{for sufficiently large } \mu x_m^{2n}.$$

The endpoint $x_m(\mu) = (m + 1/2)\pi/(2n\mu)$ is chosen in such a way that the contribution from the remainder $\mathcal{R}_n^m(\mu)$ is negligible for a fixed (bounded) μ. If we require that

$$\exp\left(-\mu x_m^{2n}\right) < \varepsilon = 10^{-\kappa}, \quad \kappa \in \mathbb{N}, \tag{48}$$

then it yields a good initial choice for m given by

$$m = m[\kappa; n, \mu_{\max}] = \operatorname{Int}\left[\frac{2n}{\pi}\mu_{\max}^{\frac{2n-1}{2n}}(\kappa \log 10)^{\frac{1}{2n}} - \frac{1}{2}\right] + 1, \tag{49}$$

where Int $[x]$ denotes the integer part of x and $\mu_{\max}$ is a bound on the largest zero we wish compute. It is clear from this analysis that the larger $\mu_{\max}$ is, the larger m will need to be, which is why this algorithm is practical only for small roots $\mu_{k,n}$.

We now approximate $\mathcal{Q}_n^m(\mu)$ by $\mathcal{Q}_n^{m,l}(\mu)$ for large l and moderate m (due to the rapid decay of the integrand),

$$\mathcal{Q}_n^m(\mu) = \sum_{k=0}^{m} \mathcal{I}_n^k(\mu) \approx \mathcal{Q}_n^{m,l}(\mu) = \sum_{k=0}^{m} \mathcal{I}_n^{k,l}(\mu),$$

where each integral $\mathcal{I}_n^k(\mu)$ is approximated by $\mathcal{I}_n^{k,l}(\mu)$ using Simpson's rule: l is the number of gridpoints and the spacing h is defined by

$$h = \frac{\Delta x_k}{l} = \frac{x_{k+1} - x_k}{l} = \frac{\pi}{2n\mu l}.$$

Since the discretization errors for $\mathcal{I}_n^{k,l}$ and $\mathcal{Q}_n^{m,l}$ are given by

$$\mathcal{I}_n^k(\mu) = \mathcal{I}_n^{k,l}(\mu) + \mathcal{O}\left(\frac{1}{l^4}\right), \quad \mathcal{Q}_n^m(\mu) = \mathcal{Q}_n^{m,l}(\mu) + \mathcal{O}\left(\frac{m}{l^4}\right),$$

we can use extrapolation with $\mathcal{Q}_n^{m,l}$ to improve the approximation of $\mathcal{F}_n^{m,l}$ as follows:

$$\mathcal{F}_n^{m,l}(\mu) = \mathcal{Q}_n^{m,2l}(\mu) + \frac{\mathcal{Q}_n^{m,2l}(\mu) - \mathcal{Q}_n^{m,l}(\mu)}{2^4 - 1} \Longrightarrow \mathcal{Q}_n^m(\mu) = \mathcal{F}_n^{m,l}(\mu) + \mathcal{O}\left(\frac{m}{l^5}\right).$$

Thus the final approximation is

$$\mathcal{F}_n(\mu) = \mathcal{F}_n^{m,l}(\mu) + \mathcal{O}\left(\frac{m}{l^5}\right) + \mathcal{O}\left(\frac{e^{-\mu x_m^{2n}}}{\mu x_m^{2n-1}}\right)$$

as $\mu x_m^{2n} \to +\infty$ and $l \to +\infty$. Clearly, the constraint on this algorithm arises from the choice of l since a moderately large value of $m \ll l$ is sufficient to make the remainder $\mathcal{R}_n^m(\mu)$ as small as desired. Moreover, it is difficult to estimate the asymptotic constant in the term $\mathcal{O}(m/l^5)$, which may be large since it involves $\frac{\partial^4}{\partial y^4} g_n(\mu, y)$. Hence the choice for l is made by doubling its value until two successive values of all the zeros $\mu_{k,n} < \mu_{\max}$ agree to 10 significant digits.

4.2. Numerical approximation of the zeros of $\mathcal{F}_n^{m,l}(\mu)$ and $\mathcal{H}_{n,m}(\mu)$ by the secant method. We use the secant method to approximate the zeros of $\mathcal{F}_n^{m,l}(\mu)$ and $\mathcal{H}_{n,m}(\mu)$, which appear in the "Numerical values" column and the "$\mathcal{H}_{n,m}^{-1}(0)$" column of Tables 1 and 2, respectively. From (42), we express $\mathcal{H}_{n,m}(\mu)$ as

$$\mathcal{H}_{n,m}(\mu) = \sum_{j=0}^{m} \alpha_{n,j} \cos\left(\mu\,(2n-1)\cos\left(\frac{\pi}{4n-2}\right) + \pi\,\frac{1-n(1+2j)}{4n-2}\right)\mu^{-j}$$

and let $\mathcal{K}(\mu)$ stand for either $\mathcal{F}_n^{m,l}(\mu)$ or $\mathcal{H}_{n,m}(\mu)$. Then the procedure consists in successively evaluating, for any $k \geq 1$ ($\mu > 0$),

$$\mu_{k,n}^0 = \frac{\pi}{4n-2}\sec\left(\frac{\pi}{4n-2}\right)\left(\frac{n-1}{2n-1} - 1 + 2k\right),$$

$$\mu_{k,n}^1 = \mu_{k,n}^0 - \frac{2\,\delta\mu}{\mathcal{K}(\mu_{k,n}^0+\delta\mu) - \mathcal{K}(\mu_{k,n}^0-\delta\mu)}\cdot\mathcal{K}(\mu_{k,n}^0) \qquad (\delta\mu = 10^{-2}),$$

$$\mu_{k,n}^{j+1} = \mu_{k,n}^j - \frac{\mu_{k,n}^j - \mu_{k,n}^{j-1}}{\mathcal{K}(\mu_{k,n}^j) - \mathcal{K}(\mu_{k,n}^{j-1})}\cdot\mathcal{K}(\mu_{k,n}^j), \qquad j \geq 1,$$

until the convergence of $\mu_{k,n}^j \to \mu_{k,n}$, which is based upon a relative-error test of the form

$$\left|\frac{\mu_{k,n}^{j+1} - \mu_{k,n}^j}{\mu_{k,n}^{j+1}}\right| < \text{tol} \; = 10^{-10}.$$

4.3. $n = 2$. If m is chosen so as to satisfy (48), then a crude initial choice for l is $l = 10^{\kappa/5}$ (typically $\kappa = 12 \Rightarrow l \approx 250$). We take $\mu_{\max} = 11$ to be a bound for the largest zero we wish to compute and take $\kappa = 12$ so that from (49), we find that $m = m[12; 2, 11] = 18$. Starting from $l = 10^{\kappa/5} \approx 250$, we double the value of l until all 10 significant figures in the column "Numerical zeros" of Table 1 do not change. The first such value is $l = 1000$. Note that for $k \leq 5$ ($\mu_{\max} \leq 6$), $m = 12$ is sufficient. One can see in Table 1 that the values computed from the asymptotic approximations are very good. Notice that the first zero $\mu_{1,2}$ is not well approximated by any of the asymptotic predictions since it is less than 1. Beyond the first zero, the asymptotic approximations improve with increasing index k. For $5 \leq k \leq 8$, $\mathcal{H}_{2,10}^{-1}(0)$ agrees with the numerical values up to 10 digits. For $k \geq 5$, $\mu_{k,2}^{(5)}$ and the numerical values agree up to 8 digits. For $k \geq 8$, the numerical and asymptotic values grow apart due to the lack of accuracy of the numerical procedure (see the comment following equation (49)). Note also that for $k \geq 8$, $\mu_{k,2}^{(5)}$ and $\mathcal{H}_{2,10}^{-1}(0)$ agree up to 10 digits ($\mathcal{H}_{2,10}^{-1}(0)$ is computed for the sake of comparison of the asymptotic and numerical estimates). In computing $\mathcal{H}_{2,10}(\mu)$, the 10 coefficients $\alpha_{2,j}, j = 1, \ldots, 10$ are determined using Appendix C. The same is done for $\mathcal{H}_{3,10}(\mu)$ below.

4.4. $n = 3$. Once again, we take $\mu_{\max} = 11$ to be a bound for the largest zero we wish to compute and take $\kappa = 12$ so that from (49), we find that $m = m[12; 3, 11] = 25$. As in the case $n = 2$, starting from $l = 250$, we double l until all 10 significant figures in the column "Numerical zeros" of Table 2 do not change. The first such value is $l = 1000$. For $k \geq 10$, there is 6-digit accuracy when we compare $\mu_{k,3}^{(3)}$ (see (3.1)) and the numerical values; for $10 \leq k \leq 14$, there is also 10-digit accuracy when comparing the numerical values with $\mathcal{H}_{3,10}^{-1}(0)$ and 7-digit accuracy between $\mu_{k,3}^{(3)}$ and $\mathcal{H}_{3,10}^{-1}(0)$ for $k \geq 16$. These results are reported in Table 2.

TABLE 1
Numerical approximation of the zeros $\mu_{k,2}$ *of* $\mathcal{F}_2(\mu) = \int_{-\infty}^{\infty} e^{\mu(4iz-z^4)}dz$.

$\mu_{k,2}$	Numerical zeros	$\mu_{k,2}^{(0)}$ (24)	$\mu_{k,2}^{(5)}$ (Corollary 2.1)	$\mathcal{H}_{2,10}^{-1}(0)$ (42)
$\mu_{1,2}$	0.8221037147	0.8061330508	0.8227392717	0.8240052094
$\mu_{2,2}$	2.0226889660	2.0153326269	2.0226917275	2.0226893916
$\mu_{3,2}$	3.2292915284	3.2245322031	3.2292916648	3.2292915324
$\mu_{4,2}$	4.4372464748	4.4337317792	4.4372464915	4.4372464749
$\mu_{5,2}$	5.6457167459	5.6429313554	5.6457167492	5.6457167459
$\mu_{6,2}$	6.8544374340	6.8521309316	6.8544374349	6.8544374340
$\mu_{7,2}$	8.0632985369	8.0613305077	8.0632985372	8.0632985369
$\mu_{8,2}$	9.2722462225	9.2705300839	9.2722462225	9.2722462225
$\mu_{9,2}$	10.4812510476	10.4797296601	10.4812510479	10.4812510479

TABLE 2
Numerical approximation of the zeros $\mu_{k,3}$ *of* $\mathcal{F}_3(\mu) = \int_{-\infty}^{\infty} e^{\mu(6iz-z^6)}dz$.

$\mu_{k,3}$	Numerical zeros	$\mu_{k,3}^{(0)}$ (47)	$\mu_{k,3}^{(3)}$ (Corollary 3.1)	$\mathcal{H}_{3,10}^{-1}(0)$ (42)
$\mu_{1,3}$	0.5006640277	0.4624572398	0.4845169688	0.46750721075
$\mu_{2,3}$	1.1311965433	1.1231104397	1.1333562062	1.1332534896
$\mu_{3,3}$	1.7905548747	1.7837636396	1.7903635764	1.7903439964
$\mu_{4,3}$	2.4492569634	2.4444168394	2.4492848081	2.4492788273
$\mu_{5,3}$	3.1089250327	3.1050700392	3.1089251416	3.1089227904
$\mu_{6,3}$	3.7689127436	3.7657232391	3.7689140713	3.7689129739
$\mu_{7,3}$	4.4290976016	4.4263764389	4.4290981557	4.4290975781
$\mu_{8,3}$	5.0894021100	5.0870296388	5.0894024443	5.0894021124
$\mu_{9,3}$	5.7497857943	5.7476828386	5.7497859980	5.7497857940
$\mu_{10,3}$	6.4102244359	6.4083360384	6.4102245680	6.4102244359
$\mu_{11,3}$	7.0707027897	7.0689892382	7.0707028789	7.0707027897
$\mu_{12,3}$	7.7312107680	7.7296424381	7.7312108304	7.7312107680
$\mu_{13,3}$	8.3917414319	8.3902956379	8.3917414769	8.3917414319
$\mu_{14,3}$	9.0522898522	9.0509488377	9.0522898854	9.0522898522
$\mu_{15,3}$	9.7128524305	9.7116020376	9.7128524558	9.7128524307
$\mu_{16,3}$	10.373426479	10.372255237	10.373426499	10.373426480

Appendix A. Proof of Lemma 2.1. To prove this lemma, we succesively substitute higher estimates in the equation: Let $\zeta = \mu^{(0)}$; then the asymptotic relation reads

$$\mu = \zeta + \frac{a_1}{\mu}\left(a_2 + \frac{a_3}{\mu} + \frac{a_4}{\mu^2} + \frac{a_5}{\mu^3} + \frac{a_6}{\mu^4}\right) + \mathcal{O}\left(\frac{1}{\mu^6}\right).$$

We have

$$\mu = \zeta + \frac{a_1 a_2}{\zeta} + \mathcal{O}\left(\frac{1}{\zeta^2}\right)$$

followed by

$$\mu = \zeta + \frac{a_1 a_2}{\zeta} + \frac{a_1 a_3}{\zeta^2} + \mathcal{O}\left(\frac{1}{\zeta^3}\right).$$

We now have

$$\begin{aligned}
\frac{a_1 a_2}{\mu} &= \frac{a_1 a_2}{\zeta}\left(1 - \frac{a_1 a_2}{\zeta^2} - \frac{a_1 a_3}{\zeta^3}\right) + \mathcal{O}\left(\frac{1}{\zeta^5}\right),\\
\frac{a_1 a_3}{\mu^2} &= \frac{a_1 a_3}{\zeta^2}\left(1 - 2\frac{a_1 a_2}{\zeta^2}\right) + \mathcal{O}\left(\frac{1}{\zeta^5}\right)
\end{aligned}$$

so that there is a $-(a_1a_2)^2/\zeta^3$ and a $-3a_1^2a_2a_3/\zeta^4$ correction term:

$$\mu = \zeta + \frac{a_1 a_2}{\zeta} + \frac{a_1 a_3}{\zeta^2} + \frac{a_1 a_4 - (a_1 a_2)^2}{\zeta^3} + \frac{a_1 a_5 - 3a_1^2 a_2 a_3}{\zeta^4} + \mathcal{O}\left(\frac{1}{\zeta^5}\right).$$

Finally, we use

$$\begin{aligned}
\frac{a_1 a_2}{\mu} &= \frac{a_1 a_2}{\zeta}\left(1 - \frac{a_1 a_2}{\zeta^2} - \frac{a_1 a_3}{\zeta^3} - \frac{a_1 a_4 - (a_1 a_2)^2}{\zeta^4} + \frac{(a_1 a_2)^2}{\zeta^4}\right) + \mathcal{O}\left(\frac{1}{\zeta^6}\right),\\
\frac{a_1 a_3}{\mu^2} &= \frac{a_1 a_3}{\zeta^2}\left(1 - 2\left(\frac{a_1 a_2}{\zeta^2} + \frac{a_1 a_3}{\zeta^3}\right)\right) + \mathcal{O}\left(\frac{1}{\zeta^6}\right),\\
\frac{a_1 a_4}{\mu^3} &= \frac{a_1 a_4}{\zeta^3}\left(1 - 3\frac{a_1 a_2}{\zeta^2}\right) + \mathcal{O}\left(\frac{1}{\zeta^6}\right)
\end{aligned}$$

so that we must add a $\left(-4a_1^2a_2a_4 + 2a_1^3a_2^3 - 2a_1^2a_3^2\right)/\zeta^5$ correction term. Thus we find

$$\begin{aligned}
\mu = \zeta + \frac{a_1}{\zeta}\Bigg(a_2 &+ \frac{a_3}{\zeta} + \frac{a_4 - a_1 a_2^2}{\zeta^2} + \frac{a_5 - 3a_1 a_2 a_3}{\zeta^3}\\
&+ \frac{a_6 - 2a_1 a_3^2 + 2a_1^2 a_2^3 - 4a_1 a_2 a_4}{\zeta^4}\Bigg) + \mathcal{O}\left(\frac{1}{\zeta^6}\right) \quad \text{as } \zeta = \mu^{(0)} \to +\infty.
\end{aligned}$$

Appendix B. Faà di Bruno's formula. For $\boldsymbol{a} = (a_1, a_2, \dots, a_n) \in \mathbb{N}^n$, following the notation in [1], we define the multinomial coefficients

$$\begin{aligned}
(n; a_1, a_2, \dots, a_n) &= \frac{n!}{a_1! a_2! \cdots a_n!},\\
(n; a_1, a_2, \dots, a_n)' &= \frac{n!}{(1!)^{a_1} a_1! (2!)^{a_2} a_2! \cdots (n!)^{a_n} a_n!}.
\end{aligned}$$

The nth derivative of the composition of two functions is given by Faà di Bruno's formula in [1, §24.1.2] and [13]:

$$\begin{aligned}
\frac{d^n}{dx^n} g(f(x)) &= \sum_{m=0}^{n} g^{(m)}(f(x)) \cdot \sideset{}{'}\sum_{\boldsymbol{a} \in \mathbb{N}^n} (n; a_1, a_2, \dots, a_n)' \cdot \prod_{k=1}^{n} \left\{f^{(k)}(x)\right\}^{a_k}\\
&= \sum_{m=0}^{n} g^{(m)}(f(x)) \cdot \sideset{}{'}\sum_{\boldsymbol{a} \in \mathbb{N}^n} \frac{n!}{a_1! a_2! \cdots a_n!} \cdot \prod_{k=1}^{n} \left\{\frac{f^{(k)}(x)}{k!}\right\}^{a_k}, \qquad (50)
\end{aligned}$$

where the second summation sign $\sum'_{\boldsymbol{a} \in \mathbb{N}^n}$ is taken over all integer n-vectors $\boldsymbol{a} = (a_1, a_2, \dots, a_n) \in \mathbb{N}^n$ such that $\sum_k k a_k = a_1 + 2a_2 + \cdots + na_n = n$ and $|\boldsymbol{a}| = \sum_k a_k = a_1 + a_2 + \cdots + a_n = m$.

TABLE 3
Coefficients $\alpha_{n,j}$ for $n = 2, \ldots, 10$ and $j = 1, \ldots, 5$.

$\alpha_{n,j}$	1	2	3	4	5
2	$\frac{7}{144}$	$\frac{385}{41472}$	$\frac{39655}{17915904}$	$\frac{665665}{10319560704}$	$\frac{-1375739365}{1486016741376}$
3	$\frac{11}{180}$	$\frac{517}{64800}$	$\frac{-22253}{174960000}$	$\frac{-158440051}{125971200000}$	$\frac{-3797666873}{4534963200000}$
4	$\frac{15}{224}$	$\frac{705}{100352}$	$\frac{-23595}{22478848}$	$\frac{-26196885}{20141047808}$	$\frac{-9089431065}{31581162962944}$
5	$\frac{19}{270}$	$\frac{931}{145800}$	$\frac{-111587}{73811250}$	$\frac{-761484451}{637729200000}$	$\frac{6741607873}{172186884000000}$
6	$\frac{115}{1584}$	$\frac{29785}{5018112}$	$\frac{-42479045}{23846068224}$	$\frac{-163420180175}{151088688267264}$	$\frac{56283394450535}{239324482215346176}$
7	$\frac{27}{364}$	$\frac{1485}{264992}$	$\frac{-188595}{96457088}$	$\frac{-138332205}{140441520128}$	$\frac{128908298475}{357844993286144}$
8	$\frac{217}{2880}$	$\frac{88753}{16588800}$	$\frac{-1487341219}{716636160000}$	$\frac{-7471144611931}{8255648563200000}$	$\cdots$
9	$\frac{35}{459}$	$\frac{1085}{210681}$	$\frac{-295295}{136521288}$	$\frac{-1787240455}{2130551220528}$	$\cdots$
10	$\frac{117}{1520}$	$\frac{23049}{4620800}$	$\frac{-78289029}{35118080000}$	$\cdots$	$\cdots$

Appendix C. Mathematica code for the computation of the coefficients $\alpha_{n,j}$. We present a code in Mathematica to compute the rational representation of the coefficients $\alpha_{n,j}$ which appear in Theorem 1.1. They are given by

$$\alpha_{n,j} = \frac{\Gamma(j+1/2)}{\sqrt{\pi}\big(n(2n-1)\big)^j} \cdot \sum_{m=0}^{2j} \left\{ \frac{(1/2-j-m)_m}{\big(n(2n-1)\big)^m} \cdot {\sum_{\sigma}}' \prod_{k=1}^{2n-2} \frac{1}{\sigma_k!} \binom{2n}{k+2}^{\sigma_k} \right\},$$

where the summation $\sum'_\sigma$ is to take place over all possible $\sigma = (\sigma_1, \ldots, \sigma_{2n-2}) \in \mathbb{N}^{2n-2}$ such that $\sigma_1 + \sigma_2 + \cdots + \sigma_{2n-2} = m$, and $\sigma_1 + 2\sigma_2 + \cdots + (2n-2)\sigma_{2n-2} = 2j$.

```
<< DiscreteMath‘Combinatorica‘;

vector[n_, j_, m_] := Module[ {dim,k2}, dim=2n-2; k2=2j;
 If[ j==0, 1, Apply[ Plus, Map[ (Apply [Times, Flatten[
 MapIndexed[((Binomial[2n, #2+2]^#1)/#1!)&, #],1]])&,
 Select[ Flatten[ Map[ Permutations, Select[
 Map[ (Join[ Table[0, {dim-Length[#]}],#])&, Partitions[m]],
 (Length[#] == dim)&]],1], (Range[dim] . # == k2)& ] ] ] ] ];

Alpha[n_,j_] := Gamma[j+1/2] / (Sqrt[Pi] (n(2n-1))^j) *
  Sum[ Pochhammer[1/2-j-m,m] / (n(2n-1))^m * vector[n,j,m], {m,0,2j}];
```

Acknowledgments. The author thanks his thesis advisor, Professor R. Caflisch, for his continuing support and interest. Special thanks go to Professor K. Holczer for his encouragement and hospitality. The author is also grateful to Professors C. K. Chu and P. Roberts for their encouragement, Dr. K. Dempsey for many valuable conversations, David Faraldo for his help in writing the Mathematica code in Appendix C, and the referee for her/his careful review and comments, which helped improve this contribution.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] L. V. Ahlfors, *Complex Analysis*, 3rd ed., McGraw–Hill, New York, 1979.

[3] N. G. Bakhoom, *Asymptotic expansions of the function* $F_k(x) = \int_0^\infty \exp(xu - u^k)du$, Proc. London Math. Soc. (2), 35 (1933), pp. 83–100.

[4] R. P. Boas, *Entire Functions*, Academic Press, New York, 1954.

[5] L. Brillouin, *Sur une méthode de calcul approchée de certaines intégrales, dite méthode de col*, Ann. Ecole Normale (3), 33 (1916), pp. 17–69.

[6] N. G. de Bruijn, *Asymptotic Methods in Analysis*, Dover, New York, 1981.

[7] ———, *The roots of trigonometric integrals*, Duke Math. J., 17 (1950), pp. 197–226.

[8] W. R. Burwell, *Asymptotic expansions of generalized hypergeometric functions*, Proc. London Math. Soc. (2), 22 (1924), pp. 57–72.

[9] M. Christ, *Examples of analytic non-hypoellipticity of* $\overline{\partial_b}$, Comm. Partial Differential Equations, 19 (1994), pp. 911–941.

[10] J. N. L. Connor and P. R. Curtis, *A method for the numerical evaluation of the oscillatory integrals associated with the cuspoid catastrophes: Application to Pearcey's integral and its derivatives*, J. Phys. A, 15 (1982), pp. 1179–1190.

[11] E. T. Copson, *Asymptotic Expansions*, Cambridge University Press, Cambridge, 1965.

[12] ———, *An Introduction to the Theory of Functions of a Complex Variable*, Oxford University Press, Oxford, 1957.

[13] C. F. Faà di Bruno, *Note sur une nouvelle formule du calcul differentiel*, Quart. J. Math., 1 (1855), pp. 359–360.

[14] D. Kaminski, *Asymptotic expansion of the Pearcey integral near the caustic*, SIAM J. Math. Anal., 20 (1989), pp. 987–1005.

[15] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[16] R. B. Paris, *A generalization of Pearcey's integral*, SIAM J. Math. Anal., 25 (1994), pp. 630–645.

[17] ———, *The asymptotic behavior of Pearcey's integral for complex variables*, Proc. Roy. Soc. London Ser. A, 432 (1991), pp. 391–426.

[18] R. B. Paris and A. D. Wood, *Asymptotics of Higher Order Differential Equations*, Pitman Research Notes in Mathematics Series, vol. 129, Longman, London, 1986.

[19] G. Pólya, *Über trigonometrische integrale mit nur reellen nullstellen*, J. Reine Angew. Math., 158 (1927), pp. 6–18.

[20] ———, *On the zeros of an integral function represented by Fourier's integral*, Messenger of Math., 52 (1923), pp. 185–188.

[21] Mathematica, Version 2.2, Wolfram Research, Inc., Champaign, IL, 1993.

[22] D. Senouf, *Complex singularities for Burgers' equation with complex viscosity and asymptotic approximations of the zeros of Fourier integrals*, Ph.D. dissertation, University of California at Los Angeles, Los Angeles, 1994.

[23] ———, *Dynamics and condensation of complex singularities for Burgers' equation*, SIAM J. Math. Anal., submitted.

[24] D. Senouf, R. Caflisch, and N. Ercolani, *Pole dynamics and oscillations for complex Burgers' equation in the small dispersion limit*, Nonlinearity, submitted.

[25] R. Wong, *Asymptotic Approximations of Integrals*, Academic Press, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 1129–1152, July 1996

SETS OF SUPERRESOLUTION AND THE MAXIMUM ENTROPY METHOD ON THE MEAN*

F. GAMBOA[†] AND E. GASSIAT[‡]

Abstract. Consider the problem of recovering a probability measure supported on a compact polish space U, endowed with a probability P, when the available measurements only concern some of its Φ-moments; Φ is here a given k-dimensional continuous real function on U. Provided the *true* moment c lies on the boundary of the convex hull of $\Phi(U)$, we exhibit a support of uniform concentration, that is, a measurable set $R_{c,\delta}(\epsilon)$ (depending on a small positive number δ) such that for any solution μ which satisfies $\| \int \Phi \, d\mu - c\|_2 \leq \epsilon$, we have $\mu(R_{c,\delta}(\epsilon)) \geq 1 - K(\epsilon)$ and $P(R_{c,\delta}(\epsilon)) \leq$ C.$K(\epsilon)$, where $K(\epsilon)$ decreases to 0 with ϵ and C is a constant number. The construction of $R_{c,\delta}(\epsilon)$ and the results are intimately connected with the maximum entropy method on the mean (MEM) developed by Gamboa and Gassiat. This method gives a general framework for superresolution theory via Pythagoras inequalities on families of dissimilarities linked with MEM. In particular cases, we prove that $K(\epsilon)$ is the exact rate of uniform concentration over $R_{c,\delta}(\epsilon)$.

Key words. superresolution, moment problem, dissimilarity, positivity constraint, maximum entropy

AMS subject classifications. 62A99, 52A40

1. Introduction.

Consider the equation

$$y_j = \int_U \Phi_j(x) \, d\mu(x) + \epsilon_j, j = 1, \ldots, k, \tag{1}$$

where U is a compact measurable space, (Φ_j) are known continuous functions, $(\epsilon_j)_{j=1,\ldots,k}$ represents noise, and we want to recover the positive measure μ. A typical example of such a situation is band-limited discrete Fourier measurements, problems in tomography, spectroscopy, astronomy, etc. To recover μ, we first suppose that all we know about the noise $(\epsilon_j)_{j=1,\ldots,k}$ is that it is bounded in *l2-norm* by a known constant ϵ. We now call *reconstruction* any method of recovery $\hat{\mu}(y)$ based on the observations $y = (y_j)_{j=1,\ldots,k}$.

The particularity of such problems is that they are ill posed. In general, the moments $\int_U \Phi_j(x) \, d\mu(x), j = 1, \ldots, k$, are not sufficient to characterize the measure μ. For example, if the Φ_j consist in the first trigonometric functions, the first k Fourier coefficients do not determine μ in general. Years ago, techniques called generically *maximum entropy techniques* (ME) were developed and seemed, in particular situations, to significantly improve the restored solutions, compared with the usual linear methods. Such techniques consist of choosing a reconstruction by minimizing a functional $J(\mu)$, subject to the constraint $\| \int \Phi \, d\mu - y\|_2 < \epsilon$, or to minimize $J(\mu) + \lambda \| \int \Phi \, d\mu - y\|_2$, which is the same by an appropriate choice of λ. Typically,

$$J(\mu) = \int_U \frac{d\mu}{dP} \log \frac{d\mu}{dP} \, dP,$$

where P is a given prior measure. These reconstructions are highly nonlinear in the observations. It has also been well established that the improvement is due to

* Received by the editors June 22, 1992; accepted for publication (in revised form) December 20, 1994.

[†] Laboratoire de Statistiques, Université Paris-Sud, Batiment 425, 91405 Orsay cedex, France; Institut Galilée, Université Paris-Nord, 93430 Villetaneuse, France.

[‡] Laboratoire de Statistiques, Université Paris-Sud, Batiment 425, 91405 Orsay cedex, France and Université d'Evry-Val d'Essonne, 91025 Evry cédex, France.

two major facts: ME takes into account a priori information by imposing implicitly important restrictions on the solution (e.g., positivity), and the *true* measure μ is sparse enough. This improvement was described in terms of *superresolution* by Frieden (see [11]). Let us explain what we mean by *resolution* and *superresolution.* When the observed moments are the Fourier coefficients, all linear translation invariant reconstructions obey the so-called *Rayleigh limit* R at fixed noise level. If μ consists of two spikes spaced by more than R, linear reconstructions may show clearly two spikes; in the opposite, if μ consists of two spikes spaced by less than R, linear reconstructions are unable to resolve clearly the two spikes. This is why R is a *resolution* limit. Now, as is shown in [11] on some examples, ME is able to resolve clearly two spikes spaced by about $R/3$: this is the so-called *superresolution phenomenon.* Though for moments other than Fourier there does not exist such a quantified resolution limit, we extend the notion to any moment problem.

Theoretical results on the subject are very few and recent; see §5 for a survey. The key idea is that, provided the *true* measure μ exhibits some *extremal* conditions (e.g., sparsity), and provided we incorporate in the reconstruction method adequate a priori information (typically the positivity of μ), the resolution of μ (when μ is interpreted as a signal) can be surprisingly high (see, for example, [18], [11], and, for a large bibliography on the subject, [7–9]). Our aim in this work is to understand clearly and quantify precisely what happens in such problems, as well as to give computational solutions for the applications. Let us first give a very simple explanation of the *superresolution phenomenon.* For this, we need to make our notations precise.

- U is a given compact Polish space.
- $\Phi := (\Phi_1, \dots, \Phi_k)$ is a given k-dimensional real function on U. We assume that Φ is continuous and that $\Phi_1 := 1$.

Define $\mathcal{M}_+(U)$ the set of all positive measures on U and $\mathcal{P}(U)$ the set of all probability distributions on U. Let

$$\mathcal{K} := \left\{ c \in \mathbb{R}^k : \exists \mu \in \mathcal{M}_+(U), \int_U \Phi \, d\mu = c \right\},$$

$$\mathcal{K}_1 := \mathcal{K} \cap \{c_1 = 1\}.$$

$\mathrm{ri}(\mathcal{K}_1)$ will denote the relative interior of $\mathcal{K}_1$ (see [24]). Obviously, $\mathcal{K}_1$ is the convex hull of $\Phi(U)$. We shall now consider only probability measures as searched measures. Define

$$\mathcal{S}(c) := \left\{ \mu \in \mathcal{P}(U) : \int_U \Phi(x) \, d\mu(x) = c \right\}$$

and

$$\mathcal{S}(c, \epsilon) := \left\{ \mu \in \mathcal{P}(U) : \left\| c - \int_U \Phi \, d\mu \right\| \leq \epsilon \right\}.$$

In some sense, as ϵ decreases to 0, $\mathcal{S}(c, \epsilon)$ *tends* to $\mathcal{S}(c)$ so that, if $\mathcal{S}(c)$ is reduced to a singleton, the diameter of $\mathcal{S}(c, \epsilon)$ for a weak convergence distance decreases to 0 with ϵ. This is formalized in Theorem 1.

Now, if $\mathcal{S}(c)$ is reduced to a singleton, then $\bar{c} = \int_U \Phi(x) \, d\mu(x)$, where μ is an atomic measure. If ϵ is small enough, any member ν of the set $\mathcal{S}(c, \epsilon)$ will be very close to μ: this is exactly a superresolution phenomenon. Let us now state the precise theorem (which extends the main Theorem of [16]), which is proved in §6. Let d be any distance on $\mathcal{P}(U)$ which is continuous for the weak convergence topology (the Prokhorov distance is an example of such distances).

THEOREM 1. *The following propositions are equivalent:*

$$\lim_{\epsilon\to 0} \sup_{\mu_1,\mu_2\in\mathcal{S}(c,\epsilon)} d(\mu_1,\mu_2) = 0, \tag{2}$$

$$\mathcal{S}(c) \textit{ is reduced to the singleton } \{\sigma_c\}, \tag{3}$$

$$\exists\sigma_c : \lim_{\epsilon\to 0} \sup_{\mu\in\mathcal{S}(c,\epsilon)} d(\mu,\sigma_c) = 0. \tag{4}$$

We will say that c is a determined point if the set $\mathcal{S}(c)$ is reduced to a singleton. Conditions on c to be determined are given in [14]. The result of Theorem 1 is qualitative. It says that superresolution phenomena appear only around determined points and for reconstructions that preserve positivity. We will call it *the strong superresolution phenomenon*, as it describes the whole signal (location and intensity). We see that the a priori information is the positivity of the measure (see the definition of $\mathcal{S}(c)$), and that the extremal condition on μ is that the point $\int_U \Phi(x)\,d\mu(x)$ is determined, which implies a special atomicity of μ. (This explains the necessity that the true measure be sparse enough for superresolution phenomena to appear.) Notice that Theorem 1 shows that *entropy* does not play any particular role in the superresolution phenomenon. Indeed, any reconstruction that satisfies the positivity constraint will exhibit the same superresolution property as soon as the moment constraint is near to be determined. It is surprising that the first formulation of superresolution as a boundary problem in the moment problem appears in [16].

Our precise interest in this paper is to make clear how the resolution of a signal may be improved by imposing positivity. To be more precise, we are not interested in the whole signal but in the location of the signal. Indeed, in many applications, the location of the signal is the significant information which is looked for. In our formulation, we want to see how good approximations of the support of μ may be found and to understand the role of the moment functions $\Phi_1,\dots,\Phi_k$ in this process. In earlier quantitative results (see §5 for references), the space U is discretized, so that some a priori resolution is imposed, and the quantification is essentially made for the intensities. Surprisingly, except for the qualitative paper of Gassiat [16], no earlier works posed the problem in terms of location, that is in terms of support of the measure.

Let us return to our problem. The set of determined points is included in the boundary of $\mathcal{K}_1$. In general the inclusion is strict. If c lies in the boundary of $\mathcal{K}_1$ (and is not necessarily determined), a *weak superresolution phenomenon* holds. Indeed, all the measures of $\mathcal{S}(c)$ are concentrated on a level set

$$\mathcal{C}(c) := \{x \in U : \langle v, \Phi(x)\rangle = 0\}$$

for a particular choice of v (see [19, Th. 1.1, p. 58]). The aim of this work is to develop similar ideas as in Theorem 1 for the subset of $\mathcal{C}(c)$,

$$\cup_{\sigma\in\mathcal{S}(c)}\mathrm{Supp}\sigma := R_c,$$

and to quantify the results. Namely, we propose a construction of *uniform concentration supports*: $R_{c,\delta}(\epsilon)$ such that any probability measure of $\mathcal{S}(c,\epsilon)$ is nearly concentrated on $R_{c,\delta}(\epsilon)$. $R_{c,\delta}(\epsilon)$ will be the approximating set of R_c; it is built as a particular level set of a Φ-polynomial $\langle v^*, \Phi\rangle$. We quantify this concentration property and, in particular cases, we get the exact rate of *superconcentration.* This enables one to locate a signal at a known precision, depending on the error ϵ, even when no *strong* superresolution phenomenon holds. All the results are based on the general reconstruction method called MEM for moment problems developed in [13]. It is interesting to notice that, as soon as one knows this reconstruction, all theorems proved here involve only elementary computations. In other words, as soon as the structure of the problem is clearly seen, the superconcentration properties do not involve sophisticated techniques; they are all "hidden" in the MEM method. In particular cases, direct calculations on appropriate level sets lead to the exact superconcentration rate without the use of the MEM dissimilarities. However, we would like to emphasize several points to show the interest of the technique developed therein. First, MEM leads to a general framework for superresolution theory in a different context: see [14] for L_1-superresolution results; see also [15] for a continuous superresolution result. This general framework also has a computational advantage: the computation of the parameter v^* of the level set, via MEM, involves strictly convex optimization, and the superresolution rate $K(\epsilon)$ is a by-product of the computation. Moreover, the idea of using level sets as approximations of the support arose directly from MEM. This lead several authors, inspired by the first version of this work, to develop the pure level set point of view [10], [20]. In general, the MEM framework exploits general probabilistic ideas: Laplace methods and developments of Cramer transforms on the boundary of their domain, which have interest by themselves.

To our knowledge, all these results are new and give new ideas to understand superresolution. Notice also that, though numerical applications of the results are easy to obtain using convex programming, explicit formulations are untractable, even on very simple examples. We will develop numerical examples and other ideas in a forthcoming paper.

The paper is organized as follows. In the next section, we give our notations and recall the basic results concerning MEM. In §3, we state and illustrate our superresolution results. In §4, we study a family of dissimilarities which gave the key idea to prove the superconcentration theorems. In §5, we give a survey on earlier results and study connections with ours. All technical proofs are collected in §6.

2. The MEM basic results.

Our work is based on results about the MEM. Let us recall what will be useful for this paper. Let P be a given reference probability distribution on U such that for any vector v in $\mathbb{R}^k, P(\langle v, \Phi\rangle = 0) = 0$. Let F be a given probability measure on $[0, +\infty[$ whose convex hull of the support equals $[0, +\infty[$. Define ψ by

$$\psi(t) := \log \int_{[0,+\infty[} \exp(tx)\, dF(x), t \in \mathbb{R},$$

and suppose that $\psi(t)$ is finite if and only if $t \in]-\infty, \alpha[, \alpha < +\infty$.

Define

$$\gamma(s) := \sup_{t\in\mathbb{R}}(st - \psi(t)).$$

For any measurable function f on $\mathbb{R}_+$, let

$$\Gamma(f) := \int_U \gamma[f(x)]\, dP(x).$$

For any μ in $\mathcal{M}_+(U)$, let

$$I(\mu) := \Gamma\left(\frac{d\mu}{dP}\right) + \alpha\left(\mu - \frac{d\mu}{dP}P\right)(U).$$

For any c and v in $\mathbb{R}^k$ and any nonnegative ϵ, let

$$H^\epsilon(v,c) := \int_U \psi(\langle v, \Phi(x)\rangle)\, dP(x) - \langle v, c\rangle + \epsilon\|v\|_{k-1},$$

where $\|v\|_{k-1}$ is the Euclidian norm of the vector $(v_2, \dots, v_k)$. Define also

$$h^*(c,\epsilon) := -\inf_{v\in\mathbb{R}^k} H^\epsilon(v,c).$$

In [13], we proved the following results. (For related works in optimization, see [1], [2].)

THEOREM 2. *Let $c \in \mathbb{R}^k$ be such that the Euclidian ball centered at c with diameter ϵ intersects* $\mathrm{ri}(\mathcal{K}_1)$. *Let v^ϵ be the unique minimizer of $H^\epsilon(v,c)$. There exists a positive measure σ^ϵ such that*

$$G^{\mathrm{MEM},\epsilon} := \psi'(\langle v^\epsilon, \Phi\rangle)P + \sigma^\epsilon \in \mathcal{S}(c,\epsilon), \tag{5}$$

Supp $\sigma^\epsilon \subset \{x; \langle v^\epsilon, \Phi(x)\rangle = \alpha\}$. *Moreover,*

$$h^*(c,\epsilon) = -H^\epsilon(v^\epsilon, c) = \inf_{f: fP\in\mathcal{S}(c,\epsilon)} \Gamma(f) = \min_{\mu\in\mathcal{S}(c,\epsilon)} I(\mu) = I(G^{MEM,\epsilon}).$$

THEOREM 3. *Assume that F weighs $\{0\}$; then*

$$\begin{aligned} h^*(c,0) &< -\log F(\{0\}) + \alpha \Leftrightarrow c \in \mathrm{ri}(\mathcal{K}_1),\\ h^*(c,0) &= -\log F(\{0\}) + \alpha \Leftrightarrow c \in \mathrm{bd}(\mathcal{K}_1),\\ h^*(c,0) &= +\infty \Leftrightarrow c \notin \mathcal{K}_1. \end{aligned}$$

The idea of such characterization of $\mathrm{bd}(\mathcal{K}_1)$ arose from the probabilistic interpretation of $G^{\mathrm{MEM},\epsilon}$ and its relation with the large deviations theory. Lewis has shown later in [21] that these theorems hold for a wider class of convex functions ψ.

3. Superresolution results.

3.1. Main theorem. Let $\overline{c}$ be a point in the boundary of $\mathcal{K}_1$. The MEM solution in $\mathcal{S}(\overline{c},\epsilon)$ is $G^{\mathrm{MEM},\epsilon}$ as defined in (5). Observe the following.

- First, $\forall x \in U, \langle v^\epsilon, \Phi(x)\rangle \le \alpha$ as follows from the definition of v^ϵ. Indeed, if not, as Φ is continuous, $H^\epsilon(v^\epsilon, \overline{c})$ is infinite, which contradicts Theorem 2.
- Second, any accumulation point of the sequence $G^{\mathrm{MEM},\epsilon}$ as ϵ decreases to 0 is a singular measure (with respect to P) $\sigma_{\overline{c}}$ such that

$$\begin{aligned} &\forall x \in \mathrm{Supp}\ \sigma_{\overline{c}}, \quad \lim_{\epsilon\to 0}\langle v^\epsilon, \Phi(x)\rangle = \alpha,\\ &\forall x \notin \mathrm{Supp}\ \sigma_{\overline{c}}, \quad \lim_{\epsilon\to 0}\langle v^\epsilon, \Phi(x)\rangle = -\infty. \end{aligned}$$

Intuitively, we may now approximate $R_{\overline{c}}$ by a set of points x such that $\langle v^\epsilon, \Phi(x)\rangle$ is near α. Precisely, define

$$g_\epsilon(x) := \alpha - \langle v^\epsilon, \Phi(x)\rangle.$$

For any positive real δ, define

$$R_{\overline{c},\delta}(\epsilon) := \{x \in U : g_\epsilon(x) \leq \delta\}.$$

$R_{\overline{c},\delta}(\epsilon)$ will be the approximating set of $R_{\overline{c}}$. We have the following theorem.

THEOREM 4. *Assume that F weighs* 0 *and set*

$$F := F(\{0\})\,\delta_0 + (1 - F(\{0\}))F_1,$$

$$\forall t \in]-\infty, \alpha[, \quad \psi_1(t) := \log \int_{]0,+\infty[} \exp(tx)\,dF_1(x).$$

Let $a := \frac{1-F(\{0\})}{F(\{0\})}$. *We have*

$$P(R_{\overline{c},\delta}(\epsilon)) \log(1 + a e^{\psi_1(\alpha-\delta)}) + \delta \sup_{\mu \in \mathcal{S}(\overline{c},\epsilon)} \mu(R^c_{\overline{c},\delta}(\epsilon)) \leq K(\epsilon),$$

where $K(\epsilon) := -\log F(\{0\}) + \alpha - h^*(\overline{c},\epsilon)$ *decreases to* 0 *with* ϵ, *and* $R^c_{\overline{c},\delta}(\epsilon)$ *denotes the complementary set of* $R_{\overline{c},\delta}(\epsilon)$ *in* U.

Observe that $R_{\overline{c},\delta}(\epsilon)$ depends on ϵ, and that δ is a parameter which has to be chosen: it might be chosen as a function of ϵ. $R_{\overline{c},\delta}(\epsilon)$ also depends on the choice of F. We will discuss this problem in a forthcoming paper.

This theorem quantifies the concentration of the supports of any solution of the moment problem near the boundary. Indeed, it says that, keeping δ fixed, you can define a measurable set $R_{\overline{c},\delta}(\epsilon)$ (which is a level set associated with an MEM solution, reachable using convex programming) such that, simultaneously, its P-measure is small, and any solution of the moment problem with perturbation ϵ is concentrated on this set except for a small part, *small* being quantified by the speed $K(\epsilon)$ in both assertions.

Example. Choose for F_1, the exponential distribution with parameter 1,

$$\psi_1(t) = \log \frac{1}{1-t}$$

so that $\alpha = 1$. Choose also $\delta := 1, F(\{0\}) = e^{-1}$, and the theorem says

$$P(R_{\overline{c},1}(\epsilon)) + \sup_{\mu \in \mathcal{S}(\overline{c},\epsilon)} \mu(R^c_{\overline{c},1}(\epsilon)) \leq K(\epsilon)$$

so that simultaneously

$$P(R_{\overline{c},1}(\epsilon)) \leq K(\epsilon)$$

and for any ϵ-*solution* μ (that is, $\mu \in \mathcal{S}(\overline{c},\epsilon)$) of the boundary point

$$\mu(R_{\overline{c},1}(\epsilon)) \geq 1 - K(\epsilon)$$

(μ is $K(\epsilon)$-*concentrated* on $R_{\overline{c},1}(\epsilon)$).

Remarks. For particular moment problems (especially T-systems), the approximating set may be chosen as the level set $\mathcal{C}(c)$ defined in §1, and it can be proved to lead to the optimal rate. This was developed later in [10] and [20]. However, the advantages of our method are that, on the basis of the observations, the level set is computed through strictly convex minimization programming together with the rate $K(\epsilon)$ (see §3.4), and our method gives a general framework for other superresolution problems using the dissimilarities defined in §4 as explained in §3.4.5; see also [14] and [15].

Proof of Theorem 4. We first prove the decreasing property of $K(\epsilon)$. To see this, observe that $h^*(c,0)$ is a convex function with domain $\mathcal{K}$, so that it is continuous on $\mathrm{ri}(\mathcal{K}_1)$. Moreover, $h^*(c,0)$ converges to $h^*(\bar{c},0)$ when c tends to $\bar{c}$ and when c stays in the relative interior of $\mathcal{K}_1$. We easily have (using, for instance, subgradient arguments)

$$h^*(\bar{c},\epsilon) = \inf_{c\in\mathcal{K}_1, \|c-\bar{c}\|\le\epsilon} h^*(c,0) = \min_{c\in\mathcal{K}_1, \|c-\bar{c}\|\le\epsilon} h(c,0) = h^*(c_\epsilon,0), \tag{6}$$

where $\|c_\epsilon - \bar{c}\| \le \epsilon$. Since h^* is strictly convex and constant on $\mathrm{bd}\mathcal{K}_1, c_\epsilon$ is in the relative interior of $\mathcal{K}_1$ for all positive ϵ, and c_ϵ tends to $\bar{c}$ when ϵ tends to 0; therefore,

$$\lim_{\epsilon\to 0} h^*(\bar{c},\epsilon) = h^*(\bar{c},0) = -\log F(\{0\}) + \alpha$$

and it follows that

$$\lim_{\epsilon\to 0} K(\epsilon) = 0.$$

Now let σ be any element of $\mathcal{S}(\bar{c},\epsilon)$. Let

$$I := \int_U \psi(\langle v^\epsilon, \Phi(x)\rangle)\, dP(x) - \log F(\{0\}) + \int_U (\alpha - \langle v^\epsilon, \Phi(x)\rangle)\, d\sigma(x).$$

First, we have

$$I = \int_U \psi(\langle v^\epsilon, \Phi(x)\rangle)\, dP(x) - \left\langle v^\epsilon, \int_U \Phi(x)\, d\sigma(x) \right\rangle + \alpha - \log F(\{0\}).$$

Using equation (6), we have

$$\begin{aligned} h(\bar{c},\epsilon) &= -\int_U \psi(\langle v^\epsilon, \Phi(x)\rangle)\, dP(x) + \langle v^\epsilon, c_\epsilon\rangle \\ &= -\int_U \psi(\langle v^\epsilon, \Phi(x)\rangle)\, dP(x) + \langle v^\epsilon, \bar{c}\rangle - \epsilon\|v^\epsilon\| \end{aligned}$$

and $\| \int_U \langle v^\epsilon, \Phi(x)\rangle\, d\sigma(x) - \bar{c}\| \le \epsilon$, so that using the Schwarz inequality, we have

$$-\int_U \psi(\langle v^\epsilon, \Phi(x)\rangle)\, dP(x) + \int_U \langle v^\epsilon, \Phi(x)\rangle\, d\sigma(x) \ge h^*(\bar{c},\epsilon)$$

and

$$I \le \alpha - \log F(\{0\}) - h^*(\bar{c},\epsilon). \tag{7}$$

Now, obviously,

$$\int_U g_\epsilon\, d\sigma \ge \delta\sigma(R^c_{\bar{c},\delta}(\epsilon)). \tag{8}$$

Also,

$$\int_U \psi(\langle v^\epsilon, \Phi\rangle)\, dP - \log F(\{0\}) = \int_U \log(1 + ae^{\psi_1(\alpha - g_\epsilon)})\, dP.$$

Using the fact that ψ_1 is increasing,

$$\int_U \psi(\langle v^\epsilon, \Phi\rangle)\,dP - \log F(\{0\}) \geq \int_{R_{\overline{c},\delta(\epsilon)}} \log(1 + ae^{\psi_1(\alpha-\delta)})\,dP,$$

so that

$$\int_U \psi(\langle v^\epsilon, \Phi\rangle)\,dP - \log F(\{0\}) \geq P(R_{\overline{c},\delta}(\epsilon))\log(1 + ae^{\psi_1(\alpha-\delta)}). \tag{9}$$

Adding (8) and (9), we have

$$P(R_{\overline{c},\delta}(\epsilon))\log(1 + ae^{\psi_1(\alpha-\delta)}) + \delta\sigma(R^c_{\overline{c},\delta}(\epsilon)) \leq K(\epsilon).$$

Taking the supremum over $\mathcal{S}(\overline{c},\epsilon)$, we get the inequality of the theorem.

Remark. The idea of considering the functional I to get inequality (7) has in itself no intuitive evidence. The idea from which it comes, which is hidden in (7), is analogous to a *Pythagoras inequality* which holds for a dissimilarity linked with the MEM construction. This will be explained in §4.

3.2. Lower bound for the superconcentration. In §§3.2 and 3.3, we will assume that U is a compact set of $\mathbb{R}^d$. Theorem 4 gives an upper bound for the superconcentration property over $R_{\overline{c},\delta}(\epsilon)$. Here we will give a lower bound for it which holds for any choice of F.

THEOREM 5. *Suppose that P is the normalized Lebesgue measure on $U, \Phi \in [C^1(U_0)]^k$, where $U \subset \operatorname{int}(U_0)$. Then there exists a positive constant C (depending only on d) such that for small enough ϵ,*

$$\sup_{\mu\in\mathcal{S}(\overline{c},\epsilon)} \mu(R^c_{\overline{c},\delta}(\epsilon)) \geq C\epsilon K(\epsilon)^{-1/d}[\log(1 + ae^{\psi_1(\alpha-\delta)})]^{1/d}.$$

3.3. Exact rate of superconcentration. Under additional assumptions, we are able to give the exact rate of superconcentration over $R_{\overline{c},\delta}(\epsilon)$. We restrict F to be member of the following family of distributions. For all $\beta > 0, F_\beta$ denotes the Poissonized distribution of the $\gamma(\beta,1)$ distribution. That is, F_β is the distribution of the random variable Z defined by

$$Z := \sum_{i=0}^{N} Y_i,$$

where $Y_0 := 0, (Y_i)$ is a sequence of independent random variables with common distribution $\gamma(\beta,1)$ and N is an independent random variable with Poisson distribution of parameter 1. Then, the log-Laplace transform of F_β is

$$\begin{aligned}\psi_\beta(\tau) &= \frac{1}{(1-\tau)^\beta} - 1 \quad \text{if } \tau < 1\\ &= +\infty \quad \text{if } \tau \geq 1.\end{aligned}$$

It is obvious that $\alpha = 1, F_\beta(\{0\}) = e^{-1}$, and ψ_β satisfies the assumptions required for the MEM construction. K_β will denote the associated K function.

We will now make an *integrability* assumption. For this purpose, let us introduce the following definition.

DEFINITION 1. *Let $\overline{c} \in \mathrm{bd}(\mathcal{K}_1)$. The critical exponent $\beta_0(\overline{c})$ is the upper bound of the set*

$$\left\{\beta \geq 0, \exists v \in \mathbb{R}^k : \forall x \in U, \langle v, \Phi(x)\rangle \geq 0; \langle v, \overline{c}\rangle = 0, \int_U \frac{dP(x)}{\langle v, \Phi(x)\rangle^\beta} < \infty\right\}.$$

Under the assumption that $\beta_0(\overline{c})$ is positive, for small ϵ and for $\beta < \beta_0(\overline{c})$, we give a precise description of $K_\beta(\epsilon)$ and the associated vector v^ϵ. Precisely, define

$$\begin{aligned}
u^\epsilon &:= (1 - v_1^\epsilon, -v_2^\epsilon, \dots, -v_k^\epsilon),\\
w^\epsilon &:= (\langle u^\epsilon, \overline{c}\rangle, u_2^\epsilon, \dots, u_k^\epsilon),\\
d(\epsilon) &:= \frac{w^\epsilon}{\|w^\epsilon\|} = (d_1(\epsilon), \tilde{d}(\epsilon)),\\
\tilde{\Phi}(x) &:= (\Phi_j(x) - \overline{c}_j)_{j=2,\dots,k},\\
\xi_{\beta,0}(d) &:= \int_U \frac{dP(x)}{(\max(0, \langle \tilde{d}, \tilde{\phi}(x)\rangle))^\beta}, \quad d = (d_1, \tilde{d}) \in \mathbb{R}^k.
\end{aligned}$$

PROPOSITION 1. *Suppose $\beta_0(\overline{c}) > 0$. Let $\beta < \beta_0(\overline{c})$. Then the following hold.*

- *$d(\epsilon)$ converges to $(0, \tilde{d}(0))$, where $\tilde{d}(0)$ is the unique minimizer of $\xi_{\beta,0}(d)$ on the intersection between the unit sphere and the hyperplane $d_1 = 0$.*
- *$K_\beta(\epsilon) \sim C_\beta \epsilon^{\beta/\beta+1}, \epsilon \to 0$ for a positive constant C_β.*

We will now assume that $\Phi \in [C^2(U)]^k$ and

- (I2) $\exists x_m \in \mathrm{int}(U), \exists \sigma \in \mathcal{S}(\overline{c})$ such that

$$\sigma(\{x_m\}) > 0,$$
$$\langle \tilde{d}(0), D^2\tilde{\Phi}(x_m)\rangle \neq 0.$$

We are now able to give the following theorem.

THEOREM 6. *Suppose that $\Phi \in [C^2(U)]^k, \beta_0(\overline{c}) > 0, \beta < \beta_0(\overline{c})$, and (I2) holds. Then, for ϵ sufficiently small, there exist two constants C_1, C_2, which depend only on β, δ, and $\overline{c}$, such that*

$$C_1 \epsilon^{\frac{\beta}{\beta+1}} \leq \sup_{\mu \in \mathcal{S}(\overline{c},\epsilon)} \mu(R^c_{\overline{c},\delta}(\epsilon)) \leq C_2 \epsilon^{\frac{\beta}{\beta+1}}.$$

In the specified situation, Theorem 6 says that the rate of uniform concentration over $R_{\overline{c},\delta}(\epsilon)$ is exactly $\epsilon^{\beta/\beta+1}$.

In specific situations, when Φ is a T-system (see also §3.4.4), it is possible to choose a different approximating support set: a level set of the supporting hyperplane at the boundary point $\overline{c}$, and simpler calculation leads to the optimal rate $\epsilon^{1/3}$. However, this particular level set is difficult to compute, even for known $\overline{c}$. However, this other point of view was developed afterwards in [10].

We show in §6 (Lemma 5) that assumption (I2) holds whenever the following geometrical assumption holds.

- (I3) $\exists x_m \in \mathrm{int}(U), \exists \sigma \in \mathcal{S}(\overline{c})$, and there exists an orthogonal basis $(b_2, \dots, b_k)$ of $\mathbb{R}^{k-1}$ such that

$$\sigma(\{x_m\}) > 0,$$
$$\langle b_j, \tilde{\Phi}\rangle \geq 0,$$
$$\langle b_j, D^2\tilde{\Phi}(x_m)\rangle \neq 0.$$

3.4. Remarks and applications.

3.4.1. Superconcentration for positive reconstructions. Consider the problem of estimating the support of μ using observation (1). With the same arguments as for Theorem 4, it is easy to get

$$P(R_{y,\delta}(\epsilon))\log(1+ae^{\psi_1(\alpha-\delta)})+\delta\sup_{\sigma\in\mathcal{S}(y,\epsilon)}\sigma(R^c_{y,\delta}(\epsilon))\leq-\log F(\{0\})+\alpha-h^*(y,\epsilon).$$

Here, as ϵ tends to 0, y tends to $\bar{c}=\int_U\Phi\,d\mu$, so that as soon as $\bar{c}$ lies on $\mathrm{bd}(\mathcal{K}_1)$, $-\log F(\{0\})+\alpha-h^*(y,\epsilon)$ tends to 0, and the inequality is again a superconcentration inequality, in which only the knowledge of the observations y and of the noise level ϵ are involved.

As $\mu\in\mathcal{S}(y,\epsilon)$, we get

$$P(R_{y,\delta}(\epsilon))\log(1+ae^{\psi_1(\alpha-\delta)})+\delta.\mu(R^c_{y,\delta}(\epsilon))\leq-\log F(\{0\})+\alpha-h^*(y,\epsilon),$$

and $R_{y,\delta}$ is a good estimator for the support of μ on the basis of y.

3.4.2. Superconcentration for a point near the boundary. All the results we have given have been written for measures of $\mathcal{S}(\bar{c},\epsilon)$, where $\bar{c}$ is a boundary point of $\mathcal{K}_1$. Let $c^*:=\bar{c}+\epsilon\,\Delta c$ with $\Delta c:=(0,\tilde{\Delta}c)$, $\|\tilde{\Delta}c\|_{k-1}=1$. We assume that c^* is an inner point of $\mathcal{K}_1$. When ϵ is small, we can ask, "What are the superconcentration properties of $\mathcal{S}(c^*)$?"

Since $\mathcal{S}(c^*)\subset\mathcal{S}(\bar{c},\epsilon)$, all the results given before stay true, changing $\mathcal{S}(\bar{c},\epsilon)$ by $\mathcal{S}(c^*)$ everywhere. However, using the MEM technique, we can give more local results. (Indeed, the previous ones do not depend on the given direction $\tilde{\Delta}c$.) Let

$$G_\beta^{\mathrm{MEM}}=\psi'_\beta(\langle v^*,\Phi(x)\rangle)P+\sigma$$

be the element of $\mathcal{S}(c^*)$ selected using MEM procedure with prior F_β (see §2). Let

$$\tilde{R}_{c^*,\delta,\beta}:=\{x\in U,1-\langle v^*,\Phi(x)\rangle\leq\delta\}.$$

Then, under some regularity assumptions (as in Theorem 6), using the Pythagoras identity (10) (see §4) and following the same ideas as before, we have

$$C_1\epsilon^{\frac{\beta}{\beta+1}}\leq\sup_{\mu\in\mathcal{S}(c^*)}\mu(\tilde{R}_{c^*,\delta,\beta})\leq C_2\epsilon^{\frac{\beta}{\beta+1}},$$

where the constants C_1 and C_2 depend only on β,δ, and c^*. For a given boundary point $\bar{c}$ and for almost all directions Δc, the constant of the superconcentration inequalities are significantly better when we use this local formulation.

3.4.3. Asymptotic behavior of $R_{\bar{c},\delta,\beta}(\epsilon)$.

LEMMA 1. *Let*

$$\mathcal{V}(\bar{c},\beta):=\left\{x\in U,-\sum_{j=2}^{k}d_j(0)\bar{c}_j+\langle\tilde{d}(0),\Phi(x)\rangle=0\right\}.$$

Suppose that $\beta_0(\bar{c})>0$. Then, for all $\beta<\beta_0(\bar{c})$,

$$\cup_{\sigma\in\mathcal{S}(\bar{c})}\mathrm{supp}\sigma\subset\varliminf_{\epsilon\longrightarrow0^+}R_{\bar{c},\delta,\beta}(\epsilon)\subset\varlimsup_{\epsilon\longrightarrow0^+}R_{\bar{c},\delta,\beta}(\epsilon)\subset\mathcal{V}(\bar{c},\beta).$$

Proof. If $x \in \overline{\lim}_{\epsilon \to 0^+} R_{\overline{c},\delta,\beta}(\epsilon)$, then there exits a decreasing sequence (ϵ_n) converging to 0 such that

$$\forall n \in N, \quad \langle v^{\epsilon_n}, \Phi(x) \rangle \leq \delta,$$

which is equivalent to

$$\forall n \in N, \quad d_1(\epsilon_n) + \langle \tilde{d}(\epsilon_n), \tilde{\Phi}(x) \rangle \leq \frac{\delta}{\|v^{\epsilon_n}\|}.$$

Taking the limit as n goes to infinity, we get $\langle \tilde{d}(0), \tilde{\Phi}(x) \rangle = 0$.

Let $x \in \cup_{\sigma \in \mathcal{S}(\overline{c})} \mathrm{supp} \sigma$. Then there exists $\sigma \in \mathcal{S}(\overline{c})$ with $\sigma(\{x\}) = p > 0$ (see [19, Thm. 3.1, p. 71). Therefore, by Theorem 4, $\sigma(R^c_{\overline{c},\delta,\beta}(\epsilon)) \leq \frac{K(\epsilon)}{\delta}$. Now, as soon as $\frac{K(\epsilon)}{\delta} < p$ we have $\sigma(R^c_{\overline{c},\delta,\beta}(\epsilon)) < p$ and $x \in R_{\overline{c},\delta,\beta(\epsilon)}$. □

3.4.4. Examples. Explicit calculations are untractable even for simple examples, but numerical applications may be directly developed. For this we refer to a forthcoming paper. However, let us give ideas about typical situations.

Analytic functions. Suppose that $U = [0,1], P$ is the uniform probability, and Φ is analytic. Then for all v in $\mathbb{R}^k$, the set

$$\mathcal{O}(v) = \{x \in [0,1], \langle v, \Phi(x) \rangle = 0\}$$

has a finite number of elements. Therefore, since we can develop the function $\langle v, \Phi(x) \rangle$ near each point x^* of $\mathcal{O}(v)$ in a convergent entire series, for all boundary point $\overline{c}$ of $\mathcal{K}_1$ we have $\beta_0(\overline{c}) > 0$. Then Proposition 4 always holds (see §6). We do not know if the assertion $\beta_0(\overline{c}) > 0$ is still true for an analytic system Φ on $U \subset \mathbb{R}^d$. In the multidimensional framework, even when the components of Φ are polynomials, the description of $\mathcal{O}(v)$ and its connected components is not easier [22], [17].

T-systems. Let $U = [0,1], \Phi(x)$ be a T-system $(\forall v \in \mathbb{R}^k, \sharp\mathcal{O}(v) \leq k-1)$, and P the uniform probability. In this case, any boundary point is determinate (see for example [19, Thm. 4.1, p. 78]). More precisely, a point $\overline{c}$ is determined if and only if (see [19, Thm. 4.1, p.78])

$$\overline{c} = \sum_{j=1}^{q} p_j \Phi(x_j), \qquad \sum_{j=1}^{q} p_j = 1, x_j \neq x_{j'}, j \neq j',$$

with $i(\overline{c}) = \sum_{j=1}^{q} \eta(x_j) \leq k-1$, where, when Φ is not periodic,

$$\begin{aligned} \eta(x) &:= 1 \quad \text{if } x \in]0,1[\\ &:= 2, \quad x = 0 \text{ or } x = 1 \end{aligned}$$

and $\eta(x) := 1$ for all x when Φ is periodic. It is possible to prove that, for any determined point $\overline{c}, \beta_0(\overline{c}) = \frac{1}{2}$, and that assumption (I3) and Theorem 6 hold as soon as one of the $x_j \neq 0, 1$ (see Lemma 5).

As an example, suppose now that $\Phi(x) = (1, \ldots, x^{k-1})$, that $i(\overline{c}) = k-1$, and $\forall j = 1 \ldots q, x_j \neq 0, 1$. Then there exists a unique nonnegative polynomial which vanishes at points $x_1, \ldots, x_q$. Therefore, we have

$$\xi_{\beta,0}[(0, \tilde{d}(0))] = \int_0^1 \frac{dx}{\prod_{j=1}^{q} (x - x_j)^{2\beta}} \left(\sum_{j=0}^{q} b_j^2 \right)^{-\frac{\beta}{2}},$$

where

$$b_j = \sum_{l+l'=j} (-1)^j \sum_{1 \le j_1 < j_2 < \cdots < j_{q-l} \le q} x_{j_1}, \dots, x_{j_{q-l}} \sum_{1 \le j_1 < j_2 < \cdots < j_{q-l'} \le q} x_{j_1}, \dots, x_{j_{q-l'}}.$$

Through this evaluation of $\xi_{\beta,0}[(0, \tilde{d}(0))]$, we see that the constant that appears in the superconcentration inequality is smaller whenever $i(\overline{c}) < k - 1$. Indeed, in this case the set of nonnegative polynomials which vanish at points $x_1, \dots, x_q$ is not reduced to a singleton.

3.4.5. Further applications. The methodology introduced in this paper to investigate superresolution phenomena can be used in other situations as soon as we dispose of the MEM reconstruction technique. This is the case in the following examples.

Markov moment problem. In this inverse problem, we replace the set of all positive measures by the set of all nonnegative measurable functions bounded by $L > 1$ (when the searched functions are probability densities (see [19, Chap. VII] for more details on Markov moment problems). The MEM method could be applied exactly as we did previously. A determined point $\overline{c}$ admits for some $d \in \mathbb{R}^k$ the representation

$$\overline{c} = L \int_U \Phi(x) 1_{\{x \in U, \langle d, \Phi(x) \rangle \ge 0\}} \, dP(x).$$

The Dirac measures are replaced here by indicator functions, so that stronger distances may be used to describe superresolution. Using the ψ-dissimilarity (see §4) for an appropriate prior we can evaluate for small ϵ

$$\sup_{f \in \mathcal{C}_{0,L}, \int \Phi f \, dP \in B(\overline{c}, \epsilon)} \| f - 1_{\{x \in U, \langle d, \Phi(x) \rangle \ge 0\}} \|_1$$

or

$$\sup_{f, g \in \mathcal{C}_{0,L}, \int \Phi f \, dP \in B(\overline{c}, \epsilon), \int \Phi f \, dP \in B(\overline{c}, \epsilon)} \| f - g \|_1,$$

where $\|f\|_1 := \int |f| \, dP$ and

$$\mathcal{C}_{0,L} := \{ f \text{ measurable}, 0 \le f \le L, P \text{ almost surely} \}.$$

We remark that the compactness assumption $\Phi_1(x) = 1$ can be removed here. Indeed, the condition $f \in \mathcal{C}_{0,L}$ is a very strong compactness assumption. We also remark that, for the Markov moment problem, the evaluation of the superresolution rate function $K(\epsilon)$ is easier. Indeed, the Lagrangian function $H^\epsilon(v, c)$ (see §2) is defined here on all of $\mathbb{R}^k$ [5].

These ideas are developed in [14] and later in [20].

Multidimensional moment problems. We will discuss here the case of superresolution rate for multidimensional moment problems. That is, keeping the same notations as in the previous sections, Φ is now a matrix-valued function from U to $\mathbb{R}^k \times \mathbb{R}^m$ with

$$\phi_{1,j}(x) := 1, \quad j = 1, \dots, m.$$

The inverse problem consists of recovering a vector-valued positive measure $\mu = (\mu_j)_{j=1,\dots,m}$ (we can also require more than the simple positivity of the components, typically $(d\mu)/(dP) \in \mathcal{C}$, where $\mathcal{C}$ is a given convex set) satisfying the linear constraint

$$\int_U \Phi(x)\, d\mu(x) = c.$$

The MEM method for this reconstruction problem has been developed in [12]. Therefore, with extra work, we can hope to find the same kind of superconcentration results.

4. A dissimilarity linked with MEM construction. In this section, we introduce and study a dissimilarity between a nonnegative function and a positive measure. This dissimilarity is linked with the MEM construction. We first begin with its definition.

DEFINITION 2. *Let G be in $\mathcal{M}_+(U)$; g will denote the Radon–Nikodým derivative of G with respect to P. Let*

$$\mathcal{M}_+^{\psi,G} := \{\textit{nonnegative measurable function } h \textit{ on } U, \psi'^{-1}(h) \in L^1(G - gP)\}.$$

Let h be in $\mathcal{M}_+^{\psi,G}$. The generalized ψ-dissimilarity $\mathcal{D}_\psi(G,h)$ between G and h is defined by

$$\mathcal{D}_\psi(G,h) := D_\psi(g,h) + \alpha(G - gP)(U) - \int_U \psi'^{-1}[h(x)]\, d(G - gP),$$

where the ψ-dissimilarity $D_\psi(g,h)$ between g and h is defined by

$$\begin{aligned} D_\psi(g,h) := & \int_U g(x)(\psi'^{-1}[g(x)] - \psi'^{-1}[h(x)])\, dP(x) \\ & - \int_U \psi[\psi'^{-1}\{g(x)\}]\, dP(x) + \int_U \psi[\psi'^{-1}\{h(x)\}]\, dP(x) \end{aligned}$$

when all the functions involved are integrable, and $+\infty$ otherwise.

The terminology of dissimilarity is justified by the following proposition (proved in §6).

PROPOSITION 2. *We have that*

$$\forall G \in \mathcal{M}_+(U), h \in \mathcal{M}_+^{\psi,G}, \mathcal{D}_\psi(G,h) \geq 0$$

$\mathcal{D}_\psi(G,h)$ vanishes if and only if

$$\frac{dG}{dP} = h \ (\textit{P almost surely}) \textit{ and } \operatorname{Supp}\left(G - \frac{dG}{dP}P\right) \subset \{x \in U, h = \infty\}.$$

We can also give a dual definition of the dissimilarities in the case of densities.

PROPOSITION 3. *For any measurable nonnegative functions g and h,*

$$\begin{aligned} D_\psi(g,h) = & \sup_{l \in C(U)} \left\{ \int_U l(x) g(x)\, dP(x) - \int_U \psi[l(x) - \psi'^{-1}\{h(x)\}]\, dP(x) \right\} \\ & + \int_U \psi[\psi'^{-1}(h(x))]\, dP(x). \end{aligned}$$

The following Pythagoras theorem gives a relation between the ψ-dissimilarity and the MEM method. It has been the key tool for finding our superconcentration result (7). Indeed, using Lemma 2, (11) leads to (7). It is a consequence of the same relation proved by Csiszar for the Kullbak divergence in [3]. (See the proof of Theorem 7.)

THEOREM 7. *Let c be an inner point of $\mathcal{K}_1$; then*

$$\begin{aligned}\forall G \in \mathcal{S}(c), \quad \mathcal{D}_\psi(G, \psi'(0)) &= I(G) \\ &= \mathcal{D}_\psi(G, \psi'(\langle v^*, \Phi\rangle)) + \mathcal{D}_\psi(G^{\mathrm{MEM}}, \psi'(0)) \\ &= \mathcal{D}_\psi(G, \psi'(\langle v^*, \Phi\rangle)) + I(G^{\mathrm{MEM}}).\end{aligned} \tag{10}$$

Let c be a point of $\mathbb{R}^k$ such that $\mathcal{S}(c, \epsilon)$ is nonvoid; then

$$\begin{aligned}&\forall G \in \mathcal{S}(c, \epsilon), \\ &\mathcal{D}_\psi(G, \psi'(0)) = I(G) \\ &\qquad \geq \mathcal{D}_\psi(G, \psi'(\langle v^\epsilon, \Phi\rangle)) + \mathcal{D}_\psi(G^{\mathrm{MEM},\epsilon}, \psi'(0)) \\ &\qquad = \mathcal{D}_\psi(G, \psi'(\langle v^\epsilon, \Phi\rangle)) + I(G^{\mathrm{MEM},\epsilon}).\end{aligned} \tag{11}$$

LEMMA 2. *Let $G \in \mathcal{M}_+(U)$; then*

$$I(G) \leq -\log F(\{0\}) + \alpha.$$

Lemma 2 is proved in [13, Eqs. (19) and (20)].

5. Earlier results on superresolution.

- The first qualitative result on superresolution appears in Gassiat [16]. It gives the qualitative results of Theorems 1 and 4 in the restrictive situation where $\Phi_1, \ldots, \Phi_k$ form a T-system. The proof relies on properties of such a system, which is clearly not necessary.
- The first quantitative superresolution result appears in Donoho et al. [9]. In this work, U is the *discretized* torus and the functions Φ_j are the sine and cosine first functions (so that the c_j are the first Fourier coefficients). The authors give an evaluation of

$$\sup_{\mu_1, \mu_2 \in \mathcal{S}(c, \epsilon)} \sum_{x \in U} |\mu_1(x) - \mu_2(x)|.$$

The techniques rely heavily on Fourier analysis.

- In Donoho [7], the author gives a superresolution inequality in a very different context. The space U is again discretized: U is the lattice ΔZ. The moment constraints are no longer of finite number: the observation is the Fourier transform of the measure, $\hat{\mu}(\omega), |\omega| \leq \Omega$, where Ω is less than the Rayleigh constant $\frac{2\pi}{\Delta}$. In this work, the measure μ is supposed to be *sparse* enough, not necessarily positive, and the set of reconstructions is supposed to be the set of measures ν (possibly signed), which possess the same *sparsity property*, and satisfy

$$||\hat{\nu} - \hat{\mu}||_{2,[-\Omega,\Omega]} \leq \epsilon.$$

As U is not compact, uniqueness is necessary here, but it is not known if uniqueness is sufficient for superresolution to hold (compare with Theorem 1 in the compact situation). The superresolution inequality given shows also how sparsity quantifies the capacity to superresolve (see [7] for exact result). The techniques rely on hard Fourier analysis and connections with interpolation problems.

- Following similar ideas, in [8] Donoho and Gassiat give superresolution inequalities when μ is supposed to be positive and sparse enough, but the set of reconstructions is now the set of all positive measures ν such that

$$\|\hat{\nu} - \hat{\mu}\|_{2,[-\Omega,\Omega]} \leq \epsilon.$$

An application is given for the case of the torus.

In our work, the ϕ_j do not need to be trigonometric functions for the result to hold. The dependence on the ϕ_j appears in the computation of the function $K(\epsilon)$, through $H^{\epsilon}(v, c)$. In our opinion, these results are very complementary to ours.

6. Proofs.

6.1. Proof of Theorem 1.

- (2) $\Rightarrow$ (3). Indeed,

$$\forall \epsilon > 0, \qquad \sup_{\mu_1,\mu_2 \in \mathcal{S}(c,\epsilon)} d(\mu_1, \mu_2) \geq \sup_{\mu_1,\mu_2 \in \mathcal{S}(c)} d(\mu_1, \mu_2).$$

- (3) $\Rightarrow$ (4). As U is compact, $\mathcal{S}(c, \epsilon)$ is a compact set of positive measures. Now, $d(., \sigma_c)$ is continuous (for the weak topology on $\mathcal{M}_+(U)$) so that

$$\forall \epsilon > 0, \quad \exists \mu_\epsilon \in \mathcal{S}(c, \epsilon) : \sup_{\mu \in \mathcal{S}(c,\epsilon)} d(\mu, \sigma_c) = d(\mu_\epsilon, \sigma_c).$$

Again using the compactness of $\mathcal{S}(c, \epsilon)$, the sequence (μ_ϵ) has at least one accumulation point μ^* when ϵ decreases to 0. Using the continuity of the application $\mu \longrightarrow \int \Phi \, d\mu, \mu^*$ lies in $\mathcal{S}(c)$. Therefore, $\mu^* = \sigma_c$ and $\lim_{\epsilon \to 0^+} d(\mu_\epsilon, \sigma_c) = 0$.

- (4) $\Rightarrow$ (2) is obvious using triangular inequality.

6.2. Proof of Theorem 5.

Using assumption (I1), we can write

$$\bar{c} = \sum_{i=1}^{q} p_i \Phi(x_i),$$

where $p_i > 0, \sum_{i=1}^{q} p_i = 1$.

Define

$$c(\epsilon) := \sum_{i=2}^{q} p_i \Phi(x_i) + [p_1 - h(\epsilon)]\Phi(x_1) + h(\epsilon)\Phi(x_1 + l(\epsilon)u),$$

where u is a point in $\mathbb{R}^d$ of modulus 1, h and l are positive and *small* functions of ϵ such that $h(\epsilon) < p_1$, and $x_1 + l(\epsilon)u$ lies in U. (This is possible with an appropriate choice of u even if $x_1 \in \mathrm{bd}(U)$.) Using the Taylor expansion until order 1, we get

$$\|c(\epsilon) - \bar{c}\| \leq h(\epsilon)l(\epsilon)\|\Phi'\|_{\infty,U_0}.$$

Therefore, set

$$h(\epsilon)l(\epsilon)\|\Phi'\|_{\infty,U_0} := \epsilon, \tag{12}$$

so that

$$\sigma_\epsilon = \sum_{i=2}^{q} p_i\, \delta_{x_i} + (p_1 - h(\epsilon))\, \delta_{x_1} + h(\epsilon)\, \delta_{x_1 + l(\epsilon)u}$$

is in $\mathcal{S}(\overline{c}, \epsilon)$.

As soon as $x_1 + l(\epsilon)u$ is in $R^{c}_{\overline{c},\delta}(\epsilon)$, we will have

$$\sup_{\mu \in \mathcal{S}(\overline{c},\epsilon)} \mu(R^{c}_{\overline{c},\delta}(\epsilon)) \geq h(\epsilon).$$

Using Theorem 4 we get

$$P(R_{\overline{c},\delta}(\epsilon)) < C_1 K(\epsilon)$$

with $C_1 := 1/\log(1 + ae^{\psi_1(\alpha - \delta)})$. For small enough $\epsilon, x_1 \in R_{\overline{c},\delta}(\epsilon)$; see Lemma 1. Now

$$P(B(x_1, r)) = r^d \mathrm{Ct}(d, P),$$

so that for $r = \left(\frac{C_1 K(\epsilon)}{\mathrm{Ct}(d,P)}\right)^{1/d}$ there exists u of modulus one and $l(\epsilon) \leq r$ such that $x_1 + l(\epsilon)u$ is in $R^{c}_{\overline{c},\delta}(\epsilon)$ and we may take

$$l(\epsilon) := [C_1 K(\epsilon)/\mathrm{Ct}(d, P)]^{1/d}.$$

Evaluate $h(\epsilon)$ using (12), and the theorem is straightforward. □

6.3. Proof of Theorem 6. When $\beta_0(\overline{c}) > 0$, we will study the behavior near 0^+ of $K_\beta(.)$ for all $\beta < \beta_0(\overline{c})$.

LEMMA 3.

$$\forall \epsilon \geq 0, \quad \forall \beta > 0, \quad K_\beta(\epsilon) = \beta^{\frac{1}{\beta+1}} \left(1 + \frac{1}{\beta}\right) \cdot \left[\inf_{d \in \mathbb{R}^k, \|d\|=1} \int_U \left(\frac{d_1 + \epsilon\|\tilde{d}\|}{\max(0, d_1 + \langle \tilde{d}\tilde{\Phi}(x)\rangle)}\right)^{\beta} dP(x)\right]^{\frac{1}{\beta+1}}$$

where $\tilde{\Phi}(x) := (\Phi_j(x) - \overline{c}_j)_{j=2,\dots,k}, d := (d_1, \tilde{d})$.

Proof. By definition,

$$\begin{aligned} K_\beta(\epsilon) &= -\log F_\beta(\{0\}) + 1 - h^*(\overline{c}, \epsilon) \\ &= 1 + 1 + \inf_{v \in \mathbb{R}^k} \left[\int_U \frac{dP(x)}{(1 - \min(1, \langle v, \Phi(x)\rangle))^\beta} - 1 - \langle v, \overline{c}\rangle + \epsilon\|v\|_{k-1}\right]. \end{aligned}$$

Making the variable changes $u := (1 - v_1, -v_2, \dots, -v_k)$ and $w := (\langle u, \overline{c}\rangle, u_2, \dots, u_k)$, we find

$$\begin{aligned} K_\beta(\epsilon) &= \inf_{w \in \mathbb{R}^k} \left[\int_U \frac{dP(x)}{(\max(0, w_1 + \langle \tilde{w}, \tilde{\Phi}(x)\rangle))^\beta} + w_1 + \epsilon\|\tilde{w}\|\right] \\ &= \inf_{d \in \mathbb{R}^k, \|d\|=1} \inf_{r > 0} \left[\frac{1}{r^\beta} \int_U \frac{dP(x)}{(\max(0, d_1 + \langle \tilde{d}, \tilde{\Phi}(x)\rangle))^\beta} + r(d_1 + \epsilon\|\tilde{d}\|)\right], \end{aligned}$$

where $d := (d_1, \tilde{d}), w := (w_1, \tilde{w})$.

Now, by an obvious optimization calculus on r, we get

$$K_\beta(\epsilon) = \inf_{d\in\mathbb{R}^k,\|d\|=1}\left[\beta^{\frac{1}{\beta+1}}\left(1+\frac{1}{\beta}\right)\left(\int_U\left[\frac{d_1+\epsilon\|\tilde{d}\|}{\max(0,d_1+\langle\tilde{d},\tilde{\Phi}(x)\rangle)}\right]^\beta dP(x)\right)^{\frac{1}{\beta+1}}\right]$$

$$=\beta^{\frac{1}{\beta+1}}\left(1+\frac{1}{\beta}\right)\left[\inf_{d\in\mathbb{R}^k,\|d\|=1}\int_U\left(\frac{d_1+\epsilon\|\tilde{d}\|}{\max(0,d_1+\langle\tilde{d},\tilde{\Phi}(x)\rangle)}\right)^\beta dP(x)\right]^{\frac{1}{\beta+1}}. \qquad \square$$

LEMMA 4. *Suppose that the critical exponent $\beta_0(\bar{c})$ is positive. Let*

$$\xi_{\beta,\epsilon}(d) := \int_U\left(\frac{\frac{d_1}{\epsilon}+\|\tilde{d}\|}{\max(0,d_1+\langle\tilde{d},\tilde{\Phi}(x)\rangle)}\right)^\beta dP(x), \epsilon>0, 0<\beta<\beta_0(\bar{c}), d\in\mathbb{R}^k,$$

$$\xi_{\beta,0}(d) = \int_U\frac{dP(x)}{\max(0,\langle\tilde{d},\tilde{\Phi}(x)\rangle)^\beta}, 0<\beta<\beta_0(\bar{c}), d\in\mathbb{R}^k.$$

Then we have the following.

- *For all $\epsilon>0$, $\xi_{\beta,\epsilon}(.)$ has a unique minimizer $d(\epsilon)$ on the unit sphere.*
- *$\xi_{\beta,0}(.)$ has a unique minimizer $(0,\tilde{d}(0))$ on the intersection between the unit sphere and the hyperplane $d_1=0$.*

Moreover, $d_1(\epsilon)$ is a positive function of ϵ and

$$\lim_{\epsilon\to0^+} d(\epsilon) = (0,\tilde{d}(0)),$$

$$\lim_{\epsilon\to0^+}\xi_{\beta,\epsilon}[d(\epsilon)] = \xi_{\beta,0}[(0,\tilde{d}(0))].$$

An application of the previous lemma leads to the following proposition.

PROPOSITION 4. *Suppose that the critical exponent $\beta_0(\bar{c})$ is positive. Then, for ϵ sufficiently small and $0<\beta<\beta_0(\bar{c})$,*

$$K_\beta(\epsilon)\sim C_\beta\epsilon^{\frac{\beta}{\beta+1}},$$

where

$$C_\beta := (\xi_{\beta,0}[(0,\tilde{d}(0))])^{\frac{1}{\beta+1}}\beta^{\frac{1}{\beta+1}}\left(1+\frac{1}{\beta}\right).$$

Lemma 4 and Proposition 4 prove Proposition 1.

Proof of Lemma 4.

- For $\epsilon>0$, uniqueness of $d(\epsilon)$ follows from uniqueness of the Lagrange multiplier v^ϵ (see §2).
- For $\epsilon=0$, let $v\in\mathbb{R}^k$ such that

$$\forall x\in U,\quad \langle v,\Phi(x)\rangle\geq0; \langle v,\bar{c}\rangle=0, \int_U\frac{dP(x)}{\langle v,\Phi(x)\rangle^\beta}<\infty.$$

Therefore,

$$v_1=-\sum_{j=2}^k v_j\bar{c}_j,$$

so that

$$\|\tilde{v}\|^{-\beta}\int_U \frac{dP(x)}{\max(0,\langle\tilde{v},\tilde{\Phi}(x)\rangle)^{\beta}}<\infty,$$

and $\inf_{d\in\mathbb{R}^k,\|d\|=1}\xi_{\beta,0}(d)<\infty$.

Now, on the convex set $\{(0,\tilde{d})\in\mathbb{R}^k,\|\tilde{d}\|\le 1\}$, the function $\xi_{\beta,0}(.)$ is strictly convex, so its minimum value is reached at a unique point $(0,\tilde{d}(0))$. Since

$$\xi_{\beta,0}(0,\tilde{d}(0))=\xi_{\beta,0}\left(\left(0,\frac{\tilde{d}(0)}{\|\tilde{d}(0)\|}\right)\right)\|\tilde{d}(0)\|^{-\beta},$$

we necessarily have $\|\tilde{d}(0)\|=1$.

We shall now prove the last points of the lemma. Let $\sigma\in\mathcal{S}(\overline{c})$; we have

$$\forall\epsilon>0,\quad\int_U(d_1(\epsilon)+\langle\tilde{d}(\epsilon),\tilde{\Phi}(x)\rangle)\,d\sigma(x)=d_1(\epsilon).$$

Therefore, since $d_1(\epsilon)+\langle\tilde{d}(\epsilon),\tilde{\Phi}(x)\rangle$ is nonnegative on $U, d_1(\epsilon)\ge 0$.

Now, since $K(\epsilon)=[\epsilon^{\beta}\xi_{\beta,\epsilon}\{d(\epsilon)\}]^{1/(\beta+1)}(1+\frac{1}{\beta})\beta^{1/(\beta+1)}$ we have that

$$\lim_{\epsilon\to0^+}d_1(\epsilon)=0,$$
$$\lim_{\epsilon\to0^+}\|\tilde{d}(\epsilon)\|=1.$$

Therefore,

$$\int_U\left(\frac{\|\tilde{d}\|}{\max(0,d_1+\langle\tilde{d},\tilde{\Phi}(x)\rangle)}\right)^{\beta}dP(x)\le\xi_{\beta,\epsilon}[d(\epsilon)].$$

Let d^* be any accumulation point of the function $d(\epsilon)$ as ϵ decreases to 0. We have, using Fatou's lemma,

$$\underline{\lim}_{\epsilon\to0^+}\xi_{\beta,\epsilon}[d(\epsilon)]\ge\int_U\frac{1}{(\max(0,\langle\tilde{d}^*,\tilde{\Phi}(x)\rangle))^{\beta}}dP(x).$$

On the other hand, $\forall\,d\in\mathbb{R}^k,\|d\|=1,\xi_{\beta,\epsilon}(d(\epsilon))\le\xi_{\beta,\epsilon}(d)$.

Taking $d_1:=0$ and minimizing the right term of the previous inequality, we find

$$\xi_{\beta,\epsilon}[d(\epsilon)]\le\xi_{\beta,0}[(0,\tilde{d}(0))].$$

Since $\xi_{\beta,0}(0,.)$ is strictly convex we have $\tilde{d}^*=\tilde{d}(0)$ and

$$\lim_{\epsilon\to0^+}\xi_{\beta,\epsilon}(d(\epsilon))=\xi_{\beta,0}(0,\tilde{d}(0)).$$

We remark that we also have $\lim_{\epsilon\to0^+}\frac{d_1(\epsilon)}{\epsilon}=0$.

Indeed, let $d_1^*=\overline{\lim}_{\epsilon\to0^+}\frac{d_1(\epsilon)}{\epsilon}$; then, by Fatou's lemma,

$$(d_1^*+1)^{\beta}\xi_{\beta,0}(0,\tilde{d}(0))\le\xi_{\beta,0}(0,\tilde{d}(0)).$$

And $d_1^*=0$. □

LEMMA 5. *Let $\bar{c} \in \mathrm{bd}(\mathcal{K}_1)$. Suppose that $U \subset \mathbb{R}^d$ and that $\Phi \in C^2(U)$. Suppose that the critical exponent $\beta_0(\bar{c})$ is positive. Let*

$$\bar{c} := \sum_{i=1}^{q} p_i \Phi(x_i),$$

where $p_i > 0$ and $\sum_{i=1}^{q} p_i = 1$.

Suppose there exists at least one x_m of the $x_j, j = 1, \ldots, r$, such that there exists an orthogonal basis $(b_2, \ldots, b_k)$ of $\mathbb{R}^{k-1}$ satisfying

- $\langle b_j, \tilde{\Phi}(x) \rangle \geq 0, \forall x \in U, j = 2, \ldots, k$;
- $x_m \in \mathrm{int}(U)$;
- $\langle b_j, D^2\tilde{\Phi}(x_m) \rangle \neq 0, j = 2, \ldots, k$. *Then,* $\langle \tilde{d}(0), D^2\tilde{\Phi}(x_m) \rangle \neq 0$.

Proof. Obviously, for $(0, \tilde{d}), (0, \tilde{e}) \in \mathbb{R}^k$,

$$\xi_{\beta,0}[(0,\tilde{d}) + (0,\tilde{e})] \leq \mathrm{Min}(\xi_{\beta,0}[(0,\tilde{d})], \xi_{\beta,0}[(0,\tilde{e})]),$$

so that

$$\xi_{\beta,0}\left(\frac{(0,\tilde{d}) + (0,\tilde{e})}{\|(0,\tilde{d}) + (0,\tilde{e})\|}\right) \leq \|(0,\tilde{d}) + (0,\tilde{e})\|^{\beta}\mathrm{Min}(\xi_{\beta,0}[(0,\tilde{d})], \xi_{\beta,0}[(0,\tilde{e})]).$$

Suppose that

$$\langle \tilde{d}(0), D^2\tilde{\Phi}(x_m) \rangle = 0$$

and let

$$\tilde{d}(0) = \sum_{j=2}^{k} \tilde{d}_j^b(0) b_j$$

be the decomposition of $\tilde{d}(0)$ on the basis $(b_2, \ldots, b_k)$.

Now, because $\langle b_j, \tilde{\Phi}(x) \rangle, j = 2, \ldots, k$, is nonnegative on U and vanishes at the point x_m, we have $\langle b_j, D^2\tilde{\Phi}(x_m) \rangle > 0, j = 2, \ldots, k$. Therefore, one of the components $(\tilde{d}_j^b(0))$ is negative. So we can assume that $\tilde{d}_2^b(0) < 0$. Thus, for $0 < \lambda < -2\tilde{d}_2^b(0)$, we have

$$\|\tilde{d}(0) + \lambda \tilde{d}_2^b(0) b_2\| < 1$$

so that

$$\xi_{\beta,0}\left(\frac{(0, \tilde{d}(0) + \lambda \tilde{d}_2^b(0) b_2)}{\|(0, \tilde{d}(0) + \lambda \tilde{d}_2^b(0) b_2)\|}\right) < \xi_{\beta,0}[(0, \tilde{d}(0))],$$

which contradicts the fact that $(0, \tilde{d}(0))$ achieves the minimum of $\xi_{\beta,0}[(0,.)]$ over $\{\|\tilde{d}\| = 1\}$. We may conclude that $\langle \tilde{d}(0), D^2\tilde{\Phi}(x_m) \rangle \neq 0$. □

Proof of Theorem 6. Without loss of generality, we may assume that $\langle \tilde{d}(0), D^2\tilde{\Phi}(x_1) \rangle \neq 0$. Define

$$c(\epsilon) := \sum_{i=2}^{r} p_i \Phi(x_i) + (1 - h(\epsilon)) p_1 \Phi(x_1) + \frac{h(\epsilon)}{2} p_1 \Phi(x_1 + l(\epsilon)u) + \Phi(x_1 - l(\epsilon)u),$$

with $u \in \mathbb{R}^d$, $\|u\| = 1$, and we choose $h(\epsilon) < 1$. We can write

$$c(\epsilon) = \int_U \Phi(x)\, d\sigma_\epsilon(x),$$

where

$$\sigma_\epsilon := \sum_{i=2}^{r} p_i\, \delta_{x_i} + p_1(1 - h(\epsilon))\delta_{x_1} + p_1 \frac{h(\epsilon)}{2}[\delta_{x_1 + l(\epsilon)u} + \delta_{x_1 - l(\epsilon)u}]$$

is a probability distribution, with support in U as soon as $l(\epsilon)$ is small enough. We have

$$c(\epsilon) = \bar{c} + \frac{h(\epsilon)}{2} p_1\{\Phi(x_1 + l(\epsilon)u) - \Phi(x_1) + \Phi(x_1 - l(\epsilon)u) - \Phi(x_1)\}.$$

Using a Taylor extension of order 2, we get

$$\|c(\epsilon) - \bar{c}\|_2 \le \frac{h(\epsilon) l^2(\epsilon)}{2} \|D^2\Phi\|_\infty. \tag{13}$$

We are now going to prove that either $x_1 + l(\epsilon)u$ or $x_1 - l(\epsilon)u$ is in $R^c_{\bar{c},\delta}(\epsilon)$ for a special value of $l(\epsilon)$ and a special choice of u. Using an exact Taylor extension of order 2, we have

$$\begin{aligned} &\alpha - \langle v_\epsilon, \Phi(x_1 + l(\epsilon)u)\rangle + \alpha - \langle v_\epsilon, \Phi(x_1 - l(\epsilon)u)\rangle \\ &\quad = 2(\alpha - \langle v_\epsilon, \Phi(x_1)\rangle) - \frac{l^2(\epsilon)}{2}\langle v_\epsilon, D^2\Phi(x_1 + \theta^* l(\epsilon)u)(u,u)\rangle \\ &\qquad - \frac{l^2(\epsilon)}{2}\langle v_\epsilon, D^2\Phi(x_1 - \theta^{**} l(\epsilon)u)(u,u)\rangle \end{aligned}$$

with

$$0 \le |\theta^*| \le 1, \qquad 0 \le |\theta^{**}| \le 1.$$

Now, using the convergence of $d(\epsilon)$ and the continuity of $D^2\tilde{\Phi}$, we have

$$\lim_{\epsilon \to 0^+} \alpha - \langle v_\epsilon, \Phi(x_1)\rangle = 0$$

and

$$-\frac{l^2(\epsilon)}{2}\langle v_\epsilon, D^2\Phi(x_1 + \theta^* l(\epsilon)u)(u,u)\rangle \sim \frac{l^2(\epsilon)}{2}\|v_\epsilon\|_{k-1}\langle \tilde{d}(0), D^2\tilde{\Phi}(x_1)(u,u)\rangle.$$

We can also write a similar formula for the other term.

Now choose u such that

$$\langle \tilde{d}(0), D^2\tilde{\Phi}(x_1)(u,u)\rangle = \eta > 0$$

and

$$l^2(\epsilon) = \frac{5\delta}{\eta}\frac{1}{\|v_\epsilon\|_{k-1}}.$$

Observe that we have $\lim_{\epsilon \to 0^+} l(\epsilon) = 0$.

For small ϵ,

$$\alpha - \langle v_\epsilon, \Phi(x_1)\rangle > -\delta$$

and

$$\frac{l^2(\epsilon)}{2}\langle v_\epsilon, D^2\Phi(x_1+\theta^* l(\epsilon)u)(u,u)\rangle = \frac{5\delta}{2\eta}\langle d(\epsilon), D^2\tilde{\Phi}(x_1+\theta^* l(\epsilon)u)(u,u)\rangle > 2\delta.$$

We can also write a similar formula for the other term.

Therefore, for small ϵ,

$$\alpha - \langle v_\epsilon, \Phi(x_1 + l(\epsilon)u)\rangle + \alpha - \langle v_\epsilon, \Phi(x_1 - l(\epsilon)u)\rangle > 2\delta,$$

so that either

$$\alpha - \langle v_\epsilon, \Phi(x_1 + l(\epsilon)u)\rangle > \delta$$

or

$$\alpha - \langle v_\epsilon, \Phi(x_1 - l(\epsilon)u)\rangle > \delta. \tag{14}$$

Now define

$$h(\epsilon) := \frac{2\epsilon}{\|D^2\Phi\|_\infty l^2(\epsilon)}.$$

Notice that $h(\epsilon) \to 0$, as $l^2(\epsilon) \sim \mathrm{Ct}\frac{\epsilon}{K(\epsilon)}$.

In view of (13), $\sigma_\epsilon \in \mathcal{S}(\bar{c}, \epsilon)$, and in view of (14),

$$\sigma_\epsilon(R^c_{\bar{c},\delta}(\epsilon)) \geq p_1\frac{h(\epsilon)}{2}.$$

Now,

$$p_1\frac{h(\epsilon)}{2} = \frac{p_1}{\|D^2\Phi\|_\infty}\frac{\eta}{5\delta}\epsilon\|v_\epsilon\|_{k-1} = C_1K(\epsilon).$$

And the theorem is proved. □

6.4. Proofs of Propositions 2 and 3. Proposition 2 follows from the study of the function

$$\theta(\tau) := \tau\{\psi'^{-1}[\tau] - \psi'^{-1}[\tau']\} - \psi[\psi'^{-1}\{\tau\}] + \psi[\psi'^{-1}\{\tau'\}].$$

For a fixed $\tau', \theta(\tau)$ is nonnegative and vanishes if and only if $\tau = \tau'$. Proposition 3 follows from the function change $k := l - \psi'^{-1}(h)$ and variational calculus (see [12] or [25]). □

6.5. Proof of Theorem 7. Let $G \in \mathcal{S}(c)$ with $I(G) < \infty$; we have

$$\begin{aligned}
\mathcal{D}_\psi(G, \psi'(\langle v^*, \Phi\rangle)) &= D_\psi\left(\frac{dG}{dP}, \psi'(\langle v^*, \Phi\rangle)\right) + \alpha\left[\left(G - \frac{dG}{dP}P\right)(U)\right] \\
&\quad - \int_U \langle v^*, \Phi(x)\rangle \left(G - \frac{dG}{dP}P\right)(dx) \\
&= \Gamma\left(\frac{dG}{dP}\right) - \int_U \langle v^*, \Phi(x)\rangle \frac{dG}{dP}\, dP(x) \\
&\quad + \int_U \psi(\langle v^*, \Phi(x)\rangle)\, dP(x) + \alpha\left[\left(G - \frac{dG}{dP}P\right)(U)\right] \\
&\quad - \int_U \langle v^*, \Phi(x)\rangle \left(G - \frac{dG}{dP}P\right)(dx) \\
&= \Gamma\left(\frac{dG}{dP}\right) - \langle v^*, c\rangle \\
&\quad + \int_U \psi(\langle v^*, \Phi(x)\rangle)\, dP(x) + \alpha\left[\left(G - \frac{dG}{dP}P\right)(U)\right] \\
&= \mathcal{D}_\psi(G, \psi'(0)) - \mathcal{D}_\psi(G^{\mathrm{MEM}}, \psi'(0)),
\end{aligned}$$

so (10) is proved.

Let $gP \in \mathcal{S}(c, \epsilon)$ with $g > 0$ and $g \in C(U)$. Let (x_i) be a sequence of points of U such that the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ converges weakly to P. Let χ_n^g be the probability distribution on $(\mathbb{R}_+^n, \mathcal{B}(\mathbb{R}_+^n))$ with density with respect to $F^{\otimes n}$:

$$\prod_{i=1}^n \exp\{Y_i \psi'^{-1}[g(x_i)] - \psi(\psi'^{-1}[g(x_i)])\},$$

where $Y_1, \dots, Y_n$ are the coordinate variables on $(\mathbb{R}_+^n, \mathcal{B}(\mathbb{R}_+^n))$. Then we have that for large enough n (see [5]),

$$E_{\chi_n^g}\left\{\frac{1}{n}\sum_{i=1}^n Y_i \Phi(x_i)\right\} \in B(c, \epsilon). \tag{15}$$

Let P_n^{ME} be the probability given by the MEM construction at stage n for the relaxed constraint problem (see [13]). Now, a nice property of P_n^{ME} is that it satisfies a Pythagoras theorem (see [4], [26]). In particular, since χ_n^g satisfies (15), we have

$$\frac{1}{n}K(\chi_n^g, F^{\otimes n}) \geq \frac{1}{n}K(P_n^{\mathrm{ME}}, F^{\otimes n}) + \frac{1}{n}K(\chi_n^g, P_n^{\mathrm{ME}}).$$

Simple calculations lead to

$$\begin{aligned}
\frac{1}{n}K(\chi_n^g, F^{\otimes n}) &= \frac{1}{n}\sum_{j=1}^n \gamma(g(x_j)), \\
\frac{1}{n}K(P_n^{\mathrm{ME}}, F^{\otimes n}) &= \frac{1}{n}\sum_{j=1}^n \gamma(\psi'[\langle v_n, \Phi(x_j)\rangle]), \\
\frac{1}{n}K(\chi_n^g, P_n^{\mathrm{ME}}) &= \frac{1}{n}\sum_{j=1}^n \{g(x_j)\{\psi'^{-1}[g(x_j)] - \langle v_n, \Phi(x_j)\rangle\} \\
&\qquad - \psi[\psi'^{-1}\{g(x_j)\}] + \psi[\langle v_n, \Phi(x_j)\rangle]\},
\end{aligned}$$

where (v_n) is a sequence of $\mathbb{R}^k$ converging to v^ϵ. Taking the limit as n goes to infinity we find (see [13]),

$$\Gamma(g) \geq I(G^{MEM,\epsilon}) + D_\psi(g, \psi'(\langle v^\epsilon, \Phi(x)\rangle)). \tag{16}$$

Now, for any $G \in \mathcal{S}(c,\epsilon)$ by Lemma 6 (see below) there exists a sequence of positive continuous function (g_n) such that (g_nP) converges weakly to G and (17) and (18) hold. Therefore, using inequality (16) for g_n, as n goes to infinity we find (11). □

LEMMA 6. *Let $G \in \mathcal{M}_+$ and h be a nonnegative measurable function on U such that $\psi'^{-1}(h)$ is continuous and $\mathcal{D}_\psi(G,h) < \infty$. Then, there exists a sequence of continuous positive functions (g_n) such that g_nP converges weakly to G and*

$$\begin{aligned}\lim_{n\to\infty} D_\psi(g_n,h) &= \lim_{n\to\infty} \mathcal{D}_\psi(g_nP,h)\\ &= \mathcal{D}_\psi(G,h),\end{aligned} \tag{17}$$

$$\begin{aligned}\lim_{n\to\infty} \mathcal{D}_\psi(g_n,\psi'(0)) &= \lim_{n\to\infty} \mathcal{D}_\psi(g_nP,\psi'(0))\\ &= \mathcal{D}_\psi(G,\psi'(0)).\end{aligned} \tag{18}$$

Proof. The existence of a sequence (g_n) such that $g_n.P$ converges weakly to G and (18) holds is proved in [13, Lem. 3]. Therefore, we will only verify that (17) holds for this sequence. This is obvious using the formula

$$D_\psi(g_n,h) = \Gamma(g_n) - \Gamma(h) + \int_U (h(x) - g_n(x))\psi'^{-1}[h(x)]\,dP(x). \qquad \square$$

Acknowledgments. The authors want to thank A. Mokkadem and P. Doukhan for helpful discussions and an anonymous referee for helpful comments.

REFERENCES

[1] J. M. BORWEIN AND A. S. LEWIS, *Duality relationships for entropy-like minimization problems*, SIAM J. Control Optim., 29 (1991), pp. 325–338.

[2] ———, *Partially-finite programming in L_1 : Entropy maximization*, SIAM J. Control Optim., 3 (1993), pp. 248–267.

[3] I. CSISZAR, *I-divergence geometry of probability distributions and minimization problems*, Ann. Probab., 3 (1975), pp. 146–158.

[4] ———, *Sanov property, generalized I-projection and a conditional limit theorem*, Ann. Probab., 12 (1984), pp. 768–793.

[5] D. DACUNHA-CASTELLE AND F. GAMBOA, *Maximum d'entropie et problème des moments*, Ann. Inst. H. Poincaré, 26 (1990), pp. 567–596.

[6] ———, *Maximum d'entropie, fonctions de type négatif et généralisation des familles exponentielles*, Tech. Report, University of Orsay, Orsay, France, 1988.

[7] D. DONOHO, *Superresolution via sparsity constraints*, SIAM J. Math. Anal., 23 (1992), pp. 1309–1331.

[8] D. DONOHO AND E. GASSIAT, *Superresolution via positivity constraints*, 1991, manuscript.

[9] D. DONOHO, I. JOHNSTONE, J. HOCH, AND A. STERN, *Maximum entropy and the nearly black object*, J. Roy. Statist. Soc. Ser. B, 54 (1992), pp. 41–82.

[10] P. DOUKHAN AND F. GAMBOA, *Superresolution rates in Prokhorov metric*, Canad. J. Math., 1996, to appear.

[11] B. R. FRIEDEN, *Restoring with maximum entropy* II: *Superresolution of photograph with diffraction-blurred impulses*, J. Opt. Soc. Amer., 62 (1972), pp. 1202–1210.

[12] F. GAMBOA AND E. GASSIAT, *Maximum d'entropie et problème des moments cas multidimensionnel*, Prob. Math. Statist., 12 (1991), pp. 67–83.

[13] F. GAMBOA AND E. GASSIAT, *Bayesian methods and maximum entropy for ill posed inverse problems*, Ann. Statist., 1996, to appear.

[14] ———, *The maximum entropy method on the mean: Applications to linear programming and superresolution*, Math. Programming, 66 (1994), pp. 103–122.

[15] F. GAMBOA AND H. GZYL, *Solving Fredholm equations by maximum entropy on the mean. Application to superresolution*, Math. Programming Ser. A., submitted.

[16] E. GASSIAT, *Problème des moments et concentration de mesure*, C. R. Acad. Sci. Paris Sér. I Math., 310 (1990), pp. 41–44.

[17] R. HART AND L. SIMON, *Nodal sets for solutions of elliptic equations*, J. Differential Geom., 30 (1989), pp. 505–522.

[18] P. A. JANSSON, ED., *Deconvolution, With Applications in Spectroscopy*, Academic Press, New York, 1984.

[19] M. G. KREIN AND A. A. NUDEL'MAN, *The Markov moment problem and extremal problems*, Amer. Math. Soc. Transl., 50 (1977), p. 71.

[20] A. S. LEWIS, *Superresolution in the Markov moment problem*, University of Waterloo, Waterloo, ON, Canada, 1993, preprint.

[21] ———, *Consistency of Moment Systems*, University of Waterloo, Waterloo, ON, Canada, 1993, preprint.

[22] A. MOKKADEM, *Proprieétés de Mélange des processus autorégressifs polynomiaux*, Ann. Inst. H. Poincaré, 26 (1990), pp. 219–260.

[23] J. NAVAZA, *Use of non local constraints in maximum entropy electron reconstruction*, Acta Cryst. Sect. A, 42 (1986), pp. 212–222.

[24] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[25] ———, *Integrals which are convex functionals* II, Pacific J. Math., 39 (1971), pp. 439–469.

[26] F. TOPSØE, *Information theoretical optimization techniques*, Kybernetika, 15 (1979), pp. 8–27.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 1153–1157, July 1996

017

A DISTRIBUTIONAL SAMPLING THEOREM*

YOUMING LIU†

Abstract. The classical Shannon sampling theorem has many extentions [*A Tutorial in Theory and Applications*, C. K. Chui, ed., Academic Press, 1992, pp. 51–70], one of which is to functions of polynomial growth. In this paper, we shall prove the following theorem:

Let $F(\omega)$ be a distribution with compact support on $[-\pi, \pi]$ and $f(t)$ be the Fourier transform. Then

$$f(t) = \sum_n f(n) \frac{\sin \pi(t-n)}{\pi(t-n)}$$

holds in the sense of (C, α) summation under a very mild condition.

The above result improves the theorems in both [*SIAM J. Math. Anal.*, 19 (1988), pp. 1198–1203] and [*Generalized Functions, Convergence Structures, and Their Applications*, B. Stankovic et al., eds., Plenum Press, 1988, pp. 349–357] given by G. G. Walter.

Key words. sampling theorem, distribution, (C, α) summation

AMS subject classifications. 40G05, 41A05, 42C15, 94A11

1. Introduction.

The classical sampling theorem says that

$$f(t) = \sum_n f(n) \frac{\sin \pi(t-n)}{\pi(t-n)} \tag{1.1}$$

for any finite and π-bandlimited signal f (i.e., $f \in L^2(R)$ and the Fourier transform F of f has compact support on $[-\pi, \pi]$). The importance of this theorem was first recognized by Shannon in communication theory. However, many signals in communication do not have finite energy, i.e., $f \notin L^2(R)$. A general class of such signals is given by the Fourier transform of distributions with compact support. Without loss of generality, we consider $A_\pi = \{\hat{F} : F \in S' \text{ and } \operatorname{supp} F \subset [-\pi, \pi]\}$, where S' denotes all linear continuous functionals on Schwartz class S and $\hat{F}$ is the Fourier transform of a distribution F.

Several versions of the sampling theorem that are appropriate for such signals have been studied by Campbell [2], Pfaffelhuber [3], Lee [4], Hoskins [5], and Walter [6]. All suffered from the same shortcoming—that (1.1) had to be modified to obtain a convergence theorem. G. G. Walter retained the elegant formula (1.1) by requiring that the convergence of the series in (1.1) be interpreted in the sense of Abel summability [7].

In this paper, we establish (1.1) for some $F \in A_\pi$ in the sense of (C, α) summability, where α depends on F.

2. Preliminaries and lemmas.

Let g be a 2π-periodic function and $g \in L^2[-\pi, \pi]$, its Fourier series $\sum_n c_n e^{inx}$. The so-called (C, α) summability of the above series means that

$$\lim_{n \to \infty} \sigma_n^\alpha(x) = \lim_n \sum_{v=-n}^{n} \frac{(\alpha)_{n-|v|}}{(\alpha)_n} c_v e^{ivx} = f(x),$$

* Received by the editors April 29, 1994; accepted for publication (in revised form) October 21, 1994.

† Center for Mathematical Sciences, Zhejiang University, Hangzhou, Zhejiang 310027, People's Republic of China. Current address: Department of Applied Mathematics, Beijing Polytechnic University, Beijing 100022, People's Republic of China.

where $\sigma_n^\alpha(x)$ is called the (C,α) partial sum of nth order and $(\alpha)_n = C_{n+\alpha}^\alpha$ or $\binom{n+\alpha}{\alpha}$. In particular, $\sigma_n^0(x) = S_n(x)$ is the usual sum and $S_n(x) = \frac{1}{\pi}\int_{-\pi}^{\pi} g(y)D_n(x-y)dy$ with $D_n(x) = \sin(n+\frac{1}{2})x/2\sin\frac{x}{2}$, the Dirichet kernel. $\sigma_n^1(x) = \sigma_n(x)$ is the $(C,1)$ mean and $\sigma_n(x) = \frac{1}{\pi}\int_{-\pi}^{\pi} g(y)F_n(x-y)dy$ with $F_n(x) = \sin^2\frac{n+1}{2}x/2(n+1)\sin^2\frac{x}{2}$ is known as the Fejer kernel. In general ([8, p. 88]), $\sigma_n^\alpha(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(y)K_n^\alpha(x-y)dy$, where the (C,α) kernel $K_n^\alpha(x) = (1/(\alpha)_n)\sum_{v=0}^n(\alpha-1)_{n-v}D_v(x)$.

To introduce our main result, we need the following lemmas.

LEMMA 1 ([8, p. 115]). *For any real number $\alpha > -1$ and complex number z,*

$$\sum_{v=0}^{n}(\alpha)_v z^v = \left[\sum_{v=0}^{n}(\alpha-1)_v z^v - (\alpha)_n z^{n+1}\right]\frac{1}{1-z}, \tag{2.1}$$

where $(\alpha)_n = \Gamma(\alpha+n+1)/\Gamma(\alpha+1)\Gamma(n+1)$ and $\Gamma(t)$ is the Gamma function.

The proof can be given by the Abel transform [8]. However, when we restrict α to the integer case as in this paper, a fundamental argument follows. In fact, by multiplying both sides of (2.1) by $1-z$ and using $(\alpha)_v = \frac{v+\alpha}{\alpha}(\alpha-1)_v$, we can find a proof after some simple calculations.

Based on Lemma 1, we can show the following.

LEMMA 2. *For any positive integer α, the (C,α) kernel $K_n^\alpha(t)$ can be given by*

$$K_n^\alpha(t) = \frac{\sin[(n+\frac{\alpha+1}{2})t - \frac{\pi\alpha}{2}]}{(\alpha)_n(2\sin\frac{t}{2})^{\alpha+1}} - \sum_{k=1}^{\alpha}\frac{(k-1)_n}{(\alpha)_n}\sin\left[\frac{\alpha-k}{2}t - \frac{\pi(\alpha-k+1)}{2}\right]\left(2\sin\frac{t}{2}\right)^{-\alpha+k-2}.$$

Proof. We have pointed out that $K_n^\alpha(x) = (1/(\alpha)_n)\sum_{v=0}^n(\alpha-1)_{n-v}D_v(x)$. For simplicity, denoting $2\sin\frac{t}{2}(\alpha)_n K_n^\alpha(t)$ by I_n, we have $I_n = \sum_{v=0}^n(\alpha-1)_{n-v}\sin(v+\frac{1}{2})t = \mathrm{Im}[z^{n+\frac{1}{2}}\sum_{v=0}^n(\alpha-1)_v z^{-v}]$, where $z = e^{it}$ and Im denotes the imaginary part of z. By Lemma 1, we can also express I_n as $I_n = \mathrm{Im}[z^{n+\frac{1}{2}}/(1-z)\sum_{v=0}^n(\alpha-2)_v z^{-v}] - (\alpha-1)_n\mathrm{Im}z^{-\frac{1}{2}}/(1-z^{-1})$. Furthermore, it also follows from Lemma 1 that

$$I_n = \mathrm{Im}\frac{z^{n+\frac{1}{2}}}{(1-z^{-1})^{\alpha-1}}\sum_{v=0}^{n} z^{-v} - \sum_{k=2}^{\alpha}(k-1)_n\mathrm{Im}\frac{z^{-\frac{1}{2}}}{(1-z^{-1})^{\alpha-k+1}}.$$

That is,

$$I_n = \mathrm{Im}\frac{z^{n+\frac{1}{2}} - z^{-\frac{1}{2}}}{(1-z^{-1})^{\alpha}} - \sum_{k=2}^{\alpha}(k-1)_n\mathrm{Im}\frac{z^{-\frac{1}{2}}}{(1-z^{-1})^{\alpha-k+1}} = I_n^1 + I_n^2,$$

where

$$I_n^1 = \mathrm{Im}\frac{z^{n+\frac{1}{2}} - z^{-\frac{1}{2}}}{(1-z^{-1})^{\alpha}} \quad \text{and} \quad I_n^2 = -\sum_{k=2}^{\alpha}(k-1)_n\mathrm{Im}\frac{z^{-\frac{1}{2}}}{(1-z^{-1})^{\alpha-k+1}}.$$

On the other hand, it is easy to see that

$$I_n^1 = \frac{1}{(2\sin\frac{t}{2})^\alpha}\left\{\sin\left[\left(n+\frac{\alpha+1}{2}\right)t - \frac{\pi\alpha}{2}\right] - \sin\left(\frac{\alpha-1}{2}t - \frac{\pi\alpha}{2}\right)\right\}$$

and

$$I_n^2 = -\sum_{k=2}^{\alpha}(k-1)_n \frac{\sin[\frac{\alpha-k}{2}t - \frac{\pi}{2}(\alpha-k+1)]}{(2\sin\frac{t}{2})^{\alpha-k+1}}$$

since $z = e^{it}$. Therefore $I_n = I_n^1 + I_n^2 = \sin[(n+\frac{\alpha+1}{2})t - \frac{\pi\alpha}{2}]/(2\sin\frac{t}{2})^{\alpha} - \sum_{k=1}^{\alpha}(k-1)_n \sin[\frac{\alpha-k}{2}t - \frac{\pi}{2}(\alpha-k+1)](2\sin\frac{t}{2})^{\alpha-k+1}$. Finally, it follows that

$$K_n^{\alpha}(t)$$
$$= \frac{\sin[(n+\frac{\alpha+1}{2})t - \frac{\pi\alpha}{2}]}{(\alpha)_n(2\sin\frac{t}{2})^{\alpha+1}} - \sum_{k=1}^{\alpha}\frac{(k-1)_n}{(\alpha)_n}\sin\left[\frac{\alpha-k}{2}t - \frac{\pi}{2}(\alpha-k+1)\right]\left(2\sin\frac{t}{2}\right)^{-\alpha+k-2}$$

from the notation $I_n = 2\sin\frac{t}{2}(\alpha)_n K_n^{\alpha}(t)$. This completes the proof of Lemma 2.

LEMMA 3. *If $\alpha \geq 1$ and $\alpha \in Z$, then (C,α) kernel $K_n^{\alpha}(t) \geq 0$ for each $t \in R$.*

In fact, for $\alpha = 1$, $K_n^{\alpha}(t) = \frac{1}{n+1}\sum_{v=0}^{n} D_v(t) = F_n(t)$, which is the Fejer kernel, and $F_n(t) \geq 0$ since $F_n(t) = \sin^2\frac{n+1}{2}t/2(n+1)\sin^2\frac{t}{2}$. In general, since $F_v = \frac{1}{v+1}[D_0 + D_1 + \cdots + D_v]$, we have that $D_v = (v+1)F_v - vF_{v-1}$ and, furthermore, $K_n^{\alpha}(t) = (1/(\alpha)_n)\sum_{v=0}^{n}(\alpha-1)_{n-v}D_v(t) = (1/(\alpha)_n)\sum_{v=0}^{n}(\alpha-1)_{n-v}[(v+1)F_v - vF_{v-1}] = (1/(\alpha)_n)[\sum_{v=0}^{n}(\alpha-1)_{n-v}(v+1)F_v - \sum_{v=0}^{n-1}(\alpha-1)_{n-v-1}(v+1)F_v] = (1/(\alpha)_n)\{(n+1)F_n + \sum_{v=0}^{n-1}[(\alpha-1)_{n-v} - (\alpha-1)_{n-v-1}](v+1)F_v\} \geq 0$. To prove Theorem 5 below, we also need the following Lemma.

LEMMA 4 (see [6]). *For any distribution F with compact support on $[-\pi,\pi]$, there is a piecewise continuous function G with the same support on $[-\pi,\pi]$ and an integer p such that*

$$F = D^p G + \sum_{j=0}^{p-1} c_j \delta^{(j)},$$

where D^p is the pth differential operator and $\delta^{(j)}$ is the jth derivative of the well-known Dirac functional δ in the distribution sense.

3. Main theorem. With the four lemmas introduced in §2, we are now ready to derive our main theorem.

THEOREM 5. *Let $f \in A_\pi = \{f = \hat{F}, F \in S'$ and $\operatorname{supp} F \subseteq [-\pi,\pi]\}$ and the corresponding G in Lemma 4 satisfy that $G(\omega)/(\omega^2-\pi^2)^{p+2} \in L^1(R)$. Then*

$$f(t) = \sum_n f(n)\frac{\sin\pi(t-n)}{\pi(t-n)}$$

holds in the sense of (C,α) summability with $\alpha > p$ and $\alpha \in Z^+$.

Proof. We can assume that $\alpha = p+1$ since $(C,p+1)$ convergence is stronger than (C,α) convergence for $\alpha > p+1$. Let $\sigma_{nt}^{\alpha}(\omega)$ be the (C,α) mean of the nth partial sum of the Fourier series of the 2π-periodic extention of $e^{it\omega}K_{[-\pi,\pi]}(\omega)$. The proof will be divided into two steps. First, we show

$$\langle D^p G, \sigma_{nt}^{\alpha}\rangle \longrightarrow_{n\to\infty} \langle D^p G, e^{it\omega}\rangle, \tag{3.1}$$

i.e.,

$$\int_{-\pi}^{\pi} G(\omega)D^p\sigma_{nt}^{\alpha}(\omega)d\omega \longrightarrow_{n\to\infty} \int_{-\pi}^{\pi} G(\omega)D^p e^{it\omega}d\omega.$$

Note that the set of all distributions with compact support is identical with the dual E' of $E = C^\infty(R)$, where E is equipped with the topology of locally uniform convergence of functions and all its derivatives ([9, Chap. I.13, Thm. 2]). Here, because $\sigma^\alpha_{nt}(\omega)$ and $e^{it\omega}$ are test functions in E, (3.1) makes sense. Since $D^p\sigma^\alpha_{nt}(\omega) \longrightarrow_{n\to\infty} D^pe^{it\omega}$ a.e. for $\alpha > p$ ([10, p. 59]), we only need to prove L^1 boundedness of $G(\omega)D^p\sigma^\alpha_{nt}(\omega)$ to use the Lebesgue dominated-convergence theorem.

Consider $D^p\sigma^\alpha_{nt}(\omega) = \int_{-\pi}^{\pi} D^pK^\alpha_n(\omega - \xi)e^{it\xi}d\xi$. Noticing that both $K^\alpha_n(\omega)$ and $D^pK^\alpha_n(\omega)$ are 2π-periodic, we have

$$D^p\sigma^\alpha_{nt}(\omega) = 2i\sin\pi t D^{p-1}K^\alpha_n(\omega-\pi) - itD^{p-1}\sigma^\alpha_{nt}(\omega)$$

by the formula of integration by parts. Similarly it is easy to see that

$$D^p\sigma^\alpha_{nt}(\omega) = \sum_{k=1}^{p} P_k(t,\sin\pi t)D^{p-k}K^\alpha_n(\omega-\pi) + P_{\alpha+1}(t,\sin\pi t)\sigma^\alpha_{nt}(\omega), \tag{3.2}$$

where $P_k(t,\sin\pi t)$ are kth-degree polynomials of t and $\sin\pi t$.

On the other hand, it is easy to show that $|D^k\sin nt/\sin^{\alpha+1}t| \le c_kn^k|\sin t|^{-\alpha-1}$. In fact, it is true for $k=1$ due to the inequality $|\sin nt| \le n|\sin t|$ for $n \in Z^+$. Now suppose it is true for $j = 1, 2, \dots, k-1$. We prove that the same is true for $j = k$. Since $D^k\sin nt = D^k((\sin nt/\sin^{\alpha+1}t)\sin^{\alpha+1}t) = \sum_{j=0}^{k-1}\binom{k}{j}D^j(\sin nt/\sin^{\alpha+1}t)D^{k-j}\sin^{\alpha+1}t + D^k(\sin nt/\sin^{\alpha+1}t)\sin^{\alpha+1}t$, it follows that $|D^k\sin nt/\sin^{\alpha+1}t| \le c_kn^k|\sin t|^{-\alpha-1}$. Therefore,

$$|D^{p-k}K^\alpha_n(\omega-\pi)| \le C_k\left|\sin\frac{\omega-\pi}{2}\right|^{-\alpha-1} \tag{3.3}$$

for $1 \le k \le p$ due to Lemma 2. Obviously, $\sigma^\alpha_{nt}(\omega)$ is bounded since $K^\alpha_n(\omega) \ge 0$ (Lemma 3) and

$$|\sigma^\alpha_n(\omega)| \le \int_{-\pi}^{\pi}|K^\alpha_n(\omega-\xi)||e^{it\xi}|d\xi = 1. \tag{3.4}$$

Combining (3.2)–(3.4), we have

$$|D^p\sigma^\alpha_{nt}(\omega)| \le \sum_{k=1}^{p} A_kC_k\left|\sin\frac{\omega-\pi}{2}\right|^{-\alpha-1} + A_{\alpha+1}$$

for any t in bounded set, where $A_{\alpha+1}$ and $A_k(1 \le k \le p)$ are constants. Furthermore, it follows that $G(\omega)D^p\sigma^\alpha_{nt}(\omega)$ is L^1 bounded from the assumption on G and the assumption at the very beginning of the proof. Hence (3.1) is proved.

Next, it is easy to see that

$$\begin{aligned}\left\langle\sum_{j=0}^{p-1} c_j\delta^{(j)}, \sigma^\alpha_{nt}\right\rangle &= \sum_{j=0}^{p-1} c_j(-1)^jD^j\sigma^\alpha_{nt}(0)\\ &\longrightarrow_{n\to\infty}\sum_{j=0}^{p-1} c_j(-1)^j(it)^j = \left\langle\sum_{j=0}^{p-1} c_j\delta^{(j)}, e^{it\omega}\right\rangle.\end{aligned}$$

It follows that $\langle F, \sigma^\alpha_{nt}\rangle \to_{n\to\infty} \langle F, e^{it\omega}\rangle = f(t)$ from (3.1) and the above limit formula. But $\langle F, \sigma^\alpha_{nt}\rangle = \langle F, (C,\alpha)\sum_{k=-n}^{n} c_k(t)e^{ik\omega}\rangle$, $c_k(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{it\omega}e^{-ik\omega}d\omega =$

$\sin\pi(t-k)/\pi(t-k)$, and $\langle F, e^{ik\omega}\rangle = f(k)$. Therefore, we have that

$$\lim_n (C,\alpha) \sum_{k=-n}^{n} f(k)\frac{\sin\pi(t-k)}{\pi(t-k)} = f(t),$$

i.e.,

$$f(t) = \sum_n f(n)\frac{\sin\pi(t-n)}{\pi(t-n)}$$

in the sense of (C,α) summability. This completes the proof of Theorem 5.

Remark 1. The condition $G(\omega)/(\omega^2-\pi^2)^{p+2} \in L^1(R)$ in Theorem 5 cannot be removed. For example, take $F(\omega) = \delta(\omega-\pi) - \delta(\omega+\pi) = DK_{[-\pi,\pi]}(\omega)$. Then F satisfies all conditions in Theorem 5 except for this L^1 condition. It is clear that the sampling theorem fails since $f(t) = \frac{1}{2\pi}\langle F(\omega), e^{i\omega t}\rangle = \frac{i}{\pi}\sin\pi t$ and $f(n) = 0$ for each integer n but $f(t)$ is not equal to 0 identically.

Remark 2. Theorem 5 improves the result in [6, Thm. 4.3]. In fact, the condition $\operatorname{supp} F \subset [-\pi+\sigma, \pi-\sigma]$ in [6] implies this L^1 condition. This theorem also improves the main result in [7] in some sense since (C,α) convergence is stronger than Abel convergence. In [7], G. G. Walter used a strange terminology on F, strongly integrable. Instead, in this paper, we used an easily understandable condition on G, i.e., L^1 condition, which is very weak. It should be pointed out that the strong integrability of F implies that $|G(\omega)/(\omega-\pi)^p|$ and $|G(\omega)/(\omega+\pi)^p|$ converge to 0 as $\omega \to \pm\pi$ [7]. Therefore, our conditions and Walter's are closely related.

REFERENCES

[1] G. G. Walter, *Wavelets and generalized functions*, in A Tutorial in Theory and Applications, C. K. Chui, ed., Academic Press, New York, 1992, pp. 51–70.

[2] L. L. Campbell, *A comparision of the sampling theorems of Kramer and Whitlaker*, J. Soc. Indust. Appl. Math., 12 (1964), pp. 117–130.

[3] E. Pfaffelhuber, *Sampling series for bandlimited generalized functions*, IEEE Trans. Inform. Theory, IT 17 (1971), pp. 650–654.

[4] A. J. Lee, *A note on the Campbell sampling theorem*, SIAM J. Appl. Math., 41 (1981), pp. 553–557.

[5] R. F. Hoskins and J. de Sousa Pinto, *Sampling expansions for functions bandlimited in the distribution sense*, SIAM J. Appl. Math., 44 (1984), pp. 605–610.

[6] G. G. Walter, *Sampling bandlimited functions of polynomial growth*, SIAM J. Math. Anal., 19 (1988) pp. 1198–1203.

[7] ———, *Abel summability for a distributional sampling theorem*, in Generalized Functions, Convergence Structures, and Their Applications, B. Stankovic et al. eds., Plenum Press, New York, 1988, pp. 349–357.

[8] J. Chen, *Trigonometric Series*, vol. 1, Shanghai Press, Shanghai, People's Republic of China, 1964.

[9] K. Yoshida, *Functional Analysis*, Springer-Verlag, Berlin, Heideberg, New York, 1980.

[10] A. Zygmund, *Trigonometric Series*, vol. 2, Cambridge University Press, Cambridge, 1959.

SIAM J. MATH. ANAL.
Vol. 27, No. 4, pp. 1158–1192, July 1996

CONSTRUCTION OF ORTHOGONAL WAVELETS USING FRACTAL INTERPOLATION FUNCTIONS*

GEORGE C. DONOVAN[†], JEFFREY S. GERONIMO[†], DOUGLAS P. HARDIN[‡], AND PETER R. MASSOPUST[§]

Abstract. Fractal interpolation functions are used to construct a compactly supported continuous, orthogonal wavelet basis spanning $L^2(\mathbb{R})$. The wavelets share many of the properties normally associated with spline wavelets, in particular, they have linear phase.

Key words. wavelets, fractal interpolation functions, linear phase

AMS subject classification. 41A15

1. Introduction. A wavelet basis of some function space, for example $L^2(\mathbb{R})$, is obtained by considering translates and dilates of one or several suitable functions [1, 5, 16, 17, 19]. Much of the recent interest in these bases has stemmed from the fact that they can be built having various useful properties such as continuity, orthogonality, compact support, vanishing moments, etc. Most of the wavelets that have been investigated to date can be constructed using the notion of multiresolution analysis [16, 17]. Let ϕ be a function in $L^2(\mathbb{R})$ and set $\phi_{k,j}(x) = \phi(2^k x - j)$. For each $k \in \mathbb{Z}$, denote by V_k the L^2-closure of the algebraic span of $\{\phi_{k,j} : j \in \mathbb{Z}\}$. The function ϕ is said to generate a multiresolution analysis if the following conditions are satisfied:

(i) $\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots$;
(ii) $\operatorname{clos}_{L^2}\left(\bigcup_{k\in\mathbb{Z}} V_k\right) = L^2$;
(iii) $\bigcap_{k\in\mathbb{Z}} V_k = \{0\}$; and
(iv) $\{\phi_{0,j} : j \in \mathbb{Z}\}$ is a Riesz basis for V_0.

If ϕ generates a multiresolution analysis, then ϕ is called a *scaling function* and will satisfy the refinement equation

$$\phi(x) = \sum c_j \phi(2x - j). \tag{1.1}$$

There also exists a $\psi \in L^2(\mathbb{R})$ such that $\overline{\operatorname{span}\{\psi(\cdot - i), i \in \mathbb{Z}\}} = W_0$, where W_0 is the orthogonal complement of V_0 in V_1 [16, 17]. Furthermore, the function ψ satisfies the equation

$$\psi(x) = \sum d_j \phi(2x - j). \tag{1.2}$$

If $\{\phi_{0,k}\}$ is an orthonormal basis for V_0, the formula

$$d_j = (-1)^j c_{1-j} \tag{1.3}$$

*Received by the editors October 1, 1993; accepted for publication (in revised form) December 6, 1994.

[†]School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332. The research of this author was partially supported by a National Science Foundation Graduate Fellowship.

[‡]Department of Mathematics, Vanderbilt University, Nashville, TN 37240.

[§]Department of Mathematics, Sam Houston State University, Huntsville, TX 77341.

gives a simple way of computing the coefficients of equation (1.2). The celebrated work of Daubechies [5] gives explicit construction of finite sequences of coefficients $\{c_j\}$ which give solutions of equation (1.1) that are orthogonal and compactly supported and have varying degrees of smoothness.

Recently, Hardin, Kessler, and Massopust [13] showed that certain classes of fractal interpolation functions (FIFs) also generate a multiresolution analysis of $L^2(\mathbb{R})$. This multiresolution analysis has certain geometric features that are similar to the multiresolution analysis generated by splines [4]. These results have been generalized to several dimensions in [8] and [9]. In [10], scaling functions were exhibited that form an orthonormal basis for the V_0 given in [10]. Here we continue the investigation of the multiresolution analysis arising from FIFs and construct orthogonal, compactly supported continuous wavelets. Wavelets with varying orders of differentiability will be considered in a later paper [7]. These wavelets fall outside the class constructed in [5] and the multiresolution analysis from which they arise yields several scaling functions instead of just one. In this case (1.1) takes the form

$$\Phi(x) = \sum C_j \Phi(2x - j), \tag{1.4}$$

where each C_j is a square matrix the size of which is determined by the number of scaling functions. Multiresolution analyses based upon several scaling functions have also appeared in the work of Micchelli [18], Goodman, Lee, and Wang [11], Goodman and Lee [12], Jia and Shen [14], and Hervé [15].

We proceed as follows. In §2, we review the relevant facts on FIFs and the multiresolution analysis arising from these function spaces. Then in §3, we exhibit and solve the equations that give scaling functions whose dilates form an orthonormal basis for a certain V_0. We also examine the smoothness of these scaling functions and exhibit their Fourier transforms. In §4, we use the scaling functions constructed above to find compactly supported, continuous, orthogonal wavelets. We investigate the support properties of these wavelets and discuss how to convert them into a wavelet basis for $L^2[0,1]$. Finally, in §5, we extend the methods developed in §§3 and 4 to integer scalings other than 2. For these cases there are a number of parameters that are free to be specified. We examine this in the case of scaling by three and exhibit a one parameter family of symmetric scaling functions.

2. Fractal interpolation functions.

Let $I = [0,1]$, $B(I)$ denote the Banach space of bounded real-valued functions on I with the ∞-norm, and let $C(I) \subset B(I)$ be the space of real-valued functions continuous on I. Let $u_i : [0,1) \to [0,1)$ and $v_i : [0,1) \times \mathbb{R} \to \mathbb{R}$, $i = 0, 1, \ldots, N-1$, be as follows:

$$u_i(x) = \frac{1}{N}(x + i), \tag{2.1}$$

$$v_i(x,y) = \lambda_i(x) + s_i y, \tag{2.2}$$

where $\lambda_i(x) \in \Pi_m$, the set of polynomials with degree at most m. It will always be assumed that $s = \max|s_i| < 1$. Let $I_i = u_i([0,1)) = [\frac{i}{N}, \frac{i+1}{N})$ for $i = 0, 1, 2, \ldots, N-1$, $\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_{N-1})$, and define $\Phi_\lambda : B(I) \to B(I)$ by

$$(\Phi_\lambda f)(x) = v_i(u_i^{-1}(x), f(u_i^{-1}(x))) \tag{2.3}$$

for $x \in I_i$, $i = 0, 1, \ldots, N-1$. Note that $f_\lambda(0) = \lambda_0(0)/(1-s_0)$ and $f_\lambda(1^-) = \lambda_{N-1}(1^-)/(1-s_{N-1})$.

Equations (2.2) and (2.3) imply that Φ_λ is a contraction on $B(I)$ with contractivity s; thus,

$$\|\Phi_\lambda f - \Phi_\lambda g\|_\infty \le s\|f-g\|_\infty, \tag{2.4}$$

and so Φ_λ has a unique attractive fixed point $f_\lambda \in B(I)$. In the event that Φ_λ satisfies the join-up conditions

$$v_{i+1}(0, f_\lambda(0)) = v_i(1^-, f_\lambda(1^-)), \qquad i = 0, 1, \ldots, N-2, \tag{2.5}$$

then f_λ is continuous and is called a fractal interpolation function (Barnsley [2]). In general, $G = \text{graph } f_\lambda$ is typically a fractal set in $\mathbb{R}^2$ made up of images of itself. To see this, let $w_i : [0,1) \times R \to [0,1) \times R$ be given by

$$w_i(x,y) = (u_i(x), v_i(x,y))$$

for $i = 0, 1, \ldots, N-1$. Then (2.3) implies that

$$G = \bigcup_{j=0}^{N-1} w_j(G).$$

Let $\beta = \bigotimes_{j=0}^{N-1} \Pi_m$ and $\lambda \in \beta$. The following theorem gives the basic correspondence between elements in β and functions in $B(I)$.

THEOREM 2.1. (See [10, 12].) *The mapping* $\lambda \overset{\theta}{\mapsto} f_\lambda$ *is a linear isomorphism from* β *to* $\theta(\beta)$.

We will be interested in the case when $m = 1$. If f_λ satisfies equation (2.5), then $f_\lambda \in C(I)$ and the space $\theta(\bigotimes_{j=0}^{N-1} \Pi_1) \cap C(I) = S_0$ is $N+1$ dimensional. Thus each element $g \in S_0$ is completely determined by $g(i/N)$, $i = 0, 1, \ldots, N$. This allows us to view θ in a slightly different manner. Given $\bar{y} \in R^{N+1}$ let $f_{\bar{y}}$ be the unique element of S_0 passing through the points $(\frac{i}{N}, y_i)$, $i = 0, 1, \ldots, N$. We shall call functions $f_{\bar{y}} \in S_0$ affine fractal interpolation functions (AFIFs).

COROLLARY 2.2. *The map* $\theta : R^{N+1} \to S_0$ *is a linear isomorphism.*

The fact that S_0 is isomorphic to R^{N+1} adds a geometric component to the multiresolution analysis associated with FIFs and will play an important role in our construction of wavelets. Let $C_b(\mathbb{R})$ be the space of bounded continuous functions on $\mathbb{R}$ Fix $s_0, s_1, \ldots, s_{N-1}$, let $\tilde{V}_0 = \{f : f|_{[i,i+1)} \text{ is an AFIF}\} \cap C_b(\mathbb{R}) \cap L^2(\mathbb{R})$, and define $f \in \tilde{V}_k \Leftrightarrow f(N^{-k}\cdot) \in \tilde{V}_0$. Then it was shown in [10] and [13] that the sequence $\{\tilde{V}_i\}$ has the following properties:

(a) $\cdots \tilde{V}_{-1} \subset \tilde{V}_0 \subset \tilde{V}_1 \cdots$,

(b) $\bigcap_{k \in R} \tilde{V}_k = \{0\}$.

Note that in contrast with [16] and [17], the spaces $\{\tilde{V}_i\}$ given above are defined independently of any particular scaling functions, which is one of the several properties these spaces share with the spline spaces S_d^r with integer knots (see [10], [11]). In fact, $\tilde{V}_0$ is spanned by sets formed from the integer translates of several scaling functions.

We say a multiresolution analysis is *continuous* and/or *compactly supported* if it is generated by a finite set of scaling functions $\{\phi^i(x)\}_{i=1}^N$, $\phi^i(x) \in L^2(\mathbb{R})$, $i = 1, \ldots, N$ such that each ϕ^i is continuous and/or compactly supported on $\mathbb{R}$. If $\langle\phi^i(\cdot),$

$\phi^j(\cdot - k)\rangle = \delta_{i,j}\delta_{k,0}$, $i = 1,\ldots,N$, then the $\{\phi^i\}$'s generate an *orthogonal* multiresolution analysis.

In order to find *orthogonal scaling functions* $\{\phi^i\}_{i=1}^N$ that generate a continuous, compactly supported, orthogonal multiresolution analysis, we develop some quadrature formulas for AFIFs. Set $I^* = \int_0^1 f_{\bar{y}} g dx$, where $f_{\bar{y}} \in S_0$ and $g \in L^2(I)$. Then from (2.3), we find

$$\begin{aligned} I^* &= \sum_{i=0}^{N-1} \int_{[i/N,(i+1)/N]} f_{\bar{y}} g dx = \sum_{i=0}^{N-1} \int_{[i/N,(i+1)/N]} v_i(u_i^{-1}(x), f_{\bar{y}}(u_i^{-1}(x))g(x)dx \\ &= \frac{1}{N} \sum_{i=0}^{N-1} \int_0^1 v_i(x, f_{\bar{y}}(x))g(u_i(x))dx. \end{aligned}$$

If we use equation (2.2) along with the assumption that $\lambda_i(x) = a_i x + b_i$, $i = 0,1,\ldots,N-1$, in the above equation, we find that

$$\begin{aligned} I^* &= \frac{1}{N} \sum_{i=0}^{N-1} \int_0^1 (a_i x + b_i) g(u_i(x))dx \\ &\quad + \frac{1}{N} \sum_{i=0}^{N-1} s_i \int_0^1 g(u_i(x)) f_{\bar{y}}(x)dx. \end{aligned} \tag{2.6}$$

If $g = 1$, then (2.6) yields [13]

$$I^* = m_0 = \int_0^1 f_{\bar{y}}(x)dx = \frac{\frac{1}{N}\sum_{i=0}^{N-1}\left(\frac{a_i}{2} + b_i\right)}{1 - \frac{1}{N}\sum_{i=0}^{N-1} s_i}, \tag{2.7}$$

while for $g(x) = x$, we find [13]

$$I^* = m_1 = \int_0^1 x f_{\bar{y}}(x)dx = \frac{\frac{1}{N^2}\sum_{i=0}^{N-1}\left[a_i(\frac{i}{2} + \frac{1}{3}) + b_i(i + \frac{1}{2}) + i s_i m_0\right]}{1 - \frac{1}{N^2}\sum_{i=0}^{N-1} s_i}. \tag{2.8}$$

With (2.7) and (2.8), integrals of two fractal functions may be computed. To this end, let $\hat{v}_i(x,y) = \hat{\lambda}_i(x) + \hat{s}_i y$ with $\hat{\lambda}_i = \hat{a}_i x + \hat{b}_i$, $\hat{u}_i = u_i$, $i = 0,1,\ldots,N-1$, and set $g = \hat{f}_{\hat{y}}$. Then [13]

$$\begin{aligned} I^* &= \int_0^1 \hat{f}_{\hat{y}}(x) f_{\bar{y}}(x)dx \\ &= \frac{\frac{1}{N}\sum_{i=0}^{N-1}\left(s_i\hat{a}_i m_1 + \hat{s}_i a_i \hat{m}_1 + s_i\hat{b}_i m_0 + \hat{s}_i b_i \hat{m}_0 + \frac{a_i\hat{a}_i}{3} + \frac{(a_i\hat{b}_i + \hat{a}_i b_i)}{2} + b_i\hat{b}_i\right)}{1 - \frac{1}{N}\sum_{i=0}^{N-1} s_i\hat{s}_i}, \end{aligned} \tag{2.9}$$

where $\hat{m}_0$ and $\hat{m}_1$ are the zeroth and first moments, respectively, of $\hat{f}_{\hat{y}}$.

Consider the $N+1$-dimensional basis $\{f_{\bar{y}_i}\}_{i=0}^N$ spanning S_0, where $\bar{y}_i = e_i$, $0 < i < N$, $\{e_i\}_{i=0}^N$ being the standard basis in R^{N+1}, $\bar{y}_0 = (1, q_1, \ldots, q_{N-1}, 0)$, and $\bar{y}_N = (0, p_1, \ldots, p_{N-1}, 1)$. The sequences $\{p_i\}$ and $\{q_i\}$ are chosen so that $\langle f_{\bar{y}_0}, f_{\bar{y}_i}\rangle = 0 = \langle f_{\bar{y}_N}, f_{\bar{y}_i}\rangle$ for $i = 1,\ldots,N-1$. Extend $f_{\bar{y}_i}, i = 1,\ldots,N-1$, to be functions in

$\tilde{V}_0$ by defining each of them to be equal to 0 for $x \notin I$. Let $\{\phi^i\}_{i=0}^{N-2}$ be a sequence of orthogonal functions in $\tilde{V}_0$ obtained from $\{f_{\bar{y}_i}\}_{i=1}^{N-1}$ by the Gram–Schmidt procedure. The fact that these functions are nonzero follows from Corollary 2.2. Set

$$\phi^{N-1}(x) = \begin{cases} f_{\bar{y}_N}(x), & x \in [0,1), \\ f_{\bar{y}_0}(x-1), & x \in [1,2), \\ 0 & \text{elsewhere} \end{cases} \tag{2.10}$$

and $\hat{\phi}^i(x) = \phi^i(x)/\left\|\phi^i(x)\right\|_{L^2}$; then we find the following.

THEOREM 2.3. *Let $\tilde{V}_0$ and $\phi^i, i = 0,\ldots,N-1$, be as above. Then $\tilde{V}_0 = \operatorname{clos}_{L^2}\operatorname{span}$ $\{\phi^i(\cdot - l) : i = 0,\ldots,N-1, l \in \mathbb{Z}\}$. Furthermore, the set $\{\hat{\phi}^i\}_{i=0}^{N-1}$ generates a continuous, compactly supported multiresolution analysis.*

Proof. From Corollary 2.2 and (2.10), we see that each $\phi^i, i = 0,\ldots,N-1$, is compactly supported and is an element of $\tilde{V}_0$. Since every $f \in \tilde{V}_0$ is determined by its values at $\frac{i}{N}, i \in \mathbb{Z}, f$ has a unique expansion in terms of $f_{\bar{y}_i}, i = 1,\ldots,N-1$, and ϕ^{N-1} and their integer translates. Thus every $f \in \tilde{V}_0$ has a unique expansion in terms of $\{\hat{\phi}^i\}_{i=0}^{N-1}$ and their integer translates. In order to show that $\{\hat{\phi}^i\}_{i=0}^{N-1}$ generates a multiresolution analysis, we must show that (a) $\{\hat{\phi}^i\}_{i=0}^{N-1}$ and its integer translates form a Riesz basis for $\tilde{V}_0$ and (b) $\operatorname{clos}_{L^2}\bigcup_{k\in Z}\tilde{V}_k = L^2$. Since the set $\{\hat{\phi}^i\}_{i=0}^{N-2}$ and its integer translates are an orthogonal set, it follows that we need only show that there exits constants A and $B, 0 < A \leq B < \infty$, such that $\forall c = \{c_i\} \in l^2, A\|c\|_{l^2} \leq \|\sum c_i\hat{\phi}^{N-1}(\cdot - i)\|_{L^2} \leq B\|c\|_{l^2}$. It is easy to show that $B = \sqrt{3}$ provides an upper bound, while the lower bound is obtained by observing that $\|\sum c_i\hat{\phi}^{N-1}(\cdot - i)\|_{L^2}^2 = \sum_i\int_i^{i+1}(c_i\hat{\phi}^{N-1}(x-i) + c_{i-1}\hat{\phi}^{N-1}(x-(i-1)))^2dx$. Since $f_{\bar{y}_0}$ and $f_{\bar{y}_N}$ are linearly independent, the matrix

$$K = \begin{pmatrix} \langle f_{\bar{y}_0}, f_{\bar{y}_0}\rangle & \langle f_{\bar{y}_0}, f_{\bar{y}_N}\rangle \\ \langle f_{\bar{y}_0}, f_{\bar{y}_N}\rangle & \langle f_{\bar{y}_N}, f_{\bar{y}_N}\rangle \end{pmatrix}$$

is positive definite. Let λ be the smallest eigenvalue of $K/\|\phi^{N-1}\|^2$; then A can be taken to be equal to $\sqrt{\lambda}$.

To show that $\bigcup_{k\in Z}\tilde{V}_k$ is dense in L^2, we note that for all $x \in \mathbb{R}$,

$$1 = \sum_i \left(\sum_{j=1}^{N-1} c_i^j f_{\bar{y}_j}(x-i)\right) + \phi^{N-1}(x-i),$$

where $c_i^j = c^j = 1 - p_j - q_j$. Now (b) follows from Proposition 3.1 in [10]. □

3. Orthogonal scaling functions. We begin by considering AFIFs with scaling $N = 2$. By Theorem 2.3, $\{\phi^i\}_{i=0}^{1}$ (note that in this case $\phi^0 = f_{\bar{y}_1}$) and its integer translates span $\tilde{V}_0$, and we look for three mutually orthogonal functions $f_{\bar{y}_0}, f_{\bar{y}_1}$, and $f_{\bar{y}_2}$. From (2.1), (2.2), (2.3), and (2.5), we find that for $f_{\bar{y}_1}$, $\lambda_0 = x$, $\lambda_1 = 1 - x$ for $f_{\bar{y}_2}$, $\lambda_0 = (p_1 - s_0)x$, $\lambda_1 = (1 - s_1 - p_1)x + p_1$, while for $f_{\bar{y}_0}$, $\lambda_0 = (q_1 + s_0 - 1)x + 1 - s_0$, $\lambda_1 = (s_1 - q_1)x + q_1 - s_1$. Substituting these values into (2.9) gives

$$\begin{gathered}\int_0^1 f_{\bar{y}_1}f_{\bar{y}_2}dx = \\ \frac{(4 - 6s_0 + 16p_1 - 2s_1s_0 - 4s_0^2 - 4s_1^2 + 4p_1s_0s_1 + 3s_0^3 + 3s_0s_1^2 - 4p_1s_0^2 - 4p_1s_1^2)}{3(2 - s_0 - s_1)(4 - s_0 - s_1)(2 - s_0^2 - s_1^2)},\end{gathered} \tag{3.1}$$

$$(3.2)\quad \int_0^1 f_{\bar{y}_1} f_{\bar{y}_0} dx = \frac{(4 - 6s_1 + 16q_1 - 2s_1s_0 - 4s_0^2 - 4s_1^2 + 4q_1s_0s_1 + 3s_1^3 + 3s_1s_0^2 - 4q_1s_0^2 - 4q_1s_1^2)}{3(2 - s_0 - s_1)(4 - s_0 - s_1)(2 - s_0^2 - s_1^2)},$$

and

$$(3.3)\quad \begin{aligned} &\int_0^1 f_{\bar{y}_0} f_{\bar{y}_2} dx \\ &= [4(p_1 + q_1)(2s_0^2 + 2s_1^2 + s_0s_1 - 2) + 8p_1q_1(s_0^2 + s_1^2 - s_0s_1 - 4) \\ &\quad + (s_0^2 + s_1^2)^2 - (s_0^2 + s_1^2 + 1)^3 - 4(s_0 + s_1)^2 + s_0^3(-2 + 2s_1 - 6q_1) \\ &\quad + s_1^3(-2 - 6p_1 + 2s_0) + 6q_1s_0(2 - s_1^2) + 6p_1s_1(2 - s_0^2) \\ &\quad + 6s_0 + 6s_1 + 2s_1s_0]/6(-2 + s_0 + s_1)(-4 + s_0 + s_1)(-2 + s_0^2 + s_1^2). \end{aligned}$$

Solving (3.1) for p_1 and (3.2) for q_1 yields

$$(3.4)\quad p_1 = \frac{-(4 - 6s_0 - 2s_1s_0 - 4s_0^2 - 4s_1^2 + 3s_0^3 + 3s_0^2s_1)}{16 + 4s_0s_1 - 4s_0^2 - 4s_1^2}$$

and

$$(3.5)\quad q_1 = \frac{-(4 - 6s_1 - 2s_1s_0 - 4s_0^2 - 4s_1^2 + 3s_1^3 + 3s_0s_1^2)}{16 + 4s_0s_1 - 4s_0^2 - 4s_1^2}.$$

If we substitute (3.4) and (3.5) into (3.3), we find that

$$\int f_{\bar{y}_0} f_{\bar{y}_2} dx = \frac{p(s_0, s_1)}{12(-4 - s_0s_1 + s_0^2 + s_1^2)(-2 + s_0 + s_1)(-4 + s_0 + s_1)},$$

where

$$(3.6)\quad \begin{aligned} p(s_0, s_1) = {} & 2s_1^4 + 6s_1^3 - 7s_1^3s_0 + 18s_1^2s_0 - 28s_1^2 - 7s_1s_0^3 + 18s_1s_0^2 \\ & - 14s_1s_0 + 12s_1 + 2s_0^4 + 6s_0^3 - 28s_0^2 + 12s_0 + 8. \end{aligned}$$

Consequently, we have the following.

LEMMA 3.1. *The AFIFs $f_{\bar{y}_0}$, $f_{\bar{y}_1}$, and $f_{\bar{y}_2}$ with $\bar{y}_0 = (1, q_1, 0)$, $\bar{y}_1 = (0, 1, 0)$ and $\bar{y}_2 = (0, p_1, 1)$ constitute an orthogonal basis for S_0 only for pairs (s_0, s_1) such that $|s_0| < 1$, $|s_1| < 1$, and $p(s_0, s_1) = 0$.*

COROLLARY 3.2. *The only pairs (s_0, s_1) such that the basis $\{f_{\bar{x}_1}, f_{\bar{x}_2}, f_{\bar{x}_3}\}$ with $\bar{x}_1 = (0, 1, a)$, $\bar{x}_2 = (0, b, 1)$, and $\bar{x}_3 = (1, c, 0)$ can be made an orthogonal basis for S_0 are those pairs for which $p(s_0, s_1) = 0$. The same is true for bases of the form $\{f_{\bar{z}_1}, f_{\bar{z}_2}, f_{\bar{z}_3}\}$ with $\bar{z}_1 = (a, 1, 0)$, $\bar{z}_2 = (0, b, 1)$, and $\bar{z}_3 = (1, c, 0)$.*

Proof. Let (s_0, s_1) be such that $\{f_{\bar{x}_1}, f_{\bar{x}_2}, f_{\bar{x}_3}\}$ is an orthogonal basis. Suppose $a \neq 0$ (otherwise the result follows from Lemma 3.1) and let $\{\hat{f}_{\bar{x}_1}, \hat{f}_{\bar{x}_2}, \hat{f}_{\bar{x}_3}\}$ be the corresponding orthonormal basis. If $\hat{f}_{\bar{x}_1}(1) = a'$ and $\hat{f}_{\bar{x}_2}(1) = b'$, set $w_0 = b'\hat{f}_{\bar{x}_1} - a'\hat{f}_{\bar{x}_2}$ and $w = a'\hat{f}_{\bar{x}_1} + b'\hat{f}_{\bar{x}_2}$, where $a^{'2} + b^{'2} = 1$. Then $\{w_0, w_1, \hat{f}_{\bar{x}_3}\}$ is an orthonormal basis of S_0 with $w_0(0) = w_0(1) = 0$ and $w_1(0) = 0$. But by Lemma 3.1, this can only happen for values (s_0, s_1) such that $p(s_0, s_1) = 0$. An analogous argument can be applied to the basis $\{f_{\bar{z}_1}, f_{\bar{z}_2}, f_{\bar{z}_3}\}$. □

From Theorem 2.3 and Lemma 3.1, we have the following.

THEOREM 3.3. *Suppose that the pair* (s_0, s_1) *is such that* $p(s_0, s_1) = 0$ *with* $|s_0| < 1$ *and* $|s_1| < 1$. *Then* $\hat{\phi}^i$, $i = 0, 1$, *generate a continuous, compactly supported, orthogonal multiresolution analysis of* $L^2(\mathbb{R})$.

It follows from Theorem 2.3 that for general pairs (s_0, s_1) with $|s_0| < 1$ and $|s_1| < 1$,

$$\Phi(x) = \begin{pmatrix} \phi^0 \\ \phi^1 \end{pmatrix} = \sum_{i=0}^{3} C_i \Phi(2x - i). \tag{3.7}$$

The 2×2 matrices C_i, $i = 0, 1, 2, 3$, may be computed by evaluating (3.7) at $x = \frac{i}{4}$, $i = 1, 2, \ldots, 8$. From the values of λ_i, $i = 0, 1$, for $f_{\bar{y}_0}$, $f_{\bar{y}_1}$, and $f_{\bar{y}_2}$ computed earlier, we find

$$\begin{aligned} C_0 &= \begin{bmatrix} s_0 + 1/2 - p & 1 \\ \frac{p - s_0}{2} + p(s_0 - p) & p \end{bmatrix}, & C_1 &= \begin{bmatrix} s_1 + 1/2 - q & 0 \\ \frac{1 - s_1 - p}{2} + p(s_1 - q) & 1 \end{bmatrix}, \\ C_2 &= \begin{bmatrix} 0 & 0 \\ \frac{1 - s_0 - q}{2} + q(s_0 - p) & q \end{bmatrix}, & C_3 &= \begin{bmatrix} 0 & 0 \\ \frac{q - s_1}{2} + q(s_1 - q) & 0 \end{bmatrix}. \end{aligned} \tag{3.8}$$

For later use, we define the inner product

$$\langle \Phi, \Phi \rangle = \int_{\mathbb{R}} \Phi\Phi^* dx = \int_{\mathbb{R}} \begin{bmatrix} \phi^0(x)\phi^0(x) & \phi^0(x)\phi^1(x) \\ \phi^1(x)\phi^0(x) & \phi^1(x)\phi^1(x) \end{bmatrix} dx = E^2,$$

where

$$E^2 = \begin{bmatrix} \|\phi^0\|^2 & 0 \\ 0 & \|\phi^1\|^2 \end{bmatrix}.$$

We will now show that even if longer supports are considered, compactly supported, continuous orthogonal scaling functions whose integer translates span V_0 can only be constructed for those values (s_0, s_1) for which $p(s_0, s_1) = 0$ with $|s_0| < 1$ and $|s_1| < 1$.

LEMMA 3.4. *Suppose for a given three-dimensional subspace* V *of* $C(I)$ *there is no orthonormal basis with two functions vanishing at one endpoint and the remaining function vanishing at the other endpoint. Then any continuous, compactly supported pair of functions* ϕ^1 *and* ϕ^2, *composed of linear combinations of the basis elements of* V *and their integer translates constructed so that* $\langle \phi^i(x), \phi^j(x - k) \rangle = \delta_{i,j}\delta_{k,0}$, *must have the property that the leftmost nonzero components of* ϕ^1 *and* ϕ^2 *are linearly dependent, as are their rightmost nonzero components.*

Proof. Suppose the supports of ϕ^1 and ϕ^2 are $[0, N]$ and $[0, M]$, respectively, with $N + M \geq 3$. Because of continuity, $\phi^1(0), \phi^2(0), \phi^1(x + N - 1)|_{x=1}$, and $\phi^2(x + M - 1)|_{x=1}$ all vanish; furthermore, $\phi^1(x)|_{[0,1]}$ and $\phi^2(x)|_{[0,1]}$ are orthogonal to $\phi^1(x + N - 1)|_{[0,1]}$ and $\phi^2(x + M - 1)|_{[0,1]}$. Now $\phi^1(x)|_{[0,1]}$ cannot be independent of $\phi^2(x)|_{[0,1]}$ nor can $\phi^1(x + N - 1)|_{[0,1]}$ be independent of $\phi^2(x + M - 1)|_{[0,1]}$ without violating the hypothesis of the lemma. □

We note that by rotation it can always be arranged that $M \neq N$.

LEMMA 3.5. *If the hypotheses of Lemma* 3.4 *are satisfied, then no such pair of functions* ϕ^1 *and* ϕ^2 *exists.*

Proof. Suppose that ϕ^1 and ϕ^2 exist and that $M < N$, and set $y = \phi^1/\|\phi^1\|_2$ and $z = \phi^2/\|\phi^2\|_2$. Then y and z are orthonormal and for a suitable basis can be represented as

$$\begin{aligned} y &= (y_1, y_2, \ldots, y_M, 0, 0, \ldots, 0) \\ &= ((y_{1,1}, 0, 0), (y_{2,1}, y_{2,2}, y_{2,3}), \ldots, (0, 0, y_{M,3}), (0, 0, 0), \ldots, (0, 0, 0)) \end{aligned}$$

and

$$z = (z_1, z_2, \dots, z_N) = ((z_{1,1}, 0, 0), (z_{2,1}, z_{2,2}, z_{2,3}), \dots, (0, 0, z_{N,3})),$$

where $2 \le M < N$ and $z_{1,1} > 0$.

Consider the rotation r defined on pairs of vectors of the above form,

$$r(y,z) = \frac{1}{\sqrt{y_{1,1}^2 + z_{1,1}^2}}(z_{1,1}y - y_{1,1}z, y_{1,1}y + z_{1,1}z) = (\tilde{y}, \tilde{z}),$$

where

$$\tilde{y} = ((0,0,0), (\tilde{y}_{2,1}, \tilde{y}_{2,2}, \tilde{y}_{2,3}), \dots, (\tilde{y}_{N,1}, \tilde{y}_{N,2}, \tilde{y}_{N,3}))$$
$$\tilde{z} = \left(\left(\sqrt{z_{1,1}^2 + y_{1,1}^2}, 0, 0\right), (\tilde{z}_{2,1}, \tilde{z}_{2,2}, \tilde{z}_{2,3}), \dots, (\tilde{z}_{N,1}, \tilde{z}_{N,2}, \tilde{z}_{N,3})\right),$$

and the shift map s defined on the range of r by

$$s(\tilde{y}, \tilde{z}) = ((\tilde{y}_2, \tilde{y}_3, \dots, \tilde{y}_N, 0), (\tilde{z}_1, \dots, \tilde{z}_N)) = (\hat{y}, \hat{z}).$$

Both s and r are continuous, $\|\hat{z}\|_2 = \|\hat{y}\|_2 = 1$, and $\hat{y}$ and $\hat{z}$ satisfy the necessary orthogonality relations. The operations also preserve continuity of the components, so the functions corresponding to $\hat{y}$ and $\hat{z}$ are continuous. By Lemma 3.4, we know that $\hat{y}_1$ and $\hat{z}_1$ are linearly dependent, so $\hat{y}_{1,2} = 0 = \hat{y}_{1,3}$. It follows that $(\hat{y}, \hat{z})$ is back in the domain of r, so we can iterate the map $s \circ r$ on (y, z) to produce a sequence $(s \circ r)^j(y,z) = (y^{(j)}, z^{(j)})$ for $j = 0, 1, 2, \dots$. Since this sequence is contained in a compact set, we can extract a convergent subsequence with limit—say (Y, Z). By continuity of the inner product, we have that $\|Y\|_2 = \|Z\|_2 = 1$, Y and Z satisfy the orthogonality relations, and the functions corresponding to Y and Z are continuous.

Observe that $z_{1,1}^{(j)}$ is monotone increasing in j. Since it is a component of a unit vector, it is bounded above by 1 and hence converges to some number $Z_{1,1}$. We also have that $(z_{1,1}^{(j)})^2$ converges (to $Z_{1,1}^2$) so that the increments of this sequence, $(z_{1,1}^{(j+1)})^2 - (z_{1,1}^{(j)})^2 = (y_{1,1}^{(j)})^2$, must converge to 0. We claim that for each $i \in \{1, 2, \dots, N-1\}$, $\lim_{j\to\infty} y_{i,1}^{(j)} = 0$. We have just shown that this is true for $i = 1$, so suppose it is true for some i. Then for each j, we have

$$|y_{i,1}^{(j+1)}| = \left| \frac{z_{1,1}^{(j)} y_{i+1,1}^{(j)} - y_{1,1}^{(j)} z_{i+1,1}^{(j)}}{\sqrt{(y_{1,1}^{(j)})^2 + (z_{1,1}^{(j)})^2}} \right|.$$

Since $y_{1,1}^{(j)}$ and $z_{1,1}^{(j)}$ are both components of unit vectors, each is no larger than 1. Thus the denominator (above) is no larger than $\sqrt{2} < 2$, and we have

$$|y_{i,1}^{(j+1)}| \ge \tfrac{1}{2}|z_{1,1}^{(j)} y_{i+1,1}^{(j)}| - \tfrac{1}{2}|y_{1,1}^{(j)} z_{i+1,1}^{(j)}|,$$

or

$$\begin{aligned} |z_{1,1}^{(j)} y_{i+1,1}^{(j)}| &\le 2|y_{i,1}^{(j+1)}| + |y_{1,1}^{(j)} z_{i+1,1}^{(j)}| \\ &\le 2|y_{i,1}^{(j+1)}| + |y_{1,1}^{(j)}|. \end{aligned}$$

By our induction hypothesis, this last expression converges to 0. Furthermore, $z_{1,1}^{(j)} \ge z_{1,1} > 0$, so $\lim_{j\to\infty} y_{i+1,1}^{(j)} = 0$ and the claim is established by induction.

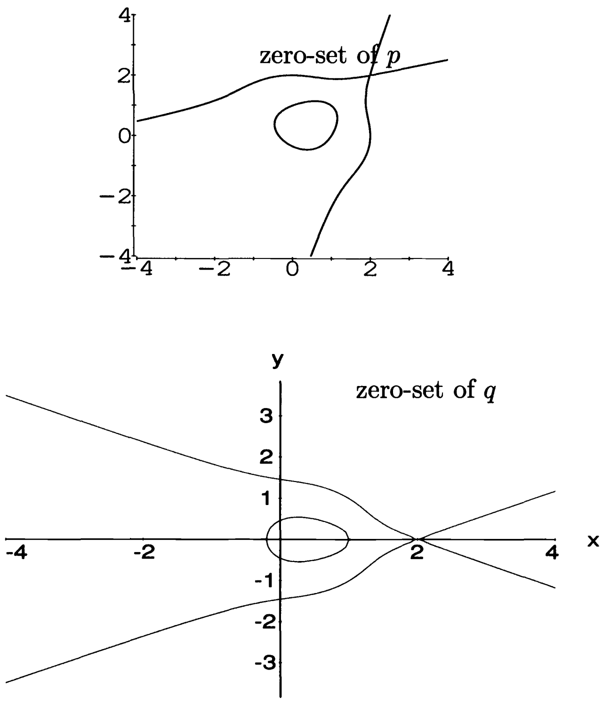


FIG. 1.

Now, our limit point (Y, Z) must have the property that $Y_{i,1} = 0$ for $i = 1, 2, \ldots, N$. But by the lemma, we also know that the leftmost nonzero 3-tuple in Y—say Y_p—is linearly dependent on $Z_1 = (Z_{1,1}, 0, 0)$. So Y_p has the form $(Y_{p,1}, 0, 0)$, and we have just shown that $Y_{p,1} = 0$. Therefore, Y must be 0, which contradicts the fact that $\|Y\| = 1$. □

With the above lemmas we are now able to prove the following.

THEOREM 3.6. *If s_0 and s_1 do not satisfy the relation $p(s_0, s_1) = 0$ (equation (3.6)), then there are no continuous, compactly supported, orthogonal scaling functions formed from the AFIFs generated by s_0 and s_1 such that the L^2-closure of the linear span these functions and their integer translates is $\tilde{V}_0$.*

Proof. It is easy to show that $\tilde{V}_0$ cannot be spanned by the integer translates of one continuous, compactly supported scaling function. This is because, if the scaling function is supported on $[0, M], M > 1$ and we consider an interval of length N, there will be at most $N + M - 1$ translates of ϕ supported on this interval. However, the same interval will contain $2N + 1$ interpolation points. Consequently, for large enough N, there will not be enough translates of ϕ to match all the necessary conditions. By Lemma 3.5 we need now only consider pairs s_0 and s_1 that allow an orthogonal basis for S_0 in which at least two basis vectors vanish at either 0 or 1. But Lemma 3.1 and Corollary 3.2 show that this can occur only in the case when $p(s_0, s_1) = 0$, which proves the result. □

Since compactly supported, continuous scaling functions constructed from AFIFs occur only for the pairs (s_0, s_1), $|s_0| < 1$, $|s_1| < 1$, such that $p(s_0, s_1) = 0$, we examine the zero-set of this polynomial. The next lemma shows that the particular set we are interested in is convex. This confirms the contour plot given in Figure 1.

LEMMA 3.7. *The zero-set of the polynomial*

$$\begin{aligned} p(s_0, s_1) = 2s_0^4 + 6s_0^3 - 7s_0^3 s_1 + 18s_0^2 s_1 - 28s_0^2 - 7s_0 s_1^3 + 18s_0 s_1^2 \\ - 14s_0 s_1 + 12s_0 + 2s_1^4 + 6s_1^3 - 28s_1^2 + 12s_1 + 8 \end{aligned}$$

has two connected components, one a convex closed curve and the other a pair of asymptotically linear curves that cross at a point.

Proof. First, we transform the polynomial using the change of variables

$$\begin{aligned} s_0 &= x - y, \\ s_1 &= x + y \end{aligned}$$

to get a new polynomial,

$$q(x, y) = 18y^4 + (-42 + 24x^2)y^2 - 10x^4 + 48x^3 - 70x^2 + 24x + 8.$$

Note that q is symmetric with respect to y for all values of x. The transformation is just a rotation by 45° and a dilation by $\sqrt{2}$, so it preserves all relevant properties of the zero-set.

Now, we want to find out where the transformed polynomial, q, takes the value 0. To do this we consider the equation $q(x, y) = 0$ and solve for y as a function of x. This gives

$$y = \pm\sqrt{\tfrac{7}{6} - \tfrac{2}{3}x^2 + \tfrac{1}{6}\sqrt{3}\sqrt{12x^4 - 32x^3 + 28x^2 - 16x + 11}},$$

or

$$y = \pm\sqrt{\tfrac{7}{6} - \tfrac{2}{3}x^2 - \tfrac{1}{6}\sqrt{3}\sqrt{12x^4 - 32x^3 + 28x^2 - 16x + 11}}.$$

The first solution gives a pair of asymptotically linear curves that cross at $x = 2$, $y = 0$. The second solution, which is real valued only for $x \in [-\frac{1}{5}, 1]$, gives two halves of a symmetric closed curve, γ, and is the one that we are primarily interested in.

Let f denote the positive branch of γ. If f is a concave down function, it follows by symmetry that γ is a convex curve. Furthermore, since the square-root function is monotone concave down, it suffices to show that f^2 $(= f \cdot f)$ is concave down. Thus, we wish to demonstrate that the second derivative of f^2 is nonpositive on $[-\frac{1}{5}, 1]$. That is,

$$0 \geq -\frac{4}{3} + \frac{\sqrt{3}(48x^3 - 96x^2 + 56x - 16)^2}{24p^{3/2}} - \frac{\sqrt{3}(144x^2 - 192x + 56)}{12p^{1/2}},$$

where $p = 12x^4 - 32x^3 + 28x^2 - 16x + 11$. Note that p is positive for all values of x, so the above expression is well defined. Also, since the denominators are powers of p, they are positive as well, and we can multiply them out to get an equivalent inequality,

$$0 \geq -4p^{3/2} - \sqrt{3}(-144x^6 + 576x^5 - 888x^4 + 832x^3 - 780x^2 + 528x - 122).$$

To show this, we approximate $\sqrt{p}$ with a linear polynomial, $q = \sqrt{3}(\frac{7}{4} - \frac{3}{4}x)$. Observe that $q \leq \sqrt{p}$ on $[-\frac{1}{5}, 1]$ since $p - q^2 = \frac{1}{16}(x-1)(192x^3 - 320x^2 + 101x - 29)$ is nonnegative

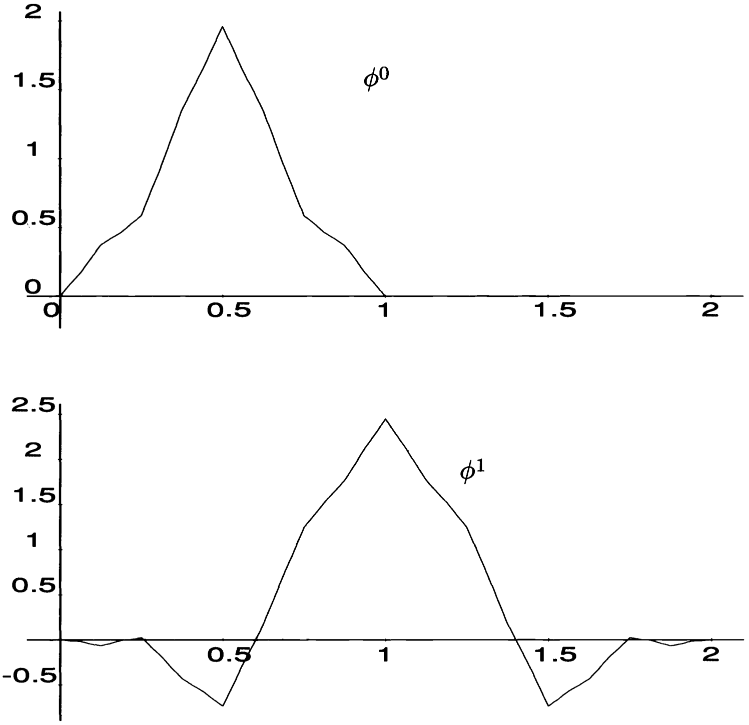


FIG. 2.

on that interval. Thus, by substituting pq for $p^{3/2}$ in the above and expanding, we have

$$\begin{aligned}
&-4p^{3/2} - \sqrt{3}(-144x^6 + 576x^5 - 888x^4 + 832x^3 - 780x^2 + 528x - 122)\\
&\le \sqrt{3}(-144x^6 + 612x^5 - 1068x^4 + 1140x^3 - 1024x^2 + 673x - 199)\\
&\le \sqrt{3}(-144x^6 + 612x^5 - 1068x^4 + 1140x^3 - 1024x^2 + 673x - 199)\\
&\quad + \tfrac{\sqrt{3}}{3125}(37683 - 19989x)\\
&= -\tfrac{4\sqrt{3}}{3125}(5x-4)^2(4500x^4 - 11925x^3 + 11415x^2 - 9729x + 9128)\\
&\le 0.
\end{aligned}$$

The last inequality is justified since both factors involving x are nonnegative for all values of x. □

Let (a, b) be the appropriate pair that solves the equations $\frac{dp}{ds_1} = 0$ and $p = 0$, which yields $a \approx -.4628$. Then by examining $p(s_0, s_1)$ and using Lemma 3.7, we see that for every $s_0 \in [a, 1)$, there is at least one and at most two values of s_1, $|s_1| < 1$ with $p(s_0, s_1) = 0$. If $s_0 = s_1$, then $x = 0$ in the polynomial $q(x, y)$ and we find $q(0, y) = -10y^4 + 48y^3 - 70y^2 + 24y + 8 = -2(5y+1)(y-1)(-2+y)^2$. Thus a solution of this equation is $y = -\frac{1}{5} = s_0 = s_1$, and equations (3.4) and (3.5) give $p_1 = q_1 = -\frac{3}{10}$. Consequently, for $s_0 = s_1 = -\frac{1}{5}$, the functions $f_{\bar{y}_0}, f_{\bar{y}_1}$, and $f_{\bar{y}_2}$, with $\bar{y}_0 = [1, -\frac{3}{10}, 0]$, $\bar{y}_1 = [0, 1, 0]$, and $\bar{y}_2 = [0, -\frac{3}{10}, 1]$, are mutually orthogonal. The scaling functions ϕ^0 and ϕ^1 given as in Theorem 3.3 are shown in Figure 2. Note that ϕ^0 is symmetric about $\frac{1}{2}$ while ϕ^1 is symmetric about 1. Consequently, individually both exhibit linear phase [4].

TABLE 1.

s_0	s_1	p_1	q_1
−.45	.24913840973997585189	−.38009192840012554360	−.12894659319741025282
−.45	.50007955918728063419	−.35582657114704781388	−.023947660348619628508
−.4	.090530851689329739269	−.36870444702962973875	−.19115180858473286044
−.4	.64049995285638175410	−.31094033110232474973	.040636075681553380445
−.35	−.0072635265148628714356	−.35283756141879955795	−.22845520471397257834
−.35	.72178391100872377451	−.27028183540476446288	.081026963799106383553
−.3	−.083203336756332486470	−.33573733406363060725	−.25696121868399870140
−.3	.78288589774512971988	−.23067148400259886260	.11334323161084481474
−.25	−.14618214061458410571	−.31804590990375360555	−.28030023201699694275
−.25	.83269626630597925447	−.19153057613718116605	.14123278956066272834
−.2	−.2	−.30000000000000000000	−.30000000000000000000
−.2	.87501353979083931481	−.15268371309061925454	.16624529998818764227
−.15	−.24668956302990954576	−.28170568236593494170	−.31686046415475332193
−.15	.91187113158837005532	−.11408486745843612530	.18920768747150983345
−.1	−.28747604282460888331	−.26321338049774688720	−.33134872545322789185
−.1	.94449789445249286835	−.075743258091117265968	.21061989468097328770
−.05	−.32315154257329379535	−.24454544031947412926	−.34375214058248711088
−.05	.97369298030356868681	−.037696611328522325658	.23080887479344090984
.0	−.35424868893540940950	−.22570811482256823492	−.35424868893540940950
.05	−.38113082998341069519	−.20669731648304742064	−.36294295781760398380
.1	−.40404262161178188244	.18750151351161996512	−.36988575837081197105
.15	−.42313982625655554203	−.16810316494975289571	−.37508507407448600005
.2	−.43850726493738378110	−.14847932732430920238	−.37851197269001982088
.25	−.45016949520593589167	−.12860173660964793697	−.38010331409967757543
.3	−.45809667756743604588	−.10843651287567490963	−.37976221812377236383
.35	−.46220699424886229565	−.087943553306688133513	−.37735680556076141315
.4	−.46236637072924777344	−.067075631539224926148	−.37271747797639839730
.45	−.45838588104810280947	−.045777188693243047605	−.36563286121393223513
.5	−.45001697146890813343	−.023982773837446344605	−.35584445910606541603
.55	−.43694444222791179535	−.0016150611545678131322	−.34304002442650905123
.6	−.41877692441505661965	.021417671915978278429	−.32684563685890545470
.65	−.39503429791626321840	.045224794037227840313	−.30681645638723853448
.7	−.36513096401300933813	.069937623736922198158	−.28242603273601441995
.75	−.32835278119092827591	.095715398034445688010	−.25305374190333264848
.8	−.28382303116974082780	.12275412910594040701	−.21796900696794990496
.85	−.23044714090374489214	.15130096524457396024	−.17630849139663963600
.9	−.16681210426698377456	.18168005062200866695	−.12703592667444104842
.95	−.090979774570901825562	.21434488601106355137	−.068856253193915809433

If we normalize $\hat{\Phi}$ so that $\langle \hat{\Phi}, \hat{\Phi} \rangle = I$, then the coefficients in (3.7) can be re-expressed as $\hat{\Phi}(x) = \sum_{i=0}^{3} \widehat{C}_i \sqrt{2} \hat{\Phi}(2x - i)$, where $\widehat{C}_i = E^{-1} C_i E / \sqrt{2}$. The $\{\widehat{C}_i\}$ for $s_0 = s_1 = -\frac{1}{5}$ are

$$\widehat{C}_0 = \begin{pmatrix} 3\sqrt{2}/10 & 4/5 \\ -1/20 & -3\sqrt{2}/20 \end{pmatrix}, \qquad \widehat{C}_1 = \begin{pmatrix} 3\sqrt{2}/10 & 0 \\ 9/20 & \sqrt{2}/2 \end{pmatrix},$$

$$\widehat{C}_2 = \begin{pmatrix} 0 & 0 \\ 9/20 & -3\sqrt{2}/20 \end{pmatrix}, \qquad \widehat{C}_3 = \begin{pmatrix} 0 & 0 \\ -1/20 & 0 \end{pmatrix}.$$

The values of p_1 and q_1 for other acceptable pairs s_0 and s_1 are given in Table 1.

We now consider the smoothness and approximation order of the function ϕ^i, $i = 0, 1$. Recall that the Hölder exponent of $f \in C(I)$ at x is

$$\alpha_x = \lim_{\epsilon \to 0} \inf\{\log|f(x) - f(y)| / \log|x - y| : y \in B(x, \epsilon)\}$$

and $\alpha = \inf\{\alpha_x, x \in I\}$ is called the Hölder exponent of f.

Let S be a closed subspace of $L^2(\mathbb{R})$, $E(f,S) = \min\{\|f - s\| : s \in S\}, \hat{f}$ be the Fourier transform of f, and $w_2^k(\mathbb{R}) = \{f \in L^2(\mathbb{R}) : \|f\|_{w_2^k} = \|(1 + |\cdot|^k)\hat{f}\| < \infty\}$. Following de Boor, DeVore, and Ron [6], we say that S provides approximation order k if for every $f \in w_2^k(\mathbb{R})$,

$$E(f, S^h) \leq C_S h^k \|f\|_{w_2^k}(\mathbb{R}),$$

where $S^h = \{s(\frac{\cdot}{h}) : s \in S\}$.

LEMMA 3.8. *Given any pair (s_0, s_1), $|s_0| < 1$, $|s_1| < 1$, let $f \in S_0$. If $|s_0| < \frac{1}{2}$ and $|s_1| < \frac{1}{2}$, then f is Lipschitz continuous, i.e., there exists an $M < \infty$ such that $\forall\, x, y \in [0,1], |f(x) - f(y)| \leq M|x - y|$. If $\max|s_i| > \frac{1}{2}$ and f is not a line, then f has Hölder exponent $\alpha = -\log\max|s_i|/\log 2$.*

We note that the second part of the lemma has already been proved by Bedford [3], and we include the proof for the convenience of the reader.

Proof. Let $s = \max_i |s_i|$, $\mathbf{i}_n = \{i_1, i_2, \ldots, i_n\}$, $i_j \in \{0,1\}$, $j = 1, \ldots, n$, and $a(\mathbf{i}_n) = i_1/2 + i_2/4 + \cdots + i_n/2^n$. Then for $x \in [a(\mathbf{i}_n), a(\mathbf{i}_n) + 1/2^n]$, we find from (2.3) that

$$\begin{aligned} f(x) &= \sum_{k=1}^{n} \left(2^{k-1}x + b_{i_k} - \sum_{m=1}^{k} \frac{i_m}{2^{m-1}}\right) \prod_{j=1}^{k-1} s_{i_j} \\ &\quad + \left(\prod_{j=1}^{n} s_{i_j}\right) f\left(2^n x - \sum_{m=1}^{n} 2^{n-m} i_m\right). \end{aligned} \tag{3.9}$$

Suppose $1/2^{n+1} \leq |x - y| \leq 1/2^n$ and $x, y \in [a(\mathbf{i}_n), a(\mathbf{i}_n) + 1/2^n]$; then the above formula shows us that

$$|f(x) - f(y)| \leq \sum_{k=1}^{n} (2s)^{k-1}|x - y| + s^n C, \tag{3.10}$$

where $C = 2\|f\|_\infty$. Suppose first that $s < \frac{1}{2}$ then

$$|f(x) - f(y)| \leq \sum_{k=1}^{n} (2s)^{k-1}|x - y| + (2s)^n 2C|x - y|,$$

where the fact that $1 \leq 2^{n+1}|x - y|$ has been used to obtain the last term in the above expression. Since $2s < 1$, we find that

$$|f(x) - f(y)| \leq \frac{1 + 2C}{1 - 2s}|x - y|. \tag{3.11}$$

If $1/2^{n+1} \leq |x - y| \leq 1/2^n$ but x and y are not in the same dyadic interval, let x_0 be the boundary point between the respective intervals that x and y are located in. Then $|f(x) - f(y)| \leq |f(x) - f(x_0)| + |f(x_0) - f(y)|$, and applying (3.11) gives

$$|f(x) - f(y)| \leq \frac{2(1 + 2C)}{1 - 2s}\,|x - y|$$

for all $x, y \in [0,1]$. This gives the Lipschitz continuity of f.

If $s = \max|s_i| > \frac{1}{2}$, assume without loss of generality that $f(0) = 0 = f(1)$. For if that is not the case, we can modify f by adding linear functions $L_1(x)$ and $L_2(x)$ so that $\hat{f} = f + L_1 + L_2$ is still in S_0, $\hat{f}(0) = \hat{f}(1) = 0$. If $1/2^{n+1} \leq |x-y| \leq 1/2^n$, $a(\mathbf{i}_n)$, $x, y \in [a(\mathbf{i}_n), a(\mathbf{i}_n) + 1/2^n]$, (3.10) implies

$$|f(x) - f(y)| \leq s^n \left(C + \sum_{k=1}^{n} (2s)^{k-n-1}\right) = s^n \left(C + \frac{2s}{1 - \frac{1}{2s}}\right).$$

Since $1/2^{n+1} \leq |x-y| \leq 1/2^n$, we find that $|x-y|^\alpha \geq 2^{-\alpha(n+1)} = 2^{-\alpha}2^{-\alpha n} = 2^{-\alpha}s^n$ with $\alpha = \frac{\ln s}{\ln 1/2}$. Hence

$$|f(x) - f(y)| \leq 2^\alpha \left(C + \frac{2s}{1 - \frac{1}{2s}}\right) |x-y|^\alpha \tag{3.12}$$

for x, y in the same dyadic interval of length $1/2^n$. If x and y are not in the same dyadic interval, then using the same argument as in the case when $2s < 1$, we find

$$|f(x) - f(y)| \leq 2^{\alpha+1} \left(C + \frac{2s}{1 - \frac{1}{2s}}\right) |x-y|^\alpha$$

for all $x, y \in [0,1]$. We need only show that α is the largest possible exponent. Suppose without loss of generality that $s = |s_0|$. Since f vanishes at 0 and 1 and $f(\frac{1}{2}) \neq 0$ (since f is not a line), there exist distinct points x_0 and $y_0 \in [0,1]$ such that $f(x_0) \neq f(y_0)$ and $(x_0 - y_0)/(f(x_0) - f(y_0)) > 0$. With $x_n = x_0/2^n$ and $y_n = y_0/2^n$, we find from (3.9) that

$$f(x_n) - f(y_n) = \sum_{k=1}^{n-1} (2s_0)^{k-1} \frac{x_0 - y_0}{2^n} + s_0^n (f(x_0) - f(y_0)).$$

Therefore,

$$\begin{aligned} |f(x_n) - f(y_n)| &= s^n |f(x_0) - f(y_0)| \left| 1 + \frac{x_0 - y_0}{f(x_0) - f(y_0)} \left(\frac{1}{2s}\right)^2 \frac{1 - (\frac{1}{2s_0})^{n-2}}{1 - \frac{1}{2s_0}} \right| \\ &\geq s^n C \geq k|x_n - y_n|^\alpha, \end{aligned}$$

which proves the result. □

We can now prove the following.

THEOREM 3.9. *Suppose the pair (s_0, s_1) is such that $|s_0| < 1$ and $|s_1| < 1$ with p_1 and q_1 satisfying (3.4) and (3.5), respectively. Then V_0 provides approximation order 2. If s_0 and s_1 are both in magnitude less than $\frac{1}{2}$, then ϕ^i, $i = 0, 1$, are both Lipschitz. If $s = \max|s_i| > \frac{1}{2}$, then ϕ^i, $i = 0, 1$, have Hölder exponent $\alpha = -\log s/\log 2$.*

Proof. In order to show that V_0 provides approximation order 2, we need only prove that the hat function

$$g(x) = \begin{cases} x, & 0 \leq x \leq 1, \\ 2 - x, & 1 \leq x \leq 2 \end{cases}$$

is in V_0 [20]. It is easy to see from (2.3) that for any pair s_0, s_1 with $0 \leq |s_0| < 1$ and $0 \leq |s_1| < 1$, $f_{[0,1/2,1]}(x) = x$. Consequently,

$$g(x) = (1/2 - p_1)\phi^0(x) + \phi^1(x) + (1/2 - q_1)\phi^0(x-1).$$

The fact that ϕ^0 and ϕ^1 are Lipschitz when $s < \frac{1}{2}$ follows from Lemma 3.7. Furthermore, Lemma 3.7 also shows that if $s > \frac{1}{2}$, then the Hölder exponent of ϕ^0 is $\log|s|/\log\frac{1}{2}$. This will also be true of ϕ^1 once it is shown that ϕ^1 is not piecewise linear. For $s > \frac{1}{2}$ this can happen only if $p_1 = q_1 = \frac{1}{2}$; however, from (3.1) we find that with $p_1 = \frac{1}{2}$,

$$\int f_{\bar y_1} f_{\bar y_2} dx = \frac{-2+s_0}{(-2+s_0+s_1)(-4+s_0+s_1)} \neq 0,$$

which is a contradiction. Therefore, $f_{\bar y_2}$ is not a line. A similar argument shows that $f_{\bar y_0}$ cannot be a line, and the result now follows. □

We complete this section by computing the Fourier transform of ϕ^0 and ϕ^1. To this end, set $g(x) = e^{ikx}$ with $N = 2$ in (2.6) to find

$$\begin{aligned}\hat f_{\bar y}(k) = \int_0^1 e^{ikx} f_{\bar y}(x)dx &= \frac{1}{2}\int_0^1 (a_0x + b_0)e^{ikx/2}dx \\ &+ \frac{1}{2}\int_0^1 e^{ik(x+1)/2}(a_1x+b_1)dx \\ &+ \frac{1}{2}(s_0 + e^{ik/2}s_1)\hat f_{\bar y}\left(\frac{k}{2}\right).\end{aligned}$$

For ϕ^0, $a_0 = 1$, $b_0 = 0$, $a_1 = -1$, and $b_1 = 1$. Consequently,

$$\hat\phi^0(k) = 8e^{ik/2}\frac{\sin^2 k/4}{k^2} + \frac{1}{2}(s_0 + e^{ik/2}s_1)\hat\phi^0\left(\frac{k}{2}\right).$$

With $C(k) = \frac{1}{2}(s_0 + e^{ik/2}s_1)$ and $h_1(k) = 8e^{ik/2}(\sin^2 k/4)/(k^2)$, we find that

$$\hat\phi^0(k) = \sum_{n=0}^{\infty}\left(\prod_{j=1}^{n} C\left(\frac{k}{2^{j-1}}\right)\right) h_1\left(\frac{k}{2^n}\right).$$

The above series converges uniformly for $k \in \mathbb{R}$ since $|C(k/2^j)| \leq (|s_0| + |s_1|)/2 = r < 1$, $j = 1, 2\ldots$, and since $|\sin x/x| \leq 1$ for all $x \in \mathbb{R}$.

To compute the Fourier transform of ϕ^1, we use the fact that ϕ^1 can be written in terms of the hat function g and ϕ^0 as shown above. Thus

$$\hat\phi^1(k) = \hat g(k) - \left(\frac{1}{2} - p_1 + \left(\frac{1}{2} - q_1\right)e^{ik}\right)\hat\phi^0(k).$$

In Figure 3, the magnitude of the Fourier transforms of ϕ^0 and ϕ^1 are plotted when $s_0 = s_1 = -\frac{1}{5}$.

4. Construction of compactly supported wavelets. Having constructed a multiresolution analysis based on the vector $\phi = \binom{\phi^0}{\phi^1}$ and its dilates and translates, we now construct compactly supported, continuous wavelets. If $\tilde W_0$ is the orthogonal complement of $\tilde V_0$ in $\tilde V_1$, then any $\Psi \in \tilde W_0$ can be written as $\Psi(x) = \binom{\psi^0(x)}{\psi^1(x)}$, where $\psi^i(x)|_{[j/2,(j+1)/2)}$, $i = 0, 1$, is an AFIF with interpolation points at the quarter integers.

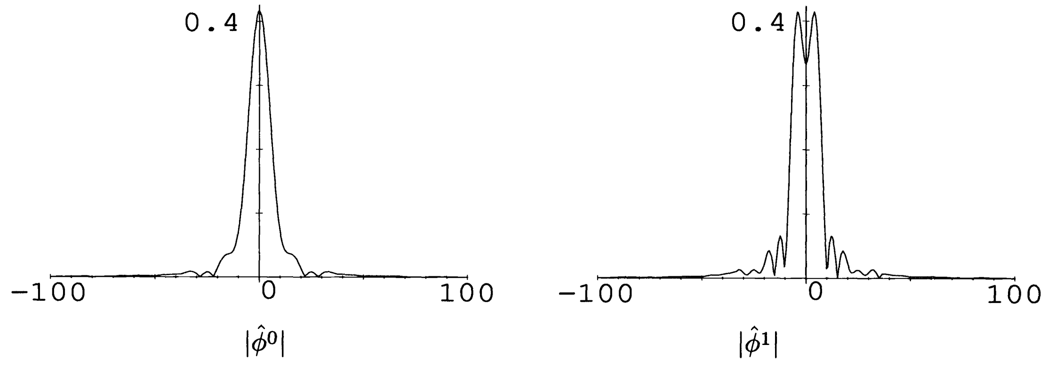


FIG. 3.

Since $\Phi(x) = \binom{\phi^0(x)}{\phi^1(x)}$ is supported on $[0,2]$, we shall look for $\Psi(x)$ to be supported on the same interval. Therefore, set

$$\psi^i(x) = \begin{cases} g_{\bar{y}_1^i}(x), & 0 \leq x \leq \dfrac{1}{2}, \\ g_{\bar{y}_2^i}(x), & \dfrac{1}{2} \leq x \leq 1, \\ g_{\bar{y}_3^i}(x), & 1 \leq x \leq \dfrac{3}{2}, \\ g_{\bar{y}_4^i}(x), & \dfrac{3}{2} \leq x \leq 1, \end{cases} \qquad i = 0, 1. \tag{4.1}$$

Here $\bar{y}_j^i$ are the values $\psi^i(x)$ taken at the quarter-integer points in the intervals indicated. In order to preserve continuity and keep the support of $\Phi(x)$ on $[0,2]$, $\bar{y}_1^i = [0, a_1^i, a_2^i]$, $\bar{y}_2^i = [a_2^i, a_3^i, a_4^i]$, $y_3^i = [a_4^i, a_5^i, a_6^i]$, and $\bar{y}_4^i = [a_6^i, a_7^i, 0]$, $i = 0, 1$. The coefficients a_j^i, $j = 1, 2, \ldots, 7$, $i = 0, 1$, are adjusted so that $\Psi(x)$ is orthogonal to $\Phi(x)$ and its translates and $\langle \psi_0^0(x), \psi^1(x) \rangle = 0$. In the case when $s_0 = -\frac{1}{5} = s_1$ and $p_1 = -\frac{3}{10} = q_1$, it follows from (2.7)–(2.9) that for $i = 0$ or 1,

$$\int_0^2 \psi^i(x)\phi^0(x)dx = \frac{5}{32}a_1^i + \frac{41}{96}a_2^i + \frac{5}{32}a_3^i + \frac{3}{64}a_4^i, \tag{4.2}$$

$$\int_0^1 \psi^i(x)\phi^0(x+1)dx = \frac{3}{64}a_4^i + \frac{5}{32}a_5^i + \frac{41}{96}a_6^i + \frac{5}{32}a_7^i, \tag{4.3}$$

$$\int_0^2 \psi^i(x)\phi^1(x)dx = -\frac{1}{48}a_1^i - \frac{1}{20}a_2^i + \frac{3}{16}a_3^i + \frac{107}{240}a_4^i + \frac{3}{16}a_5^i - \frac{1}{20}a_6^i - \frac{1}{48}a_7^i, \tag{4.4}$$

$$\int_0^2 \psi^i(x)\phi^1(x+1)dx = \frac{3}{16}a_1^i - \frac{1}{20}a_2^i - \frac{1}{48}a_3^i - \frac{1}{160}a_4^i, \tag{4.5}$$

$$\int_0^2 \psi^i(x)\phi^1(x-1)dx = -\frac{1}{160}a_4^i - \frac{1}{48}a_5^i - \frac{1}{20}a_6^i + \frac{3}{16}a_7^i, \tag{4.6}$$

and

$$\begin{aligned}\int_0^2 \psi^0(x)\psi^1(x)dx = {} & \frac{25}{96}(a_1^0a_1^1 + a_3^0a_3^1 + a_5^0a_5^1 + a_7^0a_7^1) \\ & + \frac{73}{192}(a_2^0a_2^1 + a_6^0a_6^1 + a_4^0a_4^1 + a_6^0a_6^1) \\ & + \frac{3}{128}(a_4^1a_2^0 + a_4^0a_2^1 + a_6^0a_4^1 + a_6^1a_4^0) \\ & + \frac{5}{64}(a_2^0a_1^1 + a_1^0a_2^1 + a_3^0a_2^1 + a_2^0a_3^1 \\ & \quad + a_4^0a_3^1 + a_3^0a_4^1 + a_5^0a_4^1 + a_4^0a_5^1 + a_6^0a_5^1 \\ & \quad + a_5^0a_6^1 + a_7^0a_6^1 + a_6^0a_7^1).\end{aligned} \tag{4.7}$$

The above equations once set equal to 0 will fix all but three of the unknowns. Two of these are used to normalize the integrals of ψ^0 and ψ^1. One unknown remains because of the one parameter family of rotations taking Ψ into other mother wavelets. A remarkable fact, to be shown below, is that once (4.3), (4.5), and (4.6) are satisfied, then $\langle \Psi, \Psi(\cdot + i)\rangle = 0 \ \forall\ i \in \mathbb{Z},\ i \neq 0$. A solution to the above equations that give $\langle \Psi, \Psi \rangle = I$ with Ψ having the smallest possible support is

$$\begin{aligned} y_1^0 &= \left[0, \frac{-3\sqrt{2}}{200}, \frac{9\sqrt{2}}{20}\right], & y_2^0 &= \left[\frac{9\sqrt{2}}{20}, \frac{-273\sqrt{2}}{200}, \frac{\sqrt{2}}{2}\right], \\ y_3^0 &= \left[\frac{\sqrt{2}}{2}, \frac{5\sqrt{2}}{200}, \frac{-3\sqrt{2}}{20}\right], & y_4^0 &= \left[\frac{-3\sqrt{2}}{20}, \frac{\sqrt{2}}{200}, 0\right], \end{aligned} \tag{4.8}$$

$$\begin{aligned} y_1^1 &= [0,0,0], & y_2^1 &= \left[0, \frac{3}{10}, -1\right], \\ y_3^1 &= \left[-1, \frac{48}{25}, \frac{-3}{5}\right], & y_4^1 &= \left[\frac{-3}{5}, \frac{1}{50}, 0\right] \end{aligned} \tag{4.9}$$

(see Figure 4). Since $\Psi \in \tilde{V}_1$, we have

$$\Psi(x) = \sum_{i=0}^{3} D_i \Phi(2x - i), \tag{4.10}$$

where for the particular Ψ given by (4.8) and (4.9),

$$\begin{aligned} D_0 &= \begin{pmatrix} \sqrt{3}/20 & 3\sqrt{6}/20 \\ 0 & 0 \end{pmatrix}, & D_1 &= \begin{pmatrix} -9\sqrt{3}/20 & 1/\sqrt{6} \\ 0 & -1/\sqrt{3} \end{pmatrix}, \\ D_2 &= \begin{pmatrix} 3\sqrt{3}/20 & -\sqrt{6}/20 \\ 3\sqrt{6}/10 & -\sqrt{3}/5 \end{pmatrix}, & D_3 &= \begin{pmatrix} -\sqrt{3}/60 & 0 \\ -\sqrt{6}/30 & 0 \end{pmatrix}. \end{aligned} \tag{4.11}$$

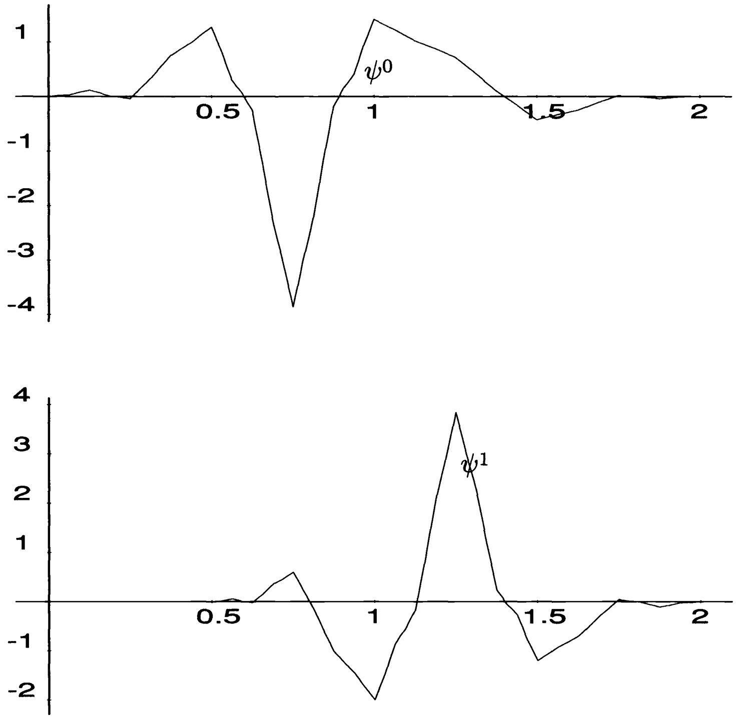


FIG. 4.

In order to see why the orthogonality of Ψ to the translates of Φ implies the orthogonality of Ψ to its nonzero integer translates, we examine the multiresolution analysis arising from Φ and consider the slightly more general case where $\Phi \in \mathbb{R}^N$, which we will use in §5. In this case, each C_i in (3.7) is an $N \times N$ matrix and Φ satisfies the N-scale dilation equation

$$\Phi(x) = \sum_{i=0}^{2N-1} C_i \Phi(Nx - i).$$

Likewise, the mother wavelet Ψ satisfies

$$\Psi(x) = \sum_{i=0}^{2N-1} D_i \Phi(Nx - i),$$

where the matrices D_i are $N(N-1) \times N$ matrices. The orthogonality of $\Phi(x)$ to its integer translates can be reexpressed as

$$(4.12) \qquad \int \Phi(x)\Phi^*(x-i)dx = \delta_{i,0} I_N = \sum_{k=0}^{2N-1} C_k C^*_{k-Ni} \qquad \forall\, i \in \mathbb{Z},$$

where I_N is the $N \times N$ identity matrix. Likewise, the orthogonality of Ψ against its translates gives

$$(4.13) \qquad \sum_{k=0}^{2N-1} D_k D^*_{k-Ni} = \delta_{i,0} I_{N(N-1)} \qquad \forall\, i \in \mathbb{Z}.$$

The fact that W_0 is orthogonal to V_0 in V_1 means

$$\sum_{k=0}^{2N-1} C_k D_{k-Ni}^* = 0 \qquad \forall\, i \in \mathbb{Z}, \tag{4.14}$$

while $V_1 = V_0 \oplus W_0$ can be expressed as

$$\sum_k C_{m-Nk}^* C_{n-Nk} + D_{m-Nk}^* D_{n-Nk} = \delta_{m,n} I_N \qquad \forall\, n, m \in \mathbb{Z} \tag{4.15}$$

(see [5]). If we set

$$H_1 = (C_0, C_1, \dots, C_{N-1}), \qquad H_2 = (C_N, C_{N+1}, \dots, C_{2N-1}),$$

$$G_1 = (D_0, D_1, \dots, D_{N-1}), \quad \text{and} \quad G_2 = (D_N, D_{N+1}, \dots, D_{2N-1}),$$

then (4.12) can be recast as

$$H_1 H_2^* = 0 \tag{4.16}$$

and

$$H_1 H_1^* + H_2 H_2^* = I_N. \tag{4.17}$$

Likewise, (4.13) and (4.14) become

$$G_1 G_2^* = 0, \tag{4.18}$$

$$G_1 G_1^* + G_2 G_2^* = I_{N(N-1)}, \tag{4.19}$$

$$H_2 G_1^* = 0, \tag{4.20}$$

$$H_1 G_2^* = 0, \tag{4.21}$$

and

$$H_1 G_1^* + H_2 G_2^* = 0. \tag{4.22}$$

Equation (4.15) can be recast as

$$H_1^* H_1 + H_2^* H_2 + G_1^* G_1 + G_2^* G_2 = I_{N^2} \tag{4.23}$$

and

$$H_1^* H_2 + G_1^* G_2 = 0. \tag{4.24}$$

The general solution to (4.16) is $H_2^* = P_1 Y$, where $P_1 : \mathbb{R}^{N^2} \to \mathbb{R}^{N^2}$ is an orthogonal projection onto the null space of H_1 and Y is any $N^2 \times N$ matrix. Likewise, from (4.21), it is easy to see that $G_2^* = P_1 X$, where X is any $N^2 \times N(N-1)$ matrix. If we set $H_1^* = (I_{N^2} - P_1)Y$, then (4.16) and (4.17) are satisfied and $Y = H_1^* + H_2^*$. Observe that $Y^*Y = I_N$. Equations (4.18)–(4.20) now suggest that $G_1^* = (I_{N^2} - P_1)X$ with $X^*X = I_{N(N-1)}$, and (4.22) says that $Y^*X = 0$. Since Y is an $N^2 \times N$ matrix, and X is an $N^2 \times N(N-1)$ matrix, it is an idea of Gil Strang that X can be obtained by letting its columns be the remaining orthonormal basis vectors for the $\mathbb{R}^{N^2}$, the first N basis vectors being the columns of Y. That is, we choose X so that the matrix $(Y\ X)$ is an orthogonal matrix. If this is done, (4.23) and (4.24) will also be satisfied since $YY^* + XX^* = I_{N^2}$. Thus equations (4.16)–(4.24) and the assumptions on G_1, G_2, H_1, and H_2 can be summarized as

$$\begin{pmatrix} Y^* \\ X^* \end{pmatrix}(Y\ X) = I_{N^2} = (Y\ X)\begin{pmatrix} Y^* \\ X^* \end{pmatrix}.$$

The filters arising from the above equations are closely related to those found in Vetterli [22] (also see Strang and Strela [21]). Thus we have shown the following.

THEOREM 4.1. *Let $\{C_i\}_{i=0}^{2N-1}$ be $N \times N$ matrices satisfying* (4.12). *Then there exist $N(N-1) \times N$ matrices $\{D_i\}_{i=0}^{2N-1}$ constructed as above so that equations* (4.13)–(4.15) *are satisfied.*

We now show that for any wavelet Ψ that is constructed using the above scaling function Φ, the minimum length of the support of any of its components is $\frac{3}{2}$ and at least one component must have a support greater than or equal to 2. Before proving the next lemma, we note that for $|s_0| < 1$, $|s_1| < 1$, and $f \in \tilde{V}_0$,

$$f(x) = \sum_i (c_i^0 \phi^0(x-i) + c_i^1 \phi^1(x-i)), \tag{4.25a}$$

where

$$c_i^1 = f(i+1) \tag{4.25b}$$

and

$$c_i^0 = f\left(i + \frac{1}{2}\right) - c_i^1 \phi^1\left(\frac{1}{2}\right) - c_{i-1}^1 \phi^1\left(\frac{3}{2}\right). \tag{4.25c}$$

LEMMA 4.2. *For any pair (s_0, s_1) such that $p(s_0, s_1) = 0$ with $|s_0| < 1$ and $|s_1| < 1$, there is no wavelet function supported on $[0,1]$ or on $[\frac{1}{2}, \frac{3}{2}]$.*

Proof. Denote by U_1 the restriction of $\tilde{V}_1$ to the interval $[0,1]$. That is, $U_1 = \{f|_{[0,1]} : f \in \tilde{V}_1\}$. We see that U_1 is a five-dimensional vector space. For notational simplicity, let

$$x_1 = \phi^1|_{[0,1]}, \qquad x_2 = \phi^1(\cdot + 1)|_{[0,1]},$$
$$x_3 = \phi^1(2\cdot +1)|_{[0,1]}, \qquad x_4 = \phi^1(2\cdot -1)|_{[0,1]}.$$

Note that $\operatorname{rank}\{\phi^0, x_1, x_2, x_3, x_4\} = 5$. Suppose that ψ^0 is a wavelet supported on $[0,1]$. Then ψ^0 is orthogonal to x_1, x_2, and ϕ^0 because of the orthogonality between wavelets and scaling functions. Also, from (4.25) applied to functions in $\tilde{V}_1$, we see that ψ^0 is orthogonal to $\phi^1(2\cdot +1)$ and $\phi^1(2\cdot -1)$ (and hence to x_3 and x_4) since ψ^0 vanishes at both 0 and 1. Thus $\psi^0 \in \{\phi^0, x_1, x_2, x_3, x_4\}^\perp = \{0\}$. This cannot be, so there is no such wavelet ψ^0.

Now suppose that ψ^0 is supported on $[\frac{1}{2}, \frac{3}{2}]$ and let

$$y_1 = \phi^0|_{[0,\frac{1}{2}]}, \qquad y_2 = \phi^1|_{[0,\frac{1}{2}]},$$
$$y_3 = \phi^0\left(\cdot + \frac{1}{2}\right)|_{[0,\frac{1}{2}]}, \qquad y_4 = \phi^1\left(\cdot + \frac{3}{2}\right)|_{[0,\frac{1}{2}]},$$
$$z_1 = \psi^0\left(\cdot + \frac{1}{2}\right)|_{[0,\frac{1}{2}]}, \qquad z_2 = \psi^0(\cdot + 1)|_{[0,\frac{1}{2}]}.$$

Note that the above are AFIFs on $[0, \frac{1}{2}]$. Since $f_{\bar{y}_2}(2\cdot)$ is orthogonal to $f_{\bar{y}_1}(2\cdot)$, it follows from (3.8) that if $s_0 \neq 0$ then y_1 and y_2 are linearly independent. A similar argument shows that if $s_1 \neq 0$ then y_3 and y_4 are linearly independent. Orthogonality between ψ^0 and $\phi^i(\cdot - j)$ dictates that z_1 must be orthogonal to y_3 and y_4. This implies that z_1 is orthogonal to all functions in $\tilde{V}_1|_{[0,\frac{1}{2}]}$ vanishing at $\frac{1}{2}$ since this space is three dimensional with basis $\{f_{\bar{y}_i}(2\cdot)\}_{i=0}^3$. Similarly, z_2 is orthogonal to all functions in $\tilde{V}_1|_{[0,\frac{1}{2}]}$ vanishing at 0. Thus it follows from (4.25) that ψ^0 is a multiple of $\phi^1(2\cdot -1)$. Now, ϕ^1 does not vanish at 1, but we require that $\langle \phi^1, \psi^0 \rangle = 0$. This cannot be, since among the functions $\phi^i(2 \cdot -j)$, of which ϕ^1 is a linear combination, the only one that does not vanish at 1 is $\phi^1(2 \cdot -1)$.

For the case of $s_0 = 0$ (or similarly for $s_1 = 0$), we can apply the quadrature formulas (2.9). Solving $p(0, s_1) = 0$ yields $s_1 = \sqrt{7} - 3$, which implies that y_3 and y_4 are linearly independent. Thus z_1 is still a multiple of the left half of $\phi^1(2\cdot)$. Rescale ψ^0 so that $z_1(\frac{1}{2}) = 1$; hence z_2 must have the form $f_{[1,r,0]}(2\cdot)$, where r is yet to be determined. Since $s_0 = 0, \phi^0(x)|_{[0,\frac{1}{2}]} = f_{[0,\frac{1}{2},1]}$, which is a line. From (2.9),

$$\langle \psi^0(\cdot + 1), \phi^0 \rangle = \langle z_2, y_1 \rangle = \frac{2 + 12r - 6s_1 + s_1^2}{6(s_1 - 2)(s_1 - 4)}.$$

In order for this to be equal to 0, $r = \frac{s_1}{2} - \frac{s_1^2}{12} - \frac{1}{6} = \sqrt{7} - 3$. We must also have $\langle \psi^0, \phi^1 \rangle = 0$. From the above remarks, we find that $\psi^0|_{[\frac{1}{2},1]} = f_{[0,p,1]}(2 \cdot -1)$, while $\phi^1|_{[\frac{1}{2},1]} = f_{[p,\frac{1-s_1+p+2s_1p}{2},1]}(2\cdot -1)$ and $\phi^1|_{[1,\frac{3}{2}]} = f_{[1,\frac{1+q}{2},q]}(2\cdot -2)$, where from (3.5) $q = (3s_1^2 + 2s_1 - 2)/(4s_1 + 8) = \sqrt{7}-3$ and from (3.4) $p = (1 - s_1^2)/(s_1^2 - 4) = (\sqrt{7} - 4)/6$. From (2.9), we find, after change of variables,

$$\begin{aligned}\langle \psi^0, \phi^1 \rangle &= \frac{1}{2}\int_0^1 f_{[0,p,1]}(x) f_{[p,\frac{1-s_1+p+2s_1p}{2},1]}(x)dx + \frac{1}{2}\int_0^1 f_{[1,r,0]}(x) f_{[1,\frac{1+q}{2},q]}(x)dx \\ &= \frac{(7 + 5\sqrt{7})(6r - 4\sqrt{7} + 25)}{756}.\end{aligned}$$

Thus in order for the above integral to be equal to 0, $r = (4\sqrt{7} - 25)/6$, which cannot be. □

This leads to the following.

THEOREM 4.3. *For any pair (s_0, s_1) with $p(s_0, s_1) = 0, |s_0| < 1$, and $|s_1| < 1$, let Ψ be a wavelet. Then the support of one component of Ψ must be of length $\geq \frac{3}{2}$, while the support of the other component must be of length ≥ 2.*

Proof. Lemma 4.2 shows that both components of Ψ must have support lengths of at least $\frac{3}{2}$. It is easy to see that not both may be supported on $[-1, \frac{1}{2}]$ since in that case either the pieces supported on $[-1, -\frac{1}{2}]$ or those on $[0, \frac{1}{2}]$ must be linearly dependent. A rotation of the components could then be used to find a wavelet whose support length is 1 which would contradict Lemma 4.2. A similar argument shows that not both components of Ψ may have support $[-\frac{1}{2}, 1]$.

Suppose $\Psi = \binom{\psi^1}{\psi^2}$, where $\operatorname{supp} \psi^1 = [-1, \frac{1}{2}]$ and $\operatorname{supp} \psi^2 = [-\frac{1}{2}, 1]$. In this case, the coefficients for the expansion of Ψ in terms of the normalized $\Phi(2\cdot)$ have the form

$$d_{-2} = \begin{bmatrix} 0 & \delta_1 \\ 0 & 0 \end{bmatrix}, \qquad d_{-1} = \begin{bmatrix} \delta_2 & \delta_3 \\ 0 & \delta_6 \end{bmatrix},$$

$$d_0 = \begin{bmatrix} \delta_4 & \delta_5 \\ \delta_7 & \delta_8 \end{bmatrix}, \qquad d_1 = \begin{bmatrix} 0 & 0 \\ \delta_9 & \delta_{10} \end{bmatrix},$$

with all others being 0. The relation $\langle \psi^0, \phi^1 \rangle = 0$, which must be satisfied, implies that $(2s_0 - 2q_1^1 + 1)\delta_5 = 0$. If δ_5 is not to be 0, then this is equivalent to $q_1^1 = \frac{1}{2} + s_0$. However, equation (3.5) also gives q_1^1 in terms of s_0 and s_1. Equating right-hand sides and simplifying gives

$$0 = \tilde{p}(s_0, s_1) = s_0^2 s_1 - 4s_0 s_1^2 + s_1^3 + 6s_0^2 + 6s_1^2 - 10s_1 - 12.$$

Of course, $p(s_0, s_1) = 0$ must still be satisfied, so

$$\begin{aligned} 0 &= -p(s_0, s_1) - \tilde{p}(s_0, s_1) \\ &= -2s_1^4 + 7s_1^3 s_0 + 7s_1 s_0^3 - 2s_0^4 - 6s_1^3 - 19s_1^2 s_0 - 14s_1 s_0^2 - 7s_0^3 \\ &\quad + 22s_1^2 + 14s_1 s_0 + 22s_0^2 - 12s_1 - 2s_0 + 4 \\ &= (\tilde{q}_2 \tilde{d}_2 + \tilde{r}_2)\tilde{d}_1 + \tilde{r}_1, \end{aligned}$$

where $\tilde{d}_1 = 3 - s_1$, $\tilde{d}_2 = 3 - 2s_0 - s_1$, $\tilde{r}_1 = -2s_0^4 + 14s_0^3 - 20s_0^2 + 58s_0 - 158$, $\tilde{r}_2 = -\frac{19}{8}s_1^3 - \frac{31}{8}s_1^2 - \frac{13}{8}s_1 + \frac{255}{8}$, and $\tilde{q}_2 = \frac{35}{8}s_1^2 + \frac{7}{4}s_1 s_0 + \frac{7}{2}s_0^2 + \frac{11}{4}s_1 - \frac{7}{4}s_0 + \frac{59}{8}$. It is easy to check that $\tilde{d}_1$, $\tilde{d}_2$, $\tilde{r}_1$, $\tilde{r}_2$, and $\tilde{q}_2$ are all positive for $(s_0, s_1) \in (-1,1)^2$, so the above equation has no valid solution, and it follows that $\delta_5 = 0$. A similar argument shows that $\delta_6 = 0$.

Now, $\langle \psi^0, \psi^1 \rangle = 0$ implies that $\delta_4 \delta_7 = 0$, which is satisfied only if $\delta_4 = 0$ or $\delta_7 = 0$. In the former case, ψ^0 is supported on $[-1, 0]$, and in the latter case, ψ^1 is supported on $[0, 1]$. By Lemma 4.2, neither case is possible. □

Although it has been shown by Daubechies [5] that with one scaling function, compactly supported, continuous wavelets cannot be symmetric, this is not the case with two scaling functions. In order to obtain wavelets supported on $[0, 2]$ that are symmetric or antisymmetric with respect to 1, it must be that $s_0 = s_1$. To see this, suppose that ψ is such a wavelet with support $[0, 2]$. Since $\psi \in \tilde{V}_1$, it follows using (2.2) and (2.3) that for $x \in [0, \frac{1}{2}], \psi(\frac{x}{2}) = a_0 x + s_0 \psi(x)$ and $\psi(2 - \frac{x}{2}) = -a_8 x + s_1 \psi(2 - x)$. The symmetry of ψ implies that $\psi(\frac{1}{2^n}) = \pm\psi(2 - \frac{1}{2^n})$, where the plus sign is used if ψ is symmetric with respect to 1 and the minus sign if it is antisymmetric with respect to 1. Comparing these equations we see that either $s_0 = s_1$, which implies that both s values are equal to $-\frac{1}{5}$, or ψ is linear on the intervals $[0, \frac{1}{2}]$ and $[\frac{3}{2}, 2]$. In the latter case, one can apply the same argument to $[\frac{1}{2}, 1]$ and $[1, \frac{3}{2}]$. These cannot both be lines, since in that case they might as well have been constructed for $s_0 = s_1 = 0$, contradicting Theorem 3.6. Suppose that one wavelet ψ^s is symmetric while the other ψ^a is antisymmetric with respect to 1 and set $\bar{y}_1^0 = [0, a_1^0, a_2^0]$, $\bar{y}_2^0 = [a_2^0, a_3^0, a_4^0]$, $\bar{y}_3^0 = [a_4^0, a_3^0, a_2^0]$, $\bar{y}_4^0 = [a_2^0, a_1^0, 0]$, $\bar{y}_1^1 = [0, a_1^1, a_2^1]$, $\bar{y}_2^1 = [a_2^1, a_3^1, 0]$, $\bar{y}_3^1 = [0, -a_3^1, -a_2^1]$, and $\bar{y}_4^1 = [-a_2^1, -a_1^1, 0]$; then equations similar to (4.2)–(4.5) ((4.6) is automatically satisfied) give $a_1^0 = 1$, $a_2^0 = -30$, $a_3^0 = 111$, $a_4^0 = -100$, $a_1^1 = 1$, $a_2^1 = -30$, and $a_3^1 = 81$. Here $\|\psi^s\|_{L^2} = \sqrt{2}\|\psi^a\|_{L^2}$. If ψ^s and ψ^a normalize to 1, then the corresponding matrices in (4.10) are

$$D_0 = \begin{bmatrix} -1/20 & -3\sqrt{2}/20 \\ -\sqrt{2}/20 & -3/10 \end{bmatrix}, \qquad D_1 = \begin{bmatrix} 9/20 & -1/\sqrt{2} \\ 9\sqrt{2}/20 & 0 \end{bmatrix},$$

$$D_2 = \begin{bmatrix} 9/20 & -3\sqrt{2}/20 \\ -9\sqrt{2}/20 & 3/10 \end{bmatrix}, \qquad D_3 = \begin{bmatrix} -1/20 & 0 \\ \sqrt{2}/20 & 0 \end{bmatrix}.$$

These wavelets are plotted in Figure 5. Consequently, ψ^s and ψ^a individually exhibit linear phase.

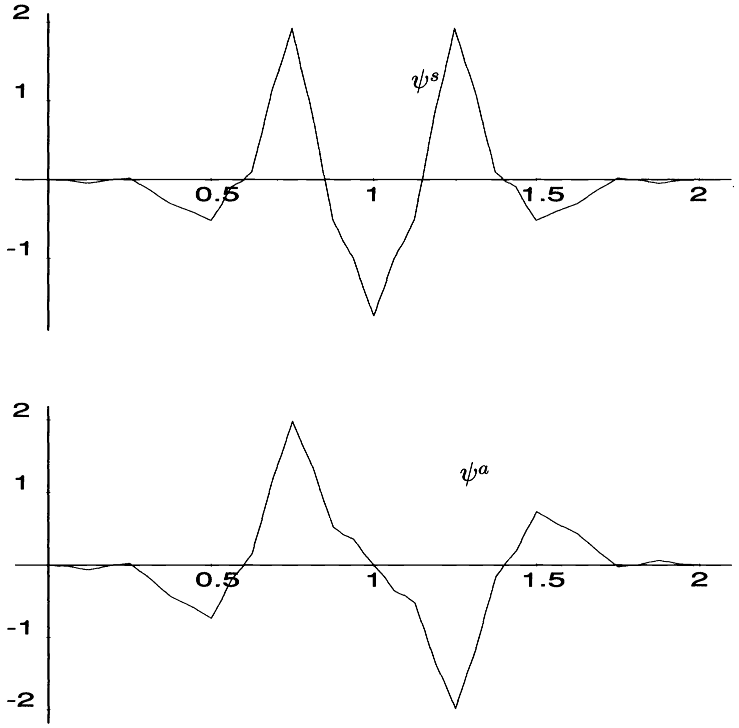


FIG. 5.

A similar computation where both wavelets are assumed to be symmetric with respect to 1 and supported on $[0,2]$ yields no solution.

Finally, we note that the multiresolution analysis arising from AFIFs is well suited for compact intervals. In fact, if we let $V_k' = \tilde{V}_k \cap L^2[0,1]$ and $\tilde{\phi}^i_{k,j} = \phi^i_{k,j}|_{[0,1]}$, then $\{V_k'\}$ provides a multiresolution analysis for $[0,1]$ and $\tilde{\phi}^i_{k,j}$, $j \in \mathbb{Z}$, $i = 0,1$, is an orthogonal basis for V_k'. To see what to do with the wavelets, write $\psi^i_j = \psi^i|_{[j-1,j]}$ and $\phi^i_j = \phi^i|_{[j-1,j]}$, $i = 0,1$, $j = 1,2$. We now rotate the wavelets, i.e.,

$$\psi^{+,0} = a\psi^0 + b\psi^1, \quad \psi^{+,1} = -b\psi^0 + a\psi^1 \tag{4.26}$$

with $|a|^2 + |b|^2 = 1$ so that

$$\langle \psi^{+,0}_j, \phi^1_j \rangle = 0, \qquad j = 1,2. \tag{4.27}$$

To see that this can be done, note that (4.27) is equivalent to

$$a\langle \psi^0_1, \phi^1_1 \rangle + b\langle \psi^1_1, \phi^1_1 \rangle = 0 \tag{4.28}$$

and

$$a\langle \psi^0_2, \phi^1_2 \rangle + b\langle \psi^1_2, \phi^1_2 \rangle = 0. \tag{4.29}$$

However, the fact that $\langle \psi^0, \phi^1 \rangle = \langle \psi^0_1, \phi^1_1 \rangle + \langle \psi^0_2, \phi^1_2 \rangle = 0$ and $\langle \psi^1, \phi^1 \rangle = \langle \psi^1_1, \phi^1_1 \rangle + \langle \psi^1_2, \phi^1_1 \rangle = 0$ shows us that (4.28) and (4.29) are not independent, which allows us to construct the desired rotation. With $\psi^{+,0}$ and $\psi^{+,1}$ constructed as in (4.26), set $\tilde{\psi}^0_{k,j} = \psi^{0,+}_{k,j}|_{[0,1]}$ and

$$\tilde{\psi}^1_{k,j} = \begin{cases} 0 & \text{if } \psi^{+,1}_{k,j} \cap [0,1]^c \neq \emptyset, \\ \psi^{+,1}_{k,j} & \text{otherwise.} \end{cases} \tag{4.30}$$

Then the nonzero components of $\{\tilde{\Psi}_{k,j}\}_{j\in\mathbb{Z}}$ form an orthogonal basis for $W'_k, k \geq 0$, where $V'_{k+1} = V'_k \oplus W'_k$. This leads to the following.

THEOREM 4.4. *$\{\tilde{\phi}^i_{k,j} = \phi^i_{k,j}|_{[0,1]} : k \geq 0,\ i = 0, 1, -i \leq j \leq 2^k - 1\}$ is an orthogonal basis for $V'_k = V_k \cap L^2[0,1]$, while $\{\tilde{\psi}^i_{k,j}, k \geq 0, i = 0, 1, i-1 \leq j \leq 2^k - (i+1)\}$ forms an orthogonal basis for W'_k. Furthermore, $V'_0 \bigoplus_{k\geq 0} W'_k = L^2[0,1]$.*

For the case where $s_0 = s_1 = -\frac{1}{5}$, it is easy to see that $\tilde{\psi}^0_{k,j}$ are just the symmetric wavelets restricted to $[0,1]$, i.e., $\tilde{\psi}^0_{k,j} = \psi^s_{k,j}|_{[0,1]}$, while $\tilde{\psi}^1_{k,j} = \psi^a_{k,j}$ if the support of $\psi^a_{k,j} \subset [0,1]$ and 0 otherwise.

5. Scaling by other integers. Many of the results of the previous sections can be used to produce scaling functions and wavelets satisfying dilation equations of the forms

$$\Phi(x) = \sum C_i \Phi(Nx - i)$$

and

$$\Psi(x) = \sum D_i \Phi(Nx - i)$$

with $N > 2$. Note that in this case the matrices D_i will in general not be square matrices. We begin by considering the $N+1$-dimensional basis $\{f_{\bar{y}_i}\}_{i=0}^N$ spanning S_0, where $\bar{y}_i = e_i$, $0 < i \leq N$, $\{e_i\}_{i=0}^N$ is the standard basis in R^{N+1}, and the last vector $\bar{y}_0$ is given by $\bar{y}_0 = [1, q_1, q_2, \ldots, q_{N-1}, 0]$. What is needed is to adjust q_i, $1 \leq i \leq N-1$, and s_j, $0 \leq j \leq N$, so that $f_{\bar{y}_0}$ is nonzero and orthogonal to $f_{\bar{y}_i}$, $1 \leq i \leq N$. Once this has been accomplished, orthogonal scaling functions ϕ^i, $0 \leq i \leq N-1$, can be obtained by applying the Gram–Schmidt procedure to the set $\{f_{\bar{y}_i}\}_{i=1}^N$. The functions ϕ^i, $0 \leq i \leq N-2$, obtained from the functions $\{f_{\bar{y}_i}\}_{i=1}^{N-1}$ will be continuous and supported on $[0,1]$ since each $f_{\bar{y}_i}$, $1 \leq i \leq N-1$, vanishes at 0 and 1. The last function ϕ^{N-1} can be obtained by subtracting from $f_{\bar{y}_N}$ its projection onto the subspace spanned by $\{\phi^i\}_{i=0}^{N-2}$ then piecing it together continuously with $f_{\bar{y}_0}$ as was done in the case $N = 2$ in §3. It follows from Theorem 5.3 below that it is sufficient to consider only a basis for S_0 of this type.

Since there are N orthogonality relations and $2N - 1$ unknowns (q_i, $1 \leq i \leq N-1$, and s_j, $0 \leq j \leq N-1$), it may be possible to impose other desirable conditions besides orthogonality and still obtain the required basis $\{\phi^i\}_{i=0}^{N-1}$. If $\Phi^*(x) = (\phi^1, \phi^2, \ldots, \phi^{N-1})$, then Φ will satisfy $\Phi(x) = \sum_{i=0}^{2N-1} C_i \Phi(Nx - i)$. $\Psi(x)$ may now be obtained from $\Phi(x)$ using Theorem 4.1 or the orthogonality equations (2.9) and solving as in §4.

In order to compute the orthogonality relations $\langle f_{\bar{y}_i}, f_{\bar{y}_0}\rangle = 0$, $1 \leq i \leq N$, $\lambda^i_j(x) = a^i_j x + b^i_j$, $0 \leq i \leq N$, $0 \leq j \leq N-1$, need to be computed. From (2.1)–(2.3) and (2.5), we find that $a^i_j = \delta_{i-1,j} - \delta_{i,j} - s_j \delta_{N,i}$ and $b^i_j = \delta_{i,j}$ for $1 \leq i \leq N$ and $0 \leq j \leq N-1$ with $q_0 = 1$ and $q_N = 0$. Also, $a^0_j = q_{j+1} - q_j + s_j$ and $b^0_j = q_j - s_j$ for $0 \leq j \leq N-1$. If $S_1 = \sum_{i=0}^{N-1} s_i$ and $S_2 = \sum_{i=0}^{N-1} i s_i$, it follows from (2.7) and (2.8) that $m^i_0 = 1/(N - S_1)$, $m^i_1 = ((N - S_1)i + S_2)/((N - S_1)(N^2 - S_1))$ for $1 \leq i \leq N-1$,

$m_0^N = (1-S_1)/(2(N-S_1))$,

$$m_1^N = \frac{N(3N-1)+(1-5N)S_1+3(1-N)S_2+2S_1^2}{6(N-S_1)(N^2-S_1)},$$

$$m_0^0 = \frac{1-S_1+2\sum_{i=1}^{N-1} q_i}{2(N-S_1)},$$

and

$$m_1^0 = \frac{S_1(S_1-(N+1))-3(N-1)S_2+N+6\sum_{i=1}^{N-1} q_i((N-S_1)i+S_2)}{6(N-S_1)(N^2-S_1)}.$$

With the above moments, (2.9) yields

$$\begin{aligned} I_{n,0} &= \langle f_{\bar{y}_n}, f_{\bar{y}_0}\rangle \\ &= \Bigg((s_{n-1}-s_n)m_1^0 + (m_0^n - m_1^n)\sum_{i=0}^{N-1} s_i(q_i - s_i) \\ &\quad + m_1^n \sum_{i=1}^{N-1} s_{i-1}q_i + s_n m_0^0 - \frac{s_{n-1}+2s_n}{6} \\ &\quad + \frac{1}{6}(q_{n-1}+4q_n+q_{n+1})\Bigg) \Bigg/ \left(N - \sum_{i=0}^{N-1} s_i^2\right), \qquad 1 \le n \le N-1, \end{aligned} \tag{5.1}$$

and

$$\begin{aligned} I_{N,0} = \Bigg(& m_1^0\left(s_{N-1} - \sum_{i=0}^{N-1} s_i^2\right) + \left(m_0^N - m_1^N - \frac{1}{6}\right)\sum_{i=0}^{N-1} s_i(q_i - s_i) \\ & + \sum_{i=1}^{N-1} s_{i-1}q_i(m_1^N - 1/3) + \frac{q_{N-1}-s_{N-1}}{6}\Bigg) \Bigg/ \left(N - \sum_{i=0}^{N-1} s_i^2\right). \end{aligned} \tag{5.2}$$

For $N=2$, (5.1) and (5.2) yield (3.2) and (3.3) with $p_1 = 0$. For $N=3$, (5.1) and (5.2) become more complicated and we shall restrict ourselves to considering s values that give symmetric or antisymmetric wavelets.

Just as when $N=2$, we require that the s values be arranged symmetrically, so s_0 must be the same as s_2. In this case, a one-parameter family of continuous, compactly supported, orthogonal scaling functions, each with linear phase, will be produced.

We proceed as before, examining what conditions s_0 must satisfy for compactly supported scaling functions to exist. Let Υ_0 be the space of vectors vanishing on both sides and Υ_1 be those vanishing on the left. Then Υ_0 is a two-dimensional space and has a basis of the form $\bar{x}_1 = [0,1,1,0]$ and $\bar{x}_2 = [0,1,-1,0]$. Let $\bar{x}_3 = [0,p_1,p_2,1]$ be a vector orthogonal to Υ_0 in Υ_1, so that $\{\bar{x}_1,\bar{x}_2,\bar{x}_3\}$ forms an orthogonal basis for Υ_1.

Now, if we can find $\bar{x}_4 = [1,q_1,q_2,0]$ orthogonal to Υ_1, then we easily generate the scaling functions

$$\phi^0 = \begin{cases} f_{\bar{x}_1} & \text{on } [0,1], \\ 0 & \text{elsewhere,} \end{cases} \quad \phi^1 = \begin{cases} f_{\bar{x}_2} & \text{on } [0,1], \\ 0 & \text{elsewhere,} \end{cases} \quad \text{and } \phi^2 = \begin{cases} f_{\bar{x}_3} & \text{on } [0,1], \\ f_{\bar{x}_4}(\cdot - 1) & \text{on } (1,2], \\ 0 & \text{elsewhere.} \end{cases}$$

If there exists no such $\bar{x}_4$, then there are no compactly supported scaling functions, as we can see from the following general result.

THEOREM 5.1. *Let V be an N-dimensional subspace of $C^0[0,1]$ such that there does not exist an orthonormal basis with $N-1$ of the basis functions vanishing at 0 and the remaining function vanishing at 1 or vice versa. Then there do not exist compactly supported $C^0(\mathbb{R})$ functions $\{\phi^j\}_{j=1}^{N-1}$ composed of linear combinations of basis elements of V and their integer translates constructed so that $\langle\phi^j(x),\phi^i(x-l)\rangle = \delta_{j,i}\delta_{0,l}$.*

Proof. The proof is by induction on N. The case $N = 3$ was established in Lemma 3.5. We only consider the step from $N = 3$ to $N = 4$, as the general argument is similar. For ease of notation set $x = \phi^1, y = \phi^2$, and $z = \phi^3$ and suppose that each is supported on a subset of $[0,M]$. Suppose V satisfies the hypotheses above. Then there are three cases we need to consider.

Case 1. One of the functions (say x) is supported on $[0,1]$, or x, y, and z can be transformed by a finite number of rotations and shifts so that this is the case. First, perform this transformation if necessary. We see that the components of y and z are restricted to a three-dimensional space, which has no orthonormal basis of vectors with two functions vanishing on the left and one vanishing on the right or vice versa. Thus by Lemmas 3.4 and 3.5, the scaling functions y and z cannot exist.

Case 2. All three functions have support $[0,2]$ or longer, and the leftmost components, $\{x_1,y_1,z_1\}$, have rank 1. In this case, we may begin to apply the rotate-and-shift strategy of Lemma 3.5 on y and z. If for some finite j, $y_1^{(j)}$ and $z_1^{(j)}$ are not linearly dependent, then we go to Case 3 below. Otherwise, the proof of Lemma 3.5 applies and we are done.

Case 3. All three functions have support $[0,2]$ or longer, and the leftmost components, $\{x_1,y_1,z_1\}$, have rank 2. Note that $\{x_1,y_1,z_1\}$ cannot have rank 3 because then the rightmost nonzero component of any of them would be orthogonal to all functions vanishing on the left, which we assumed was impossible. Here we apply an argument similar to the one found in Lemma 3.5 using two rotation maps instead of one. Many of the details in this proof are analogous to those in the proof of Lemma 3.5, so they are omitted here.

Since two among x_1, y_1, and z_1 must be linearly independent, we may assume that y_1 and z_1 are independent. We can rotate y and z so that y_1 and z_1 are orthogonal, and for an appropriate basis, we have

$$\begin{aligned} x &= ((x_{1,1},x_{1,2},0,0),x_2,x_3,\ldots,x_M),\\ y &= ((y_{1,1},0,0,0),y_2,y_3,\ldots,y_M),\\ z &= ((0,z_{1,2},0,0),z_2,z_3,\ldots,z_M), \end{aligned}$$

where $y_{1,1} > 0$ and $z_{1,2} > 0$. As before, we have a rotation r_1 which transforms x and y into

$$\begin{aligned} x' &= ((0,x'_{1,2},0,0),x'_2,\ldots,x'_M),\\ y' &= ((y'_{1,1},y'_{1,2},0,0),y'_2,\ldots,y'_M), \end{aligned}$$

leaving z unchanged, and a rotation r_2 which transforms x' and z into

$$\begin{aligned} x'' &= (0,x''_2,\ldots,x''_M),\\ z' &= ((0,z'_2,0,0),z'_2,\ldots,z'_M), \end{aligned}$$

leaving y' unchanged. We also have the shift map s, which takes x'' to

$$(x''_2,x''_3,\ldots,x''_M,0) = x''',$$

leaving y' and z' unchanged. The resulting vectors x''', y', and z' are themselves orthogonal scaling functions, and thus the leftmost components cannot have rank 3. Clearly, y'_1 and z'_1 are linearly independent, so their rank is 2. Now we iterate the map $s \circ r_2 \circ r_1$ to obtain the sequence $(x^{(j)}, y^{(j)}, z^{(j)}) = (s \circ r_2 \circ r_1)^j (x, y, z)$, which must have some limit point (X, Y, Z). Analogously to Lemma 3.5, both $\{y^{(j)}_{1,1}\}$ and $\{z^{(j)}_{1,2}\}$ are monotone increasing, from which it follows that $X_{k,1} = X_{k,2} = 0$ for $k = 1, 2, \ldots, M$ and hence that $X = 0$, which we know is not possible.

To proceed to higher N, we need only consider the analog of Case 3 above since the other cases are eliminated by the induction hypothesis. The analog of Case 3 is when all N functions have support $[0, 2]$ or longer and the dimension of the space spanned by the leftmost components of these functions is $N - 2$. In this case, $N - 2$ rotations are needed to reduce by iteration one of the functions to 0, thereby forcing a contradiction. □

Next, we investigate what conditions must be imposed on s_0 and s_1 in order that such a basis should exist. The vectors $[0, 1, 0, 0]$, $[0, 0, 1, 0]$, and $[0, 0, 0, 1]$ form a basis (not orthonormal) for Υ_1, so it suffices to check for orthogonality between $\bar{x}_3$ and each of these vectors. From (5.1) and (5.2), we find (where $\bar{y}_i, i = 0, 1, 2, 3$ are used), after some manipulation,

$$\begin{aligned} I_{1,0} = {} & 8s_0s_1q_2 - 108q_1 - 27q_2 + 12s_0 + 42s_1 + 60s_0q_1 - 48s_0q_2 - 8s_0s_1 \\ & - 24s_1q_1 - 24s_1q_2 + 76s_0^2 + 31s_1^2 + 8s_1s_0q_1 - 24s_0^3 - 6s_1^3 - 27 \\ & + 20s_0^2q_2 - 12s_0^2s_1 - 12s_0s_1^2 - s_1^2q_2 + 8s_1^2q_1 - 16q_1s_0^2 = 0, \end{aligned}$$

$$\begin{aligned} I_{2,0} = {} & 8s_0s_1q_2 - 27q_1 - 108q_2 + 12s_0 + 24s_1 - 48s_0q_1 + 60s_0q_2 - 8s_0s_1 \\ & - 24s_1q_1 - 24s_1q_2 + 28s_0^2 + 16s_1^2 + 8s_1s_0q_1 - 16s_0^2q_2 \\ & + 8s_1^2q_2 - s_1^2q_1 + 20q_1s_0^2 = 0, \end{aligned}$$

$$\begin{aligned} I_{3,0} = {} & -8s_0s_1q_2 - 27q_2 + 48s_0 + 12s_0q_1 + 12s_0q_2 - 4s_0s_1 + 24s_1q_1 \\ & + 42s_1q_2 - 50s_0^2 - 21s_1^2 - 8s_1s_0q_1 - 12s_0^2s_1q_2 - 32s_0^3 - 8s_1^3 \\ & + 76s_0^2q_2 - 16s_0^2s_1 - 16s_0s_1^2 + 31s_1^2q_2 + 16s_1^2q_1 + 28q_1s_0^2 \\ & - 24s_0^3q_2 + 8s_0^3s_1 + 6s_0^2s_1^2 + 4s_1^3s_0 - 6s_1^3q_2 - 12s_1^2s_0q_2 \\ & + 8s_0^4 + s_1^4 = 0. \end{aligned}$$

Solving the first two for q_1 and q_2 gives

$$q_1 = \frac{32s_0^3 - 76s_0^2 - 32s_0^2s_1 + 24s_0s_1 + 16s_0s_1^2 - 24s_1 + 36 - 56s_1^2 - 16s_1^3}{(6s_0 + 21s_1 + 45)(2s_0 - s_1 - 3)},$$

$$q_2 = \frac{40s_0^3 - 20s_0^2 - 4s_0^2s_1 + 12s_0s_1 + 20s_0s_1^2 - 36s_0 - 30s_1 - 19s_1^2 - 9 - 2s_1^3}{(6s_0 + 21s_1 + 45)(2s_0 - s_1 - 3)}.$$

If we substitute these expressions into the third equation and simplify, we get the desired relationship,

$$\begin{aligned} 0 = {} & 48s_0^4 - 256s_0^3 + 16s_0^2s_1^2 + 192s_0^2s_1 + 400s_0^2 + 16s_0s_1^3 - 96s_0s_1^2 \\ & - 272s_0s_1 - 192s_0 + s_1^4 + 16s_1^3 + 70s_1^2 + 48s_1 + 9. \end{aligned}$$

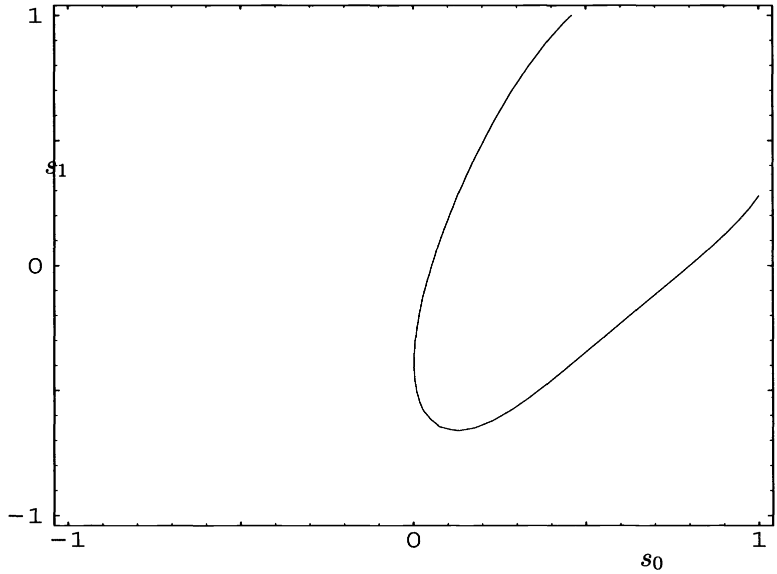


FIG. 6.

Let the polynomial on the right be denoted by $\hat{p}(s_0, s_1)$. The zero set of $\hat{p}$ is shown in Figure 6. From the above, the following is now clear.

THEOREM 5.2. *Continuous, compactly supported, orthogonal scaling functions* $\{\phi^j\}_{j=0}^2$ *can be constructed so that this set and its integer translates span* $\tilde{V}_0$ $(N = 3, s_2 = s_0)$ *if and only if* $|s_0| < 1$, $|s_1| < 1$, *and* $\hat{p}(s_0, s_1) = 0$.

As an example, we construct the scaling functions and wavelets generated from $s_0 = s_2 = \frac{3}{7}$, $s_1 = -\frac{3}{7}$. It is easy to check that these values do satisfy the conditions of Theorem 5.3, so the construction above gives the scaling functions shown in Figure 7. We then generate the wavelets, which are shown in Figure 8. The interpolation values for these functions are given in Tables 2 and 3. Table 4 gives the matrices in the dilation equation for the scaling functions obtained from $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4$, $(s_0, s_1, s_2$ arbitrary) given above.

For general N, we consider two special cases: Case A, $s_i = 0$, $0 \le i \le N-2$, $s_{N-1} = s$; and Case B, $s_i = s$, $0 \le i \le N-1$.

For Case A, $S_1 = s$ and $S_2 = (N-1)s$. Therefore,

$$m_0^0 = \frac{(1 - s + 2\sum_{i=1}^{N-1} q_i)}{2(N-s)}$$

and

$$m_1^0 = \frac{(s^2 - s(3N^2 - 5N + 4)) + N + 6\sum_{i=1}^{N-1} q_i((N-s)i + (N-1)s)}{6(N^2 - s)(N-s)},$$

and (5.1) yields

$$(5.3)\quad I_{n,0} = \frac{(m_0^n - m_1^n)s(q_{N-1} - s)) + \frac{1}{6}(q_{n-1} + 4q_n + q_{n+1})}{N - s^2}, \qquad 1 \le n \le N-2.$$

With the benefit of hindsight, we set $q_{N-1} = s$ and solve

$$(5.4)\qquad I_{n,0} = q_{n-1} + 4q_n + q_{n+1}, \qquad 1 \le n \le N-2,$$

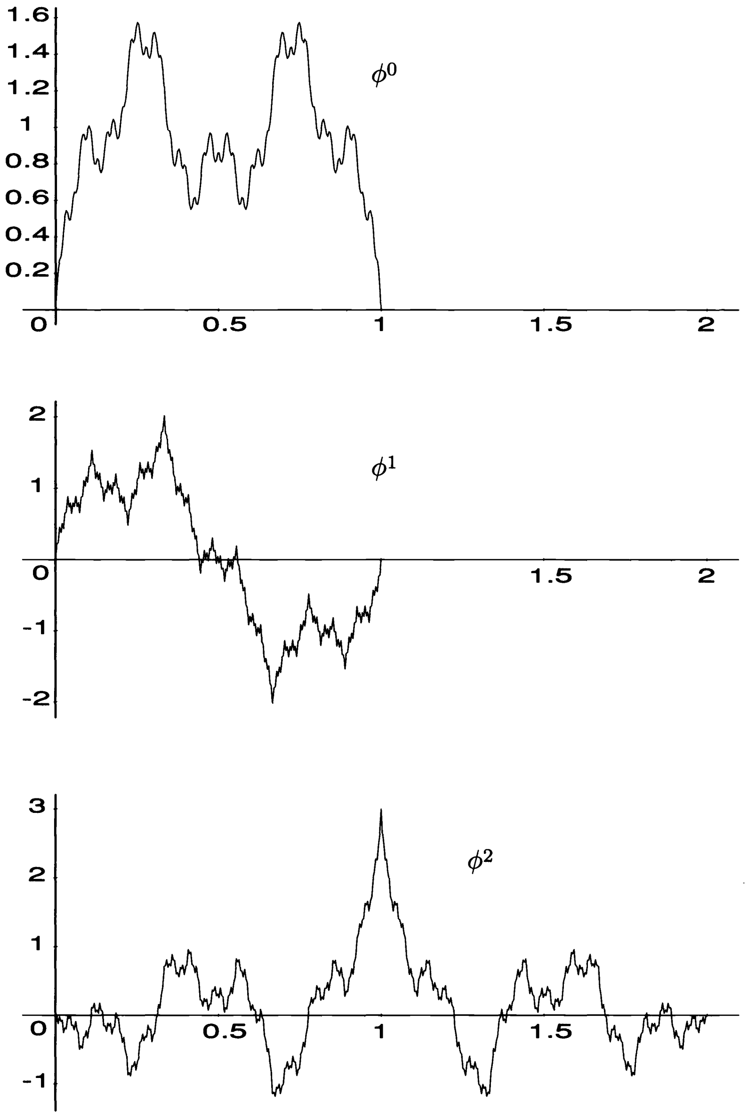


FIG. 7.

with boundary conditions $q_0 = 1$ and $q_{n-1} = s$. The solution is

$$q_n = \frac{U_{n-1}s}{U_{N-2}} + \frac{U_{N-n-2}}{U_{N-2}},$$

where U_i is the Chebyshev polynomial of the second kind evaluated at $x = -2$, i.e., $U_i = (\lambda^{i+1} - \lambda^{-(i+1)})/(\lambda - \lambda^{-1})$ with $\lambda = -2 + \sqrt{3}$. Using (5.4), we also find that

$$\sum_{n=1}^{N-1} q_n = \frac{1}{6}(5q_{N-1} + q_{N-2} + q_1 - 1) \tag{5.5}$$

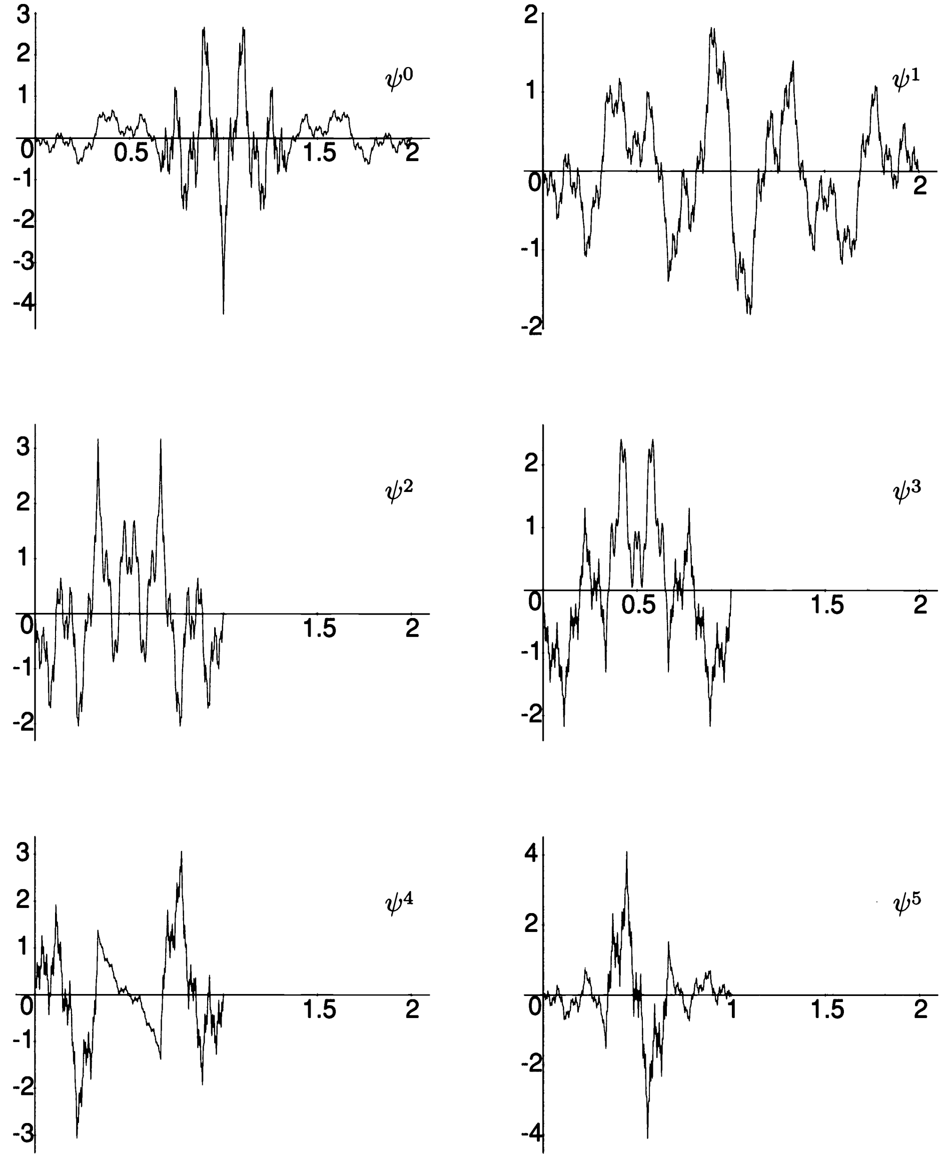


FIG. 8.

TABLE 2.

	0	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	$\frac{5}{9}$	$\frac{2}{3}$	$\frac{7}{9}$	$\frac{8}{9}$	1	$\frac{10}{9}$	$\frac{11}{9}$	$\frac{4}{3}$	$\frac{13}{9}$	$\frac{14}{9}$	$\frac{5}{3}$	$\frac{16}{9}$	$\frac{17}{9}$	2
ϕ_1	0	$\frac{32\sqrt{2}}{49}$	$\frac{46\sqrt{2}}{49}$	$\frac{6\sqrt{2}}{7}$	$\frac{24\sqrt{2}}{49}$	$\frac{24\sqrt{2}}{49}$	$\frac{6\sqrt{2}}{7}$	$\frac{46\sqrt{2}}{49}$	$\frac{32\sqrt{2}}{49}$	0	0	0	0	0	0	0	0	0	0
ϕ_2	0	$\frac{160\sqrt{2}}{147}$	$\frac{50\sqrt{2}}{147}$	$\frac{10\sqrt{2}}{7}$	$\frac{-20\sqrt{2}}{147}$	$\frac{20\sqrt{2}}{147}$	$\frac{-10\sqrt{2}}{7}$	$\frac{-50\sqrt{2}}{147}$	$\frac{-160\sqrt{2}}{147}$	0	0	0	0	0	0	0	0	0	0
ϕ_3	0	$\frac{17}{147}$	$\frac{-128}{147}$	$\frac{5}{7}$	$\frac{32}{147}$	$\frac{121}{147}$	$\frac{-8}{7}$	$\frac{17}{147}$	$\frac{40}{147}$	3	$\frac{40}{147}$	$\frac{17}{147}$	$\frac{-8}{7}$	$\frac{121}{147}$	$\frac{32}{147}$	$\frac{5}{7}$	$\frac{-128}{147}$	$\frac{17}{147}$	0

TABLE 3.

	0	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	$\frac{5}{9}$	$\frac{2}{3}$	$\frac{7}{9}$	$\frac{8}{9}$
ψ_1	0	$\frac{17\sqrt{2}}{294}$	$\frac{-64\sqrt{2}}{147}$	$\frac{5\sqrt{2}}{14}$	$\frac{16\sqrt{2}}{147}$	$\frac{121\sqrt{2}}{294}$	$\frac{-4\sqrt{2}}{7}$	$\frac{-149\sqrt{2}}{147}$	$\frac{272\sqrt{2}}{147}$
ψ_2	0	$\frac{17\sqrt{6}}{294}$	$\frac{-64\sqrt{6}}{147}$	$\frac{5\sqrt{6}}{14}$	$\frac{16\sqrt{6}}{147}$	$\frac{121\sqrt{6}}{294}$	$\frac{-4\sqrt{6}}{7}$	$\frac{-44\sqrt{6}}{147}$	$\frac{104\sqrt{6}}{147}$
ψ_3	0	$\frac{2\sqrt{129}}{301}$	$\frac{-50\sqrt{129}}{301}$	$\frac{12\sqrt{129}}{43}$	0	0	$\frac{12\sqrt{129}}{43}$	$\frac{-50\sqrt{129}}{301}$	$\frac{2\sqrt{129}}{301}$
ψ_4	0	$\frac{-698\sqrt{43}}{2107}$	$\frac{422\sqrt{43}}{2107}$	$\frac{-60\sqrt{43}}{301}$	$\frac{12\sqrt{43}}{49}$	$\frac{12\sqrt{43}}{49}$	$\frac{-60\sqrt{43}}{301}$	$\frac{422\sqrt{43}}{2107}$	$\frac{-698\sqrt{43}}{2107}$
ψ_5	0	$\frac{586\sqrt{17049}}{39781}$	$\frac{-934\sqrt{17049}}{39781}$	$\frac{60\sqrt{17049}}{5683}$	0	0	$\frac{-60\sqrt{17049}}{5683}$	$\frac{934\sqrt{17049}}{39781}$	$\frac{-586\sqrt{17049}}{39781}$
ψ_6	0	$\frac{-7762\sqrt{5683}}{835401}$	$\frac{8182\sqrt{5683}}{835401}$	$\frac{-808\sqrt{5683}}{39781}$	$\frac{8\sqrt{5683}}{147}$	$\frac{-8\sqrt{5683}}{147}$	$\frac{808\sqrt{5683}}{39781}$	$\frac{-8182\sqrt{5683}}{835401}$	$\frac{7762\sqrt{5683}}{835401}$

	1	$\frac{10}{9}$	$\frac{11}{9}$	$\frac{4}{3}$	$\frac{13}{9}$	$\frac{14}{9}$	$\frac{5}{3}$	$\frac{16}{9}$	$\frac{17}{9}$	2
ψ_1	$-3\sqrt{2}$	$\frac{272\sqrt{2}}{147}$	$\frac{-149\sqrt{2}}{147}$	$\frac{-4\sqrt{2}}{7}$	$\frac{121\sqrt{2}}{294}$	$\frac{16\sqrt{2}}{147}$	$\frac{5\sqrt{2}}{14}$	$\frac{-64\sqrt{2}}{147}$	$\frac{17\sqrt{2}}{294}$	0
ψ_2	0	$\frac{-104\sqrt{6}}{147}$	$\frac{44\sqrt{6}}{147}$	$\frac{4\sqrt{6}}{7}$	$\frac{-121\sqrt{6}}{294}$	$\frac{-16\sqrt{6}}{147}$	$\frac{-5\sqrt{6}}{14}$	$\frac{64\sqrt{6}}{147}$	$\frac{-17\sqrt{6}}{294}$	0
ψ_3	0	0	0	0	0	0	0	0	0	0
ψ_4	0	0	0	0	0	0	0	0	0	0
ψ_5	0	0	0	0	0	0	0	0	0	0
ψ_6	0	0	0	0	0	0	0	0	0	0

TABLE 4

Dilation matrices for $n = 3$, s_0, s_1, s_2 arbitrary.

$$C_0 = \begin{pmatrix} \frac{2s_0+1-(p_1+p_2)}{2} & \frac{-(1+3(p_1-p_2)}{6} & 1 \\ \frac{1-(p_1+p_2)}{2} & \frac{6s_0-(1+3(p_1-p_2))}{6} & 1 \\ \frac{(p_1+p_2-1)(s_0-p_1)}{2} & \frac{(p_1-p_2+1/3)(s_0-p_1)}{2} & p_1 \end{pmatrix}$$

$$C_1 = \begin{pmatrix} \frac{2s_1+2-(p_1+p_2+q_1+q_2)}{2} & \frac{-(p_1-p_2+q_1-q_2)}{2} & 1 \\ \frac{p_1+p_2-(q_1+q_2)}{2} & s_1+1/3+\frac{p_1-p_2-(q_1-q_2)}{2} & -1 \\ \frac{(p_1+p_2)(s_1+1-p_2)-s_1-p_1(q_1+q_2)}{2} & \frac{(p_1-p_2)(1/3+s_1-p_2)+s_1/3-p_1(q_1-q_2)}{2} & p_2 \end{pmatrix}$$

$$C_2 = \begin{pmatrix} s_2+1/2-\frac{(q_1+q_2)}{2} & \frac{q_2-q_1+1/3}{2} & 0 \\ \frac{q_1+q_2-1}{2} & \frac{2s_2-1/3+q_1-q_2}{2} & 0 \\ \frac{1-p_1+s_2(p_1+p_2-1)-p_2(q_1+q_2)}{2} & \frac{4p_2-3p_1-1+3s_2(p_1-p_2+1/3)-3p_2(q_1-q_2)}{6} & 1 \end{pmatrix}$$

$$C_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \frac{s_0(q_1+q_2-1)-q_1(p_1+p_2)-q_2+1}{2} & \frac{3s_0(q_1-q_2-1/3)-3q_1(p_1-p_2)+3q_2-4q_1+1}{6} & q_1 \end{pmatrix}$$

$$C_4 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \frac{(s_1-q_1)(q_1+q_2-1)+q_2(1-p_1-p_2)}{2} & \frac{(s_1-q_1)(q_1-q_2-1/3)-q_2(p_1-p_2+1/3)}{2} & q_2 \end{pmatrix}$$

$$C_5 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \frac{(s_2-q_2)(q_1+q_2-1)}{2} & \frac{(s_2-q_2)(q_1-q_2-1/3)}{2} & 0 \end{pmatrix}$$

and

$$\sum_{n=1}^{N-1} nq_n = \frac{((5N-4)q_{N-1} + (N-1)q_{N-2} - 1)}{6}. \tag{5.6}$$

To obtain $I_{N-1,0} = 0$ with $q_{N-1} = s$, we find from (5.1) that

$$s\left(m_1^0 + m_0^0 - \frac{1}{3}\right) + \frac{1}{6}(q_{N-2} + s) = 0. \tag{5.7}$$

Likewise, from (5.2), we find that $I_{N,0} = 0$ only if $s(s-1)m_1^0 = 0$. Since $s \neq 0$ or 1, we find

$$m_1^0 = 0. \tag{5.8}$$

Substitute (5.5) and (5.8) into (5.7), and (5.5) and (5.6) into (5.8) to find that (5.7) and (5.8) are both equal to 0 if and only if s satisfies the equation

$$s^2 + (N+1)(U_{N-3} + 2U_{N-2})s + N = 0, \qquad N \geq 2. \tag{5.9}$$

This leads to the following.

THEOREM 5.3. *Suppose $N \geq 2, s_i = 0, 0 \leq i \leq N-2$, and $s_{N-1} = s = (-b + \sqrt{b^2 - N})/2$ with $b = (N+1)(U_{N-3} + 2U_{N-2})$ with $U_i = (\lambda^{i+1} - \lambda^{-(i+1)})/(\lambda - \lambda^{-1})$ with $\lambda = -2 + \sqrt{3}$. Then there exist continuous, compactly supported, orthogonal scaling functions $\{\phi^i\}_{i=0}^{N-1}$ supported on $[0,2]$ such that this set and its integer translates span $\tilde{V}_0$.*

A set of mother wavelets $\{\psi^j\}_{j=1}^{N(N-1)}$ may be constructed using Theorem 4.1 or the orthogonality equations (2.9).

For Case B, $S_1 = Ns$ and $S_2 = N(N-1)s/2$. Therefore,

$$m_0^n = \frac{1}{N(1-s)}, \quad m_1^n = \frac{s(N-2n-1)+2n}{2N(1-s)(N-s)}, \quad 1 \leq n \leq N-1,$$

$$m^N = \frac{1-Ns}{2N(1-s)}, \quad m_1^N = \frac{4Ns^2 - s(3N+4N+1) + 6N - 2}{12N(1-s)(N-s)}, \quad m_0^0 = \frac{1-Ns+2a}{2N(1-s)},$$

and

$$m_1^0 = \frac{2Ns^2 - s(3N^2 - 4N + 5) + 2 + 12(1-s)a + 6(N-1)sb}{12N(1-s)(N-s)},$$

where $a = \sum_{i=1}^{N-1} q_i$ and $b = \sum_{i=1}^{N-1} iq_i$. The orthogonality conditions become

$$\begin{aligned} I_{n,0} = 0 = {} & s[s^2(N^2 - N(2n-1)) - 2s(N^2 - n(N+1) + 1) + 3N - N^2 - 2n] \\ & + 4(N-s)sa \\ & + \frac{N(1-s)(N-s)}{3}(q_{n-1} + 4q_n + q_{n+1}), \qquad 1 \leq n \leq N-1, \end{aligned} \tag{5.10}$$

and

$$\begin{aligned} I_{N,0} = 0 = {} & m_1^0 s(1-Ns) + \left(m_0^N - m_1^N - \frac{1}{6}\right)s(a - Ns + 1) \\ & + sa\left(m_1^N - \frac{1}{3}\right) + \frac{q_{N-1} - s}{6} \end{aligned} \tag{5.11}$$

with $n = N-1$ in (5.10), and we find

(5.12)
$$a = \frac{s^3(N^2-3N)+4s^2+s(N^2-N-2)-\frac{N}{3}(1-s)(N-s)(q_{N-2}+4q_{N-1})}{4s(N-s)}.$$

Therefore, for $1 \le n \le N-2$, (5.10) becomes

$$0 = q_{n-1} + 4q_n + q_{n+1} - d_n, \tag{5.13}$$

where $d_n = (q_{N-2}+4q_{N-1}) + (6/N(N-s))s(sN-1)(N-n-1)$. With $\bar{q}^* = (q_1, q_2, \dots, q_{N-2})$, H the $N-2 \times N-2$ tridiagonal matrix with 4 on the diagonal and 1 on the off diagonal, and $\bar{c}^* = (d_1 - 1, d_2, \dots, d_{N-3}, d_{N-2} - q_{N-1})$, we find that

$$\bar{q} = H^{-1}\bar{c}. \tag{5.14}$$

It follows from $HH^{-1} = I$ that

$$H^{-1}_{n,m} = \begin{cases} -\frac{U_{n-1}U_{N-2-m}}{U_{N-3}}, 0 \le n \le m \le N-1, \quad N \ge 3, \\ H^{-1}_{n,m} = H^{-1}_{m,n}, 0 \le m \le n \le N-1. \end{cases} \tag{5.15}$$

Here U_i is as in Case A above. The equation above gives $q_i, 1 \le i \le N-2$, in terms of q_{N-2} and q_{N-1}. Since $H^{-1}_{N-2,i} = -U_{i-1}/U_{N-3}$, we find that $\sum_{i=1}^{N-2} H^{-1}_{N-2,i} = \frac{1}{6}(U_{N-2} - U_{N-3} - 1)/U_{N-3}$. Solving (5.14) for q_{N-2} then substituting in the above calculations yields

$$q_{N-2} = \frac{(4 - 2U_{N-3} - 6U_{N-2})q_{N-1} + 6 - \frac{36s(sN-1)}{N(N-s)}\sum_{i=1}^{N-2} U_{i-1}(N-n-1)}{5U_{N-3} + U_{N-2} - 1}.$$

To solve for q_{N-1}, sum equation (5.10) from $n = 0$ to $N-2$ to find

$$a = \frac{s^2(3\,N^2 - 3\,N) + s(3\,N^2 - 8\,N) + 6}{N(6\,sN - 12\,s + 6\,N)} - \frac{(q_{N-1} + q_1)\,(s-1)\,N}{6\,sN - 12\,s + 6\,N}.$$

Eliminate q_1 from this equation using (5.12), and then comparing it with (5.12) gives a solution for q_{N-1}. Using a symbolic manipulation language such as Maple to solve these equations and using the recurrence formula for Chebyshev polynomials of the second kind, we find

$$\begin{aligned} q_{N-2} = &-\frac{s^3N((21\,N^2 - 93\,N + 104\,)U_{N-3} + (6\,N^2 - 24\,N + 28\,)U_{N-4} - 6\,N^2 + 12\,N + 4)}{d} \\ &+ \frac{2\,s^2((45\,N^2 - 45\,N + 52\,)U_{N-3} + (12\,N^2 12\,N + 14\,)U_{N-4} + 6\,N^3 + 2)}{d} \\ &+ \frac{3\,sN((7\,N^2 - N - 30\,)U_{N-3} + (2\,N^2 - 8\,)U_{N-4} - 2\,N^2 + 4\,N) - 12\,N^3}{d} \end{aligned}$$

and

$$\begin{aligned} q_{N-1} = &\frac{-s^3N(4 + 3\,N^2 - 9\,N) - 2\,s^2(3\,N^3 - 2) + 3\,sN^2(N-3) + 6\,N^3}{d} \\ &+ \frac{U_{N-4}(s^3N(213\,N^2 - 369\,N + 388) - s^2(336\,N^2 - 336\,N + 388) - sN(213\,N^2 - 33\,N - 336))}{d} \\ &- \frac{U_{N-5}(s^2(90\,N^2 - 90\,N + 104) - s^3N(57\,N^2 - 99\,N + 104) + sN(57\,N^2 - 9\,N - 90))}{d}, \end{aligned}$$

where

$$\begin{aligned} d = 2\,N(s-N)(s(2 + U_{N-4}(168\,N - 194) + U_{N-5}(45\,N - 52)) \\ + N(168\,U_{N-4} + 45\,U_{N-5})). \end{aligned}$$

Note that d is nonzero for $-1 \leq s \leq 1$. What remains is to solve (5.11). If we multiply (5.10) by $n-1$, sum for $n = 1$ to $N-1$, and then solve for $\sum_{n=0}^{N-1} nq_n = b$, we find, after using (5.12),

$$b = \frac{s^2(N^3 - 3N^2 + 2N) - s(N^2 - 3N + 1) - N}{6\,N - 6\,s} + \frac{(2\,N-1)\;Nq_{N-1}}{6}.$$

Substituting the above equations in (5.11) and simplifying using the recurrence formula for the Chebychev polynomials yields

$$\begin{aligned} 0 = N(s^3(U_{N-4}(35\,N^2 - 12\,N^3 + 35\,N^2 - 36\,N - 7) \\ + U_{N-3}(-45\,N^3 + 130\,N^2 - 135\,N - 26) - N^2 - 1) \\ + s^2(U_{N-4}(-12\,N^3 + 45\,N^2 + 2\,N + 21) \\ + U_{N-3}(7\,N - 45\,N^3 + 168\,N^2 + 7\,N + 78) - 3\,N^2 + 2\,N + 3) \\ + -6\,sN((3\,U_{N-4} + 11\,U_{N-3})(N+1) + 1) + 6\,N^2)(s-1)(s+1)/d_1, \end{aligned}$$

which is equal to 0 for $|s| < 1$ when

$$\begin{aligned} p_N(s) = s^3(U_{N-4}(35\,N^2 - 12\,N^3 - 36\,N - 7) + U_{N-3}(130\,N^2 - 45\,N^3 - 135\,N - 26) - N^2 - 1) \\ + s^2(U_{N-4}(21 - 12\,N^3 + 2\,N + 45\,N^2) + U_{N-3}(7\,N + 168\,N^2 - 45\,N^3 + 78) + 3 + 2\,N - 3\,N^2) \\ - 6\,sN(3\,U_{N-4}(N+1) + 11\,U_{N-3}(N+1) + 1) + 6\,N^2 = 0. \end{aligned}$$

Here

$$\begin{aligned} d_1 = s(24\, + (2016\,N - 2328\,)U_{N-4} + (540\,N - 624\,)U_{N-5}) \\ + 2016\,U_{N-4}N + 540\,U_{N-5}N(s-N)^2. \end{aligned}$$

Note that d_1 is nonzero for $-1 \leq s \leq 1$. To see that the above cubic has a zero for $|s| < 1$, we evaluate the polynomial p_N at $s = \pm 1$. Thus,

$$p_N(1) = -2\;(N-1)^2\,((45\,N - 26)\,U_{N-3} + (12\,N - 7)\,U_{N-4} - 1)$$

and

$$p_N(-1) = 4\;(N+1)^2\,(1 + 7\,U_{N-4} + 26\,U_{N-3})\,.$$

Since $|U_{N-3}| > |U_{N-4}| \geq 0$ and $|U_{N-3}| \geq 1$ for $N \geq 3$ and since sign $U_{N-3} = (-1)^{N+1}$, we see that sign $p_N(1) = (-1)^N$, while sign $p_N(-1) = (-1)^{N+1}$ for $N \geq 3$.

It now follows from $p_N(0) > 0$ that there is a real root s_N of p_N with $|s_N| < 1$ and sign $s_N = (-1)^{N+1}$, $N \geq 3$.

Thus we have shown the following.

THEOREM 5.4. *For $N \geq 3$ and $s_i = s, 0 \leq i \leq N-1$, with s a real root of $p_N, |s| < 1$, there exist continuous, compactly supported, orthogonal scaling functions $\{\phi^i\}_{i=0}^{N-1}$ supported on $[0,2]$ such that this set and its integer translates span $\tilde{V}_0$.*

We now give some examples from the above theorem (here we have factored a constant out of $p_N(s)$). For $N = 3$, $p_3(s) = 9s^3 - 7s^2 + 15s - 1$, $q_1 = -\frac{4}{3}(9s^2 + 2s - 3)/((3s+5)(s-3))$, and $q_2 = \frac{1}{3}(18s^3 - 9s^2 - 22s - 3)/((3s+5)(s-3))$. For $N =$

4, $p_4(s) = 53s^4+3s^2+51s+1$, $q_3 = \frac{1}{4}(48s^3 - 25s^2 - 61s + 2)/((5s + 7)(s - 4))$, $q_2 = \frac{1}{2}(2s + 1)(3s + 1)/(5s + 7)$, and $q_1 = -\frac{3}{4}(4s^3 + 33s^2 + 9s - 10)/((5s + 7)(s - 4))$. Finally, for $N = 5$, we have $p_5(s) = 7153s^3 + 3061s^2 + 4595s - 25$.

Acknowledgments. The second and third authors would like to thank Ingrid Daubechies and Gil Strang for their encouragement and insights.

REFERENCES

[1] P. Auscher, *Wavelet bases for $L^2(\mathbb{R})$ with rational dilation factor*, in Wavelets and Their Applications, G. Beylkin et al., eds., Jones and Bartlett, Boston, 1992, pp. 439–452.

[2] M. F. Barnsley, *Fractal functions and interpolation*, Constr. Approx., 2 (1986), pp. 303–329.

[3] T. Bedford, *Hölder exponents and box dimension for self-affine fractal functions*, Constr. Approx., 5 (1989), pp. 33–48.

[4] C. K. Chui, *An Introduction to Wavelets,* Academic Press, Boston, 1992.

[5] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[6] C. de Boor, R. A. DeVore, and A. Ron, *The structure of finitely generated shift-invariant spaces $L^2(\mathbb{R}^d)$*, J. Funct. Anal., 119 (1994), pp. 37–78.

[7] G. Donovan, J. S. Geronimo, and D. P. Hardin, *Intertwining multiresolution analyses and the construction of piecewise polynomial wavelets*, SIAM J. Math. Anal., 27 (1996), to appear.

[8] J. S. Geronimo and D. P. Hardin, *Fractal interpolation surfaces and a related 2-D multiresolution analysis*, J. Math. Anal. Appl., 176 (1993), pp. 561–586.

[9] J. S. Geronimo, D. P. Hardin, and P. R. Massopust, *An application of Coxeter groups to the construction of wavelet bases in $\mathbb{R}^n$*, in Lecture Notes in Pure and Applied Mathematics 157: Contemporary Aspects of Fourier Analysis, W. U. Bray, P. S. Milojević, and C. V. Stanojević, eds., Marcel Dekker, New York, Basel, 1992, pp. 187–196.

[10] ———, *Fractal functions and wavelet expansions based on several scaling functions*, J. Approx. Theory, 78 (1994), pp. 373–401.

[11] T. N. T. Goodman, S. L. Lee, and W. A. Wang, *Wavelets in wandering subspaces*, Trans. Amer. Math. Soc., 338 (1991), pp. 639–654.

[12] T. N. T. Goodman and S. L. Lee, *Wavelets of multiplicity r*, Trans. Amer. Math. Soc., 342 (1994), pp. 307–324.

[13] D. P. Hardin, B. Kessler, and P. R. Massopust, *Multiresolution analyses based on fractal functions*, J. Approx. Theory, 71 (1992), pp. 104–120.

[14] R. Q. Jia and Z. Shen, *Multiresolution analysis and wavelets*, Proc. Edinburgh Math. Soc. (2), 37 (1994), pp. 271–300.

[15] L. Hervé, *Multi-resolution analysis of multiplicity d: Applications to dyadic interpolation*, Appl. Comput. Harmonic Anal., 1 (1994), pp. 299–315.

[16] S. Mallat, *Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.

[17] I. Meyer, *Ondelettes et Opérateurs*, Hermann, Paris, 1989.

[18] C. Micchelli, *Using the refinement equation for the construction of pre-wavelets* VI: *Shift invariant subspaces*, in Approximation Theory, Spline Functions, and Applications, S. P. Singh, ed., NATO ASI Series C, 356 (1992), pp. 213–222.

[19] M. B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael, eds., *Wavelets and Their Applications,* Jones and Bartlett, Boston, 1992.

[20] G. Strang and G. Fix, *An Analysis of the Finite Element Method*, Wellesley–Cambridge Press, Wellesley, MA, 1973.

[21] G. Strang and V. Strela, *Short wavelets and matrix dilation equations*, IEEE Trans. SP., 43 (1995), pp. 108–115.

[22] M. Vetterli, *Wavelets and filter banks for discrete-time signal processing*, in Wavelets and their Applications, M. B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael, eds., Jones and Bartlett, Boston, 1992.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1193–1211, Sepetmber 1996

CONTINUOUS DEPENDENCE ON INITIAL DATA FOR DISCONTINUOUS SOLUTIONS OF THE NAVIER–STOKES EQUATIONS FOR ONE-DIMENSIONAL, COMPRESSIBLE FLOW*

DAVID HOFF†

Abstract. We prove that discontinuous solutions of the Navier–Stokes equations for one-dimensional, compressible fluid flow depend continuously on their initial data. Perturbations in the different components are measured in various fractional Sobolev norms; L^2 bounds are then obtained by interpolation. This improves upon earlier results in which continuous dependence was known only in a much stronger topology, one inappropriately strong for the physical model. More generally, we derive a bound for the difference between exact and approximate weak solutions in terms of their initial differences and of the weak truncation error associated with the approximate solution.

Key words. continuous dependence, discontinuous solutions, Navier–Stokes equations

AMS subject classifications. 35Q30, 35R05, 65M15

1. Introduction.

We prove the continuous dependence on initial data of discontinuous solutions of the Navier–Stokes equations for one-dimensional, compressible fluid flow,

$$(1.1)\qquad \begin{cases} v_t - u_x = 0, \\ u_t + p(v,e)_x = \left(\frac{\varepsilon u_x}{v}\right)_x, \\ e_t + u_x p(v,e) = \frac{\varepsilon u_x^2}{v} + \lambda\left(\frac{T(e)_x}{v}\right)_x, \end{cases}$$

with Cauchy data

$$(1.2)\qquad (v,u,e)\big|_{t=0} = (v_0,u_0,e_0)$$

under assumptions consistent with the known existence theory. Here v, u, e, p, and T represent, respectively, the specific volume, velocity, specific internal energy, pressure, and temperature in the fluid, $t \geq 0$ is time, and $x \in \mathbb{R}$ is the Lagrangean coordinate (thus the lines $x =$ constant correspond to particle trajectories). ε and λ are positive viscosity and heat-conduction coefficients.

The key point of interest here is that the topology of continuous dependence is that of $L^2(\mathbb{R})$, which is particularly appropriate for the physical problem. By contrast, the only other known continuous dependence result, Theorem 4.1 of [2], is formulated in an exceptionally strong norm, one which dominates the *variation* in perturbations of the discontinuous quantity v. This norm appeared to be natural for certain technical reasons related to the fact that only incomplete smoothing occurs, owing to the degeneracy of the viscosity matrix and also to the particular rates of (partial) smoothing near the initial layer $t = 0$. On the other hand, this very strong norm is not at all appropriate from the physical point of view. For example, the mass measure is one

*Received by the editors March 21, 1994; accepted for publication (in revised form) February 16, 1995. This research was supported in part by NSF grant DMS-9201597.

†Department of Mathematics, Indiana University, Bloomington, IN 47405-5701.

of the physically observable quantities, and the density $1/v$ is its derivative. The result of [2] therefore requires that the local variation in the derivative of an observable quantity be small; and this we must regard as unsatisfactory. There is also a very practical reason for clarifying the issue of continuous dependence; this is the fact that error bounds for approximate solutions can be derived only in norms in which the problem is known to be well posed. In particular, for the system under consideration here, error bounds for certain finite difference approximations are derived in Zhao and Hoff [4]; the rates of convergence obtained are unrealistically low, however, precisely because they are formulated in a norm which is inappropriately strong.

The goal of the present paper is therefore to show that, under assumptions consistent with the known existence theory of [1] and [2], discontinuous solutions depend continuously on their initial values in L^2, which is clearly a more suitable norm for the physical problem. This result is stated in the theorem below and is proved in §§2 and 3. (Actually, we prove a more general result, in which the difference between an exact weak solution and an approximate weak solution is bounded in terms of the truncation error associated with the latter and the difference in initial values.) The key idea is to substitute an adjoint-equation analysis for the more usual approach based on direct L^2-energy estimates. The adjoint functions are estimated in (positive) fractional Sobolev norms; error bounds are obtained by duality in (negative) fractional Sobolev norms, and L^2 information is then extracted by interpolation. As we shall see, this device has the effect of splitting nonintegrable singularities at $t = 0$ into the integrable products of singularities at the initial and forward times.

The present paper generalizes the result of Hoff and Zarnowski [3], in which L^2-continuous dependence was proved for the isentropic/isothermal version of (1.1), which consists of the first two equations only and with $p = p(v)$. The full Navier–Stokes system discussed here presents a number of technical complexities and difficulties not encountered in the simpler case, owing particularly to the coupling of the square of the velocity gradient into the energy equation in (1.1).

We now give a precise formulation of our results. First, concerning the functions p and T appearing in (1.1), we assume that there is a compact rectangle $R = [\underline{v}, \bar{v}] \times [\underline{e}, \bar{e}]$ contained in the open, positive quadrant of v-e space, such that

$$p \in C^1(R), \qquad T \in C^2([\underline{e},\bar{e}]), \tag{1.3}$$

and that there is a positive constant C such that

$$\begin{cases} C^{-1} \le T'(e), \\ C^{-1}e \le T(e) \le Ce \end{cases} \tag{1.4}$$

for $e \in [\underline{e}, \bar{e}]$.

Next, in order to describe weak solutions having different values at $x = \pm\infty$, we introduce states $U_\pm = [v_\pm, u_\pm, e_\pm]^t$ with $(v_\pm, e_\pm) \in R$, and we construct a smooth function $\tilde{U}(x) = [\tilde{v}(x), \tilde{u}(x), \tilde{e}(x)]^t$ such that

$$\tilde{U}(x) = U_\pm, \qquad \pm x \ge 1. \tag{1.5}$$

Weak solutions $U = [v, u, e]^t$ on $\mathbb{R} \times [0, \bar{t}]$ with initial value U_0 and end-states $U_\pm$ should then satisfy the following minimal regularity conditions:

$$U - \tilde{U} \in C([0, \bar{t}\,]; L^2(\mathbb{R})), \tag{1.6}$$

$$U(\cdot,0) = U_0, \tag{1.7}$$

$$(v,e) \in R \quad \text{a.e.}, \tag{1.8}$$

$$u - \tilde{u}, e - \tilde{e} \in L^2((0,\bar{t}\,); H^1(\mathbb{R})). \tag{1.9}$$

We let X denote the test-function space

$$X([t_1,t_2]) = (L^1 \cap L^\infty \cap H^1 \cap C^1)(\mathbb{R} \times [t_1,t_2]),$$

and we give the following weak formulation of the system (1.1)–(1.2).

DEFINITION. *Let $U = [v,u,e]^t$ satisfy* (1.6)–(1.9) *above. Then U is a weak solution of* (1.1) *on $\mathbb{R} \times [0,\bar{t}\,]$ with initial value U_0 and end-states $U_\pm$ provided that, for all intervals $[t_1,t_2] \subseteq [0,\bar{t}\,]$ and all test functions $\varphi, \psi, \chi \in X([t_1,t_2])$,*

$$\begin{aligned} \mathcal{L}_1(t_1,t_2,\varphi;U) \equiv \int (v - \tilde{v})\varphi\, dx\Big|_{t_1}^{t_2} + \int_{t_1}^{t_2} \int [(\tilde{v} - v)\varphi_t + u\varphi_x]\, dxdt \\ = 0; \end{aligned} \tag{1.10}$$

$$\begin{aligned} \mathcal{L}_2(t_1,t_2,\psi;U) &\equiv \int (u - \tilde{u})\psi\, dx\Big|_{t_1}^{t_2} \\ &\quad + \int_{t_1}^{t_2} \int \left[(\tilde{u} - u)\psi_t - p(v,e)\psi_x + \frac{\varepsilon u_x \psi_x}{v}\right] dxdt \\ &= 0; \end{aligned} \tag{1.11}$$

(1.12)

$$\begin{aligned} \mathcal{L}_3(t_1,t_2,\chi;U) &\equiv \int (e - \tilde{e})\chi\, dx\Big|_{t_1}^{t_2} \\ &\quad + \int_{t_1}^{t_2} \int \left[(\tilde{e} - e)\chi_t + \left(u_x p(v,e) - \frac{\varepsilon u_x^2}{v}\right)\chi + \frac{\lambda T(e)_x \chi_x}{v}\right] dxdt \\ &= 0. \end{aligned}$$

The existence of weak solutions and their additional regularity properties will be discussed below.

Solutions and approximate solutions will be compared in various fractional-norm Sobolev spaces. To describe these, we let $\beta \in \mathbb{R}$ and $\varphi \in \mathcal{S}$ ($\mathcal{S}$ is the Schwartz class of smooth, rapidly decreasing functions), and we define

$$|\varphi|_\beta = \left[\int_{\mathbb{R}} (1 + |\xi|)^{2\beta} |\hat{\varphi}(\xi)|^2 d\xi\right]^{1/2},$$

where $\hat{}$ denotes the Fourier transform. We then take H^β to be the completion of $\mathcal{S}$ with respect to the norm $|\cdot|_\beta$. Observe that if $\beta_1 \leq \beta_2$, then $|\varphi|_{\beta_1} \leq |\varphi|_{\beta_2}$, so that $H^{\beta_2} \subseteq H^{\beta_1}$. $|\cdot|_0$ is the usual L^2-norm, which we shall always denote by $\|\cdot\|$. Standard properties of the Fourier transform show that, given a nonnegative integer k, there is a positive constant $C = C(k)$ such that

$$C^{-1}|\varphi|_k \leq \sum_{0\leq j\leq k} \left\| \frac{d^j\varphi}{dx^j} \right\| \leq C|\varphi|_k.$$

We now choose indices α, β, and δ satisfying

$$\left\{ \begin{array}{l} \beta \in (0,1), \\ 0 < \alpha < \delta < \min\{\beta, 1/2\}. \end{array} \right. \tag{1.13}$$

The difference between an exact weak solution U and an approximate solution $U^h = [v^h, u^h, e^h]^t$ will be measured in the norm given by

$$\begin{aligned} E(t) = \|\Delta v(\cdot,t)\| &+ |\Delta u(\cdot,t)|_{-\alpha} + |\Delta e(\cdot,t)|_{-\beta} \\ &+ t^{\delta/2}\|\Delta u(\cdot,t)\| + t^{\beta/2}\|\Delta e(\cdot, t)\|, \end{aligned} \tag{1.14}$$

where $\Delta v = v^h - v$, etc. The constants α, β, and δ will be fixed throughout.

Clearly, an estimate for $E(t)$ must involve not only the initial difference $E(0)$, but also some measure of the extent to which the approximate solution U^h fails to be an exact weak solution. This measure, which we call the weak truncation error, is essentially the norm of the functional $\mathcal{L}_1 \oplus \mathcal{L}_2 \oplus \mathcal{L}_3$ over the space X^3, suitably topologized. Specifically, given $(\varphi, \psi, \chi) \in X([t_1, t_2])^3$, we first define the norm

(1.15)

$$\begin{aligned} |\!|\!|(\varphi, \psi, \chi)|\!|\!|_{[t_1,t_2]} = \sup_{t_1\leq t\leq t_2} \Big[&\|\varphi(\cdot, t)\| + |\psi(\cdot,t)|_\alpha + |\chi(\cdot,t)|_\beta \\ &(t_2 - t)^{(1-\alpha)/2}\|\psi_x(\cdot,t)\| + (t_2 - t)^{(1-\beta)/2}\|\chi_x(\cdot,t)\| \Big] \\ &+ \left(\int_{t_1}^{t_2} \int \left[\psi_x^2 + \chi_x^2 + (t_2 - t)^{1-\alpha}\psi_t^2 + (t_2 - t)^{1-\beta}\chi_t^2 \right] dxdt \right)^{1/2}. \end{aligned}$$

Then given an approximate solution $U^h = [v^h, u^h, e^h]^t$ satisfying (1.6)–(1.9), we define the weak truncation error

(1.16)

$$Q_1(t_1, t_2; U^h) = \sup \frac{|\mathcal{L}_1(s_1, s_2, \varphi; U^h)| + |\mathcal{L}_2(s_1, s_2, \psi; U^h)| + |\mathcal{L}_3(s_1, s_2, \chi; U^h)|}{|\!|\!|(\varphi, \psi, \chi)|\!|\!|_{[s_1,s_2]}}.$$

The sup here is taken over intervals $[s_1,s_2] \subseteq [t_1, t_2]$ and over test functions $(\varphi, \psi, \chi) \in X([s_1,s_2])^3$. It is clear that the first three terms in the definition (1.15) of $|\!|\!|(\varphi, \psi, \chi)|\!|\!|$ are dual to the first three terms in the definition (1.14) of $E(t)$. The other terms in (1.15) appear for technical reasons and reflect anticipated rates of smoothing for the adjoint system corresponding to (1.1). We shall also have need for the following auxiliary error functionals:

$$Q_2(t_1, t_2; U^h)^2 = \sup \left[\limsup_{\eta\to 0} |\mathcal{L}_2(s_1, s_2, j_\eta * t^\delta \Delta u; U^h)| \right], \tag{1.17}$$

where $j_\eta(x,t)$ is the standard space–time mollifier and the outer sup is over intervals $[s_1, s_2] \subseteq (t_1, t_2)$; and

$$Q_3(t_1, t_2; U^h) = \sup |\mathcal{L}_3(s_1, s_2, t^{\beta/2}\chi; U^h)|, \tag{1.18}$$

where the sup is again taken over intervals $[s_1, s_2] \subseteq [t_1, t_2]$ and over test functions $\chi \in X([s_1, s_2])$ for which

(1.19)
$$\sup_{s_1 \leq t \leq s_2} [\|\chi(\cdot,t)\| + (s_2 - t)^{1/2}\|\chi_x(\cdot,t)\|] + \left[\int_{s_1}^{s_2} \int [\chi_x^2 + (s_2 - t)\chi_t^2] \; dxdt\right]^{1/2} \leq 1.$$

Finally, we define the total truncation error,

$$Q = Q_1 + Q_2 + Q_3. \tag{1.20}$$

We expect that, in most contexts, Q_1 will be the dominant term in Q. Observe that, if U^h is an *exact* weak solution with end-states $U_\pm$, then $Q(t_1, t_2; U^h) = 0$.

Next, we describe the regularity properties of weak solutions which will be required for our analysis. Besides the minimal conditions (1.6)–(1.9), we shall assume that there are positive constants C_0 and r such that

$$\sup_{0 \leq t \leq \bar{t}} \|U(\cdot,t) - \tilde{U}\|^2 + \int_0^{\bar{t}} \int (u_x^2 + e_x^2) \; dxdt \leq C_0; \tag{1.21}$$

$$\begin{cases} v_t, u_x, e_x \in C((0, \bar{t}\,] \; ; L^2(\mathbb{R})) \text{ and} \\ \|v_t(\cdot,t)\|, \|u_x(\cdot,t)\|, \|e_x(\cdot,t)\| \leq C_0^{1/2} t^{-1/4}; \end{cases} \tag{1.22}$$

$$\begin{cases} v_t, u_x \in L^\infty_{\text{loc}} ((0,\bar{t}] \; ; L^\infty(\mathbb{R})) \text{ and} \\ \displaystyle\int_0^{\bar{t}} t^{\min\{\delta,\beta/2\}} \|u_x(\cdot,t)\|^2_{L^\infty(\mathbb{R})} dt \leq C_0, \\ \displaystyle\int_{t_1}^{t_2} \|v_t(\cdot,t)\|_{L^\infty(\mathbb{R})} dt \leq C_0^{1/2} |t_2 - t_1|^r; \end{cases} \tag{1.23}$$

$$u_t \in L^2((\tau, \bar{t}); L^2(\mathbb{R})) \text{ for all } \tau > 0. \tag{1.24}$$

Some discussion of these assumptions is in order. First, the local existence theory of [1] shows that (even when p and T are somewhat more general than in (1.3) and (1.4)) a local solution exists and satisfies all but two of the conditions (1.6)–(1.12) and (1.21)–(1.24), with C_0 a multiple of the initial norm $\|U_0 - \tilde{U}\|^2 + [\text{Var}(v_0)]^2 + [\text{Var}(u_0)]^2$, which is assumed to be finite. The two exceptions are the continuity (1.6) of $e(\cdot, t)$ into $L^2(\mathbb{R})$ at $t = 0$ and the smoothing rate in (1.22) for $\|e_x(\cdot,t)\|$. The second of these is proved in [2, Thm. 3.2], but under the additional assumptions that p and T satisfy the conditions (1.3) and (1.4), that $e_0 \in BV$, and that

$$\lambda^{1/2} C_0 \|T''\|^2_{L^\infty([\underline{e}, \bar{e}])} \ll 1, \tag{1.25}$$

where C_0 is now a multiple of $\|U_0 - \tilde{U}\|^2 + [\text{Var}(U_0)]^2$. Once the smoothing rate $\|e_x(\cdot,t)\| \sim t^{-1/4}$ has been established, an easy energy argument can be applied to the third equation in (1.1) to obtain that $\int_0^{\bar{t}} \int_{\mathbb{R}} t^{1/2+\theta} e_t^2 \, dxdt < \infty$ for any $\theta > 0$ (this is the analogue for e of the estimates (1.17) and (1.20) in [1] for u). The continuity (1.6) of $e(\cdot,t)$ into L^2 at $t = 0$ then follows easily. We observe that the condition (1.25) is vacuous in the polytropic case that T is a constant multiple of e and is in any event rather mild since λ is presumably small. We also point out that the rates of smoothing in (1.22) for u_x and e_x are optimal, even for solutions of the heat equation with initial data in $L^2 \cap BV$.

Concerning the assumptions (1.23), we appeal to the result (1.4) of Theorem 1.1 in [1], which shows that, for any $\theta > 0$,

$$\sup_{0<t\leq \bar{t}} t^{1/2+\theta}(\|u_x(\cdot,t)\|_{L^\infty} + \|v_t(\cdot,t)\|_{L^\infty}) < \infty.$$

This provides more than adequate justification for the assumptions (1.23).

In the present paper, we simply assume that a weak solution U exists and satisfies the conditions (1.6)–(1.12) and (1.21)–(1.24); it will be unnecessary to make explicit reference to particular properties of U_0 or to relate U_0 to the constant C_0. The smallness condition (1.25), on the other hand, will be made an explicit hypothesis in our main theorem, stated below. It appears to be necessary in order to achieve the required estimates for solutions of the adjoint of the first variation of the energy equation in (1.1) (see Lemma 2.2 below).

We can now state our main result.

THEOREM. *Assume that p and T satisfy the conditions* (1.3)–(1.4) *and let $\tilde{U}$ be as described above in* (1.5). *Then there are positive constants C and ζ such that, if U is a weak solution of* (1.1) *with end-states $U_\pm$, satisfying* (1.6)–(1.12) *and* (1.21)–(1.24), *if U^h satisfies* (1.6), (1.8), (1.9), *and* (1.21)–(1.24), *and if*

$$\lambda^{1/2} C_0 \|T''\|^2_{L^\infty([\underline{e},\bar{e}])} < \zeta, \tag{1.26}$$

then for intervals $[t_1, t_2] \subseteq [0, \bar{t}]$,

$$\sup_{t_1\leq t\leq t_2} E(t) + \left[\int_{t_1}^{t_2}\int t^\delta \Delta u_x^2\, dxdt\right]^{1/2} \leq C\left[E(t_1) + Q(t_1,t_2;U^h)\right]. \tag{1.27}$$

The error functional $E(t)$ here is as defined above in (1.14), *with $\Delta u = u^h - u$, etc., and Q is the total truncation error, defined in* (1.16)–(1.20). *The constant ζ depends only on $\underline{v}$, $\inf_{[\underline{e},\bar{e}]} T'(e)$ (see* (1.3)–(1.4)), *and on an upper bound for λ. The constant C depends on ϵ, λ, $p\big|_R$, $T\big|_{[\underline{e},\bar{e}]}$, α, β, δ, and r, and on upper bounds for C_0 and the reciprocal of the difference between the two sides of* (1.26).

(1.27) shows in particular that

$$\sup_{0\leq t\leq \bar{t}} \left[\|\Delta v(\cdot,t)\| + |\Delta u(\cdot,t)|_{-\alpha} + |\Delta e(\cdot,t)|_{-\beta}\right] \leq C[E(0) + Q(0,\bar{t};U^h)], \tag{1.28}$$

where

$$E(0) = \|\Delta v_0\| + |\Delta u_0|_{-\alpha} + |\Delta e_0|_{-\beta}.$$

Bounds for Δu and Δe in L^2 can be recovered in either of two ways, both of which entail the introduction of an initial layer. First, we may simply appeal to the definition (1.14) of E to obtain

$$\sup_{0\leq t\leq \bar{t}} \left[\|\Delta v(\cdot,t)\| + t^{\delta/2}\|\Delta u(\cdot,t)\| + t^{\beta/2}\|\Delta e(\cdot,t)\|\right] \leq C[E(0) + Q(0,\bar{t};U^h)]. \tag{1.29}$$

Alternatively, we may apply the bounds (1.22) for u_x and e_x and interpolate as follows:

$$\begin{aligned}\|\Delta u(\cdot,t)\| = |\Delta u|_0 &\leq |\Delta u|_1^{\alpha/(1+\alpha)} |\Delta u|_{-\alpha}^{1/(1+\alpha)} \\ &\leq Ct^{-\alpha/4(1+\alpha)} |\Delta u|_{-\alpha}^{1/(1+\alpha)},\end{aligned}$$

so that

$$|\Delta u|_{-\alpha} \geq C^{-1} t^{\alpha/4} \|\Delta u\|^{1+\alpha}.$$

Treating Δe in a similar way, we then find from (1.28) that

$$\begin{aligned}\sup_{0\leq t\leq \bar{t}} &\left[\|\Delta v(\cdot,t)\| + t^{\alpha/4}\|\Delta u(\cdot,t)\|^{1+\alpha} + t^{\beta/4}\|\Delta e(\cdot,t)\|^{1+\beta}\right] \\ &\leq C[E(0) + Q(0,\bar{t};U^h)].\end{aligned} \tag{1.30}$$

(1.30) may be an improvement over (1.29), depending upon the context.

The bounds (1.29) and (1.30) degenerate when t is close to zero. This difficulty may be dealt with as follows. The result (1.20) in Theorem 1.1 of [1] shows that, for every $\theta > 0$, there is a constant $C = C(\theta)$ such that

$$\int_0^{\bar{t}} \int t^{(1+\theta)/2} u_t^2 \, dx dt \leq C. \tag{1.31}$$

The same bound holds for e_t, at least under the hypotheses of the present theorem, as discussed above. It then follows easily that

$$\|u(\cdot,t) - u_0\|, \; \|e(\cdot,t) - e_0\| \leq Ct^{(1-\theta)/4}.$$

If the bound (1.31) holds as well for u_t^h and e_t^h (as when U^h itself is an exact weak solution), we can then simply triangulate to obtain that, for any $\theta > 0$, there is a constant $C = C(\theta)$ such that

$$\|\Delta u(\cdot,t)\| + \|\Delta e(\cdot,t)\| \leq C\left[\|\Delta u_0\| + \|\Delta e_0\| + t^{(1-\theta)/4}\right]. \tag{1.32}$$

(1.32) may be an improvement over (1.29) or (1.30) when t is close to zero.

In the case that U^h is itself an exact weak solution, $Q(0,\bar{t};U^h) = 0$, and (1.28)–(1.30) imply the uniqueness of weak solutions of the initial-value problem for (1.1) as well as the continuous dependence of the solution on the initial condition. Alternatively, when U^h is an approximation to U, the results (1.28)–(1.30) allow for the derivation of bounds for the error $U^h - U$ from bounds for the weak truncation error Q, which is a measure of the *consistency* of the numerical procedure used to generate U^h, and which presumably can be estimated in terms of mesh parameters or ranks of projection operators.

In §2, we prove the above theorem be estimating the various terms appearing in the definition (1.14) of E. The key estimates are obtained by subtracting the weak equations $\mathcal{L}_j(\cdot\,; U) = 0$ from the corresponding definitions (1.10)–(1.12) of the functionals $\mathcal{L}_j(\cdot\,; U^h)$ and choosing the test functions to satisfy an appropriate adjoint system. These adjoint equations are solved backwards in time, with "initial" data given in the appropriate dual classes. We state a number of important properties of these adjoint solutions, which we then apply to prove the main estimate (1.27). Bounds for the adjoint solutions follow from standard energy methods and interpolation theorems and are derived in §3.

2. Proof of the theorem. In this section, we prove the theorem described in §1 by estimating the various terms appearing in the definition (1.14) of $E(t)$. These estimates will be obtained in a sequence of lemmas directly from the weak forms (1.10)–(1.12), in which the test functions φ, ψ, and χ will be chosen to satisfy appropriate adjoint equations. The hypotheses of the theorem will be in force throughout this section, and C will always denote a generic positive constant as described in the theorem.

We begin with a bound for the term $t^{\beta/2}\|\Delta e(\cdot,t)\|$.

LEMMA 2.1. *There is a constant C such that, for $0 \le t_1 \le t_2 \le \bar{t}$,*

$$\begin{aligned} t_2^{\beta/2}\|\Delta e(\cdot,t_2)\| \le C\Bigg[& t_1^{\beta/2}\|\Delta e(\cdot,t_1)\| + \sup_{t_1\le t\le t_2}\left(\|\Delta v(\cdot,t)\| + |\Delta e(\cdot,t)|_{-\beta}\right) \\ & + \left(\int_{t_1}^{t_2}\int t^{\delta}\Delta u_x^2 dxdt\right)^{1/2} + Q_3(t_1,t_2;U^h)\Bigg] \end{aligned} \tag{2.1}$$

Proof. Without loss of generality, $[t_1,t_2] \subseteq (0,\bar{t})$. We subtract the equation $\mathcal{L}_3(t_1,t_2,t^{\beta/2}\chi;U) = 0$ from the definition (1.12) of $\mathcal{L}_3(t_1,t_2,t^{\beta/2}\chi;U^h)$ and rearrange to obtain

$$\begin{aligned} & t^{\beta/2}\int \Delta e(x,t)\chi(x,t)\,dx\Big|_{t_1}^{t_2} \\ &= \int_{t_1}^{t_2}\int t^{\beta/2}\Bigg[\Delta e\chi_t - \lambda\frac{\Delta T_x}{v^h}\chi_x - \lambda\left(\frac{1}{v^h}-\frac{1}{v}\right)T_x\chi_x \\ &\quad - \left(\Delta u_x p^h + \Delta p u_x - \varepsilon\frac{(u_x^h)^2 - u_x^2}{v^h} - \varepsilon\left(\frac{1}{v^h}-\frac{1}{v}\right)u_x^2\right)\chi\Bigg]\,dxdt \\ &\quad + \frac{\beta}{2}\int_{t_1}^{t_2}\int t^{(\beta/2)-1}\Delta e\chi\,dxdt + \mathcal{L}_3(t_1,t_2,t^{\beta/2}\chi;U^h). \end{aligned} \tag{2.2}$$

We shall choose the function χ so that the first two terms on the right here cancel, modulo lower order terms. Specifically, we let j_η be the standard space–time mollifier and we define $e_\eta = j_n * e$, $e_\eta^h = j_\eta * e^h$, and $v_\eta^h = j_\eta * v^h$. We let $A(x,t) = T[e^h(x,t), e(x,t)]$ be the usual divided difference, and $A_\eta = T[e_\eta^h, e_\eta]$. We then take χ to be the solution of the following adjoint system, solved backwards in time, with data specified at $t = t_2$:

$$\begin{cases} \chi_t + \lambda A_\eta\left(\dfrac{\chi_x}{v_\eta^h}\right)_x = 0, & t \le t_2, \\ \chi(x,t_2) = H(x). \end{cases} \tag{2.3}$$

For the present, we take H in the Schwartz class $\mathcal{S}(\mathbb{R})$. The first two terms on the right side of (2.2) then become

$$-\lambda\int_{t_1}^{t_2}\int t^{\beta/2}\left[\left(\frac{1}{v^h}-\frac{1}{v_\eta^h}\right)\Delta T_x\chi_x + \Big((A-A_\eta)\Delta e\Big)_x\frac{\chi_x}{v_\eta^h}\right]dxdt.$$

We thus obtain from (2.2) that

$$\begin{aligned} t_2^{\beta/2}\left|\int \Delta e(x,t_2)H(x)\,dx\right| &\leq t_1^{\beta/2}\left|\int (\Delta e\chi)(x,t_1)\,dx\right| \\ &+C\int_{t_1}^{t_2}\int t^{\beta/2}\Big[\left|\chi_x\Delta T_x(v^h-v^h_\eta)\right|+\left|\Big((A-A_\eta)\Delta e\Big)_x\chi_x\right| \\ &\qquad +|\Delta vT_x\chi_x|+|\Delta u_x\chi|+|\Delta pu_x\chi| \\ &\qquad +|(u^h_x+u_x)\Delta u_x\chi|+|\Delta vu^2_x\chi|\Big]\,dxdt \\ &+C\left|\int_{t_1}^{t_2}\int t^{(\beta/2)-1}\Delta e\chi\,dxdt\right|+|\mathcal{L}_3(t_1,t_2,t^{\beta/2}\chi;U^h)|. \end{aligned} \tag{2.4}$$

We shall bound each of the terms on the right side of (2.4) in terms of $\|H\|$ and then let $\eta\to 0$. First, however, we need to obtain various estimates for the adjoint function χ in terms of its data H.

LEMMA 2.2. *There are positive constants ζ and C, as described in the statement of the theorem, and independent of η, such that, if (1.26) holds, then the solution χ of the system (2.3) satisfies*

$$\begin{aligned} &\sup_{0\leq t\leq t_2}\left[\|\chi(\cdot,t)\|+(t_2-t)^{1/2}\|\chi_x(\cdot,t)\|\right] \\ &\qquad +\left(\int_0^{t_2}\left[\|\chi_x(\cdot,t)\|^2+(t_2-t)\|\chi_t(\cdot,t)\|^2\right]dt\right)^{1/2} \\ &\leq C\|H\| \end{aligned} \tag{2.5}$$

and

$$\sup_{0\leq t\leq t_2}\|\chi_x(\cdot,t)\|+\left(\int_0^{t_2}\|\chi_t(\cdot,t)\|^2\,dt\right)^{1/2}\leq C|H|_1. \tag{2.6}$$

(2.5) and (2.6) follow from standard energy estimates; details are deferred to §3.

We now apply Lemma 2.2 to bound each of the terms on the right side of (2.4) in terms of $\|H\|$; we then let $\eta\to 0$ and take the sup over $H\in\mathcal{S}$. We shall present the details for five representative terms. Similar arguments apply to the others.

Applying Lemma 2.2, we may bound the first term in the double integral in (2.4) by

$$C\left[\int_{t_1}^{t_2}\int \Delta T_x^2(v^h-v^h_\eta)^2dxdt\right]^{1/2}.$$

The integrand here approaches zero a.e. as $\eta\to 0$ and is bounded pointwise by $C\Delta T_x^2\in L^1(\mathbb{R}\times[t_1,t_2])$. This term therefore approaches zero as $\eta\to 0$. Next, we split the fifth term in the double integral in (2.4) by writing $|\Delta p|\leq C(|\Delta v|+|\Delta e|)$. Applying (1.22) and Lemma 2.2, we may then bound the second of the terms that result by

$$\begin{aligned} \int_{t_1}^{t_2}\int t^{\beta/2}|\Delta eu_x\chi|dxdt &\leq C\sup_{t_1\leq t\leq t_2}(t^{\beta/2}\|\Delta e(\cdot,t)\|)\int_{t_1}^{t_2}\|\chi(\cdot,t)\|^{1/2}\|\chi_x(\cdot,t)\|^{1/2}\|u_x(\cdot,t)\|\,dt \\ &\leq C\|H\|\sup_t(t^{\beta/2}\|\Delta e\|)\int_{t_1}^{t_2}(t_2-t)^{-1/4}t^{-1/4}dt \\ &\leq C|t_2-t_1|^{1/2}\|H\|\sup_t(t^{\beta/2}\|\Delta e\|). \end{aligned}$$

Applying (1.22) and Lemma 2.2 again, we may bound the next term in (2.4) by

$$\begin{aligned}\int_{t_1}^{t_2}\int t^{\beta/2}|u_x\Delta u_x\chi|dxdt &\le C\left(\int\int t^{\delta}\Delta u_x^2\right)^{1/2}\left(\int_{t_1}^{t_2} t^{\beta-\delta}\|\chi\|\|\chi_x\|\|u_x\|^2dt\right)^{1/2}\\ &\le C\|H\|\left(\int_{t_1}^{t_2}\int t^{\delta}\Delta u_x^2\right)^{1/2}\left(\int_{t_1}^{t_2} t^{\beta-\delta-1/2}(t_2-t)^{-1/2}dt\right)^{1/2}\\ &\le C\|H\|\left(\int_{t_1}^{t_2}\int t^{\delta}\Delta u_x^2\right)^{1/2}\end{aligned}$$

since $\beta > \delta$ by (1.13). Next, we bound the second last term in (2.4) by

$$\begin{aligned}&C\int_{t_1}^{t_2} t^{(\beta/2)-1}|\Delta e(\cdot,t)|_{-\beta}|\chi(\cdot,t)|_{\beta}\,dt\\ &\qquad\le C\sup_{t_1\le t\le t_2}|\Delta e(\cdot,t)|_{-\beta}\int_{t_1}^{t_2} t^{(\beta/2)-1}\|\chi(\cdot,t)\|^{1-\beta}\|\chi_x(\cdot,t)\|^{\beta}\,dt\\ &\qquad\le C\|H\|\sup|\Delta e|_{-\beta}\int_{t_1}^{t_2} t^{(\beta/2)-1}(t_2-t)^{-\beta/2}dt\\ &\qquad\le C\|H\|\sup|\Delta e|_{-\beta}\end{aligned}$$

since $\beta > 0$. Finally, since $\mathcal{L}_3$ is linear in the test function, we can apply the definition (1.18) of Q_3 to bound the last term in (2.4) by $\|H\|Q_3(t_1,t_2;U^h)$. Assembling all these estimates, then letting $\eta \to 0$ and taking the sup over $H \in \mathcal{S}$, we obtain from (2.4) that

$$\begin{aligned}t_2^{\beta/2}\|\Delta e(\cdot,t_2)\| \le C\Bigg[&t_1^{\beta/2}\|\Delta e(\cdot,t_1)\| + |t_2-t_1|^{1/2}\sup_{t_1\le t\le t_2}\left(t^{\beta/2}\|\Delta e(\cdot,t)\|\right)\\ &+\sup_{t_1\le t\le t_2}\left(\|\Delta v(\cdot,t)\| + |\Delta e(\cdot,t)|_{-\beta}\right)\\ &+\left(\int_{t_1}^{t_2}\int t^{\delta}\Delta u_x^2\right)^{1/2} + Q_3(t_1,t_2;U^h)\Bigg].\end{aligned}$$

A simple Gronwall-type estimate then enables us to eliminate the second term on the right here. This proves the lemma. □

In the following lemma, we extend the result (2.1) to include a bound for the term $t^{\delta/2}\|\Delta u(\cdot,t)\|$ appearing in the definition (1.14) of E.

Lemma 2.3. *There is a positive constant C, as described in the statement of the theorem, such that, for $0 \le t_1 \le t_2 \le \bar{t}$,*

$$\begin{aligned}&\sup_{t_1\le t\le t_2}\left(t^{\delta/2}\|\Delta u(\cdot,t)\| + t^{\beta/2}\|\Delta e(\cdot,t)\|\right) + \left(\int_{t_1}^{t_2}\int t^{\delta}\Delta u_x^2dxdt\right)^{1/2}\\ &\qquad\le C\Bigg[t_1^{\delta/2}\|\Delta u(\cdot,t_1)\| + t_1^{\beta/2}\|\Delta e(\cdot,t_1)\|\\ &\qquad\quad+\sup_{t_1\le t\le t_2}\left(\|\Delta v(\cdot,t)\| + |\Delta u(\cdot,t)|_{-\alpha} + |\Delta e(\cdot,t)|_{-\beta}\right)\\ &\qquad\quad+Q_2(t_1,t_2;U^h) + Q_3(t_1,t_2;U^h)\Bigg].\end{aligned}\tag{2.7}$$

Proof. Without loss of generality, $[t_1,t_2]\subseteq(0,\bar{t})$. A simple approximation argument shows first that we may take $\psi = t^\delta\Delta u$ in the definition (1.11) of $\mathcal{L}_2(t_1,t_2,\psi;U^h)$ and also that $\mathcal{L}_2(t_1,t_2,t^\delta\Delta u;U)=0$. We subtract the latter equation from the definition of $\mathcal{L}_2(t_1,t_2,t^\delta\Delta u;U^h)$ and rearrange, applying the fact that

$$\int_{t_1}^{t_2}\int \Delta u\Delta u_t\,dxdt = \frac{1}{2}\int \Delta u(x,\cdot)^2dx\Big|_{t_1}^{t_2},$$

which holds because $\Delta u\in C([t_1,t_2];L^2(\mathbb{R}))$ and $\Delta u_t\in L^2(\mathbb{R}\times[t_1,t_2])$ by (1.6) and (1.24). The result is that

(2.8)

$$\begin{aligned}\frac{1}{2}t_2^\delta\|\Delta u(\cdot,t_2)\|^2 &+ \varepsilon\int_{t_1}^{t_2}\int\frac{t^\delta\Delta u_x^2}{v^h}\\ &= \frac{1}{2}t_1^\delta\|\Delta u(\cdot,t_1)\|^2 + \int_{t_1}^{t_2}\int\left[\frac{1}{2}\delta t^{\delta-1}\Delta u^2 + t^\delta\Big(\Delta p\Delta u_x + O(\Delta v)u_x\Delta u_x\Big)\right]\\ &\qquad + \mathcal{L}_2(t_1,t_2,t^\delta\Delta u;U^h).\end{aligned}$$

A simple interpolation allows us to bound the first term in the double integral on the right by

$$\begin{aligned}C&\int_{t_1}^{t_2} t^{\delta-1}|\Delta u(\cdot,t)|_{-\alpha}^{2/(1+\alpha)}|\Delta u(\cdot,t)|_1^{2\alpha/(1+\alpha)}dt\\ &\le C\sup_{t_1\le t\le t_2}|\Delta u(\cdot,t)|_{-\alpha}^{2/(1+\alpha)}\int_{t_1}^{t_2} t^{\delta-1-\alpha\delta/(1+\alpha)}\left(t^\delta\int\Delta u_x^2dx\right)^{\alpha/(1+\alpha)}dt\\ &\le C\sup|\Delta u|_{-\alpha}^{2/(1+\alpha)}\left(\int_{t_1}^{t_2}\int t^\delta\Delta u_x^2\right)^{\alpha/(1+\alpha)}\end{aligned}$$

since $\delta>\alpha$ by (1.13). The second term in the double integral in (2.8) is bounded by

$$\begin{aligned}C&\int_{t_1}^{t_2}\int t^\delta(|\Delta v|+|\Delta e|)|\Delta u_x|dxdt\\ &\le C\left(\int_{t_1}^{t_2}\int t^\delta\Delta u_x^2\right)^{1/2}\left[\sup_{t_1\le t\le t_2}\|\Delta v(\cdot,t)\| + \left(\int_{t_1}^{t_2}t^{\delta-\beta}\left(t^\beta\int\Delta e^2dx\right)dt\right)^{1/2}\right]\\ &\le C\left(\int_{t_1}^{t_2}\int t^\delta\Delta u_x^2\right)^{1/2}\left[\sup\|\Delta v\| + |t_2-t_1|^{(\delta-\beta+1)/2}\sup(t^{\beta/2}\|\Delta e\|)\right].\end{aligned}$$

Next, we can bound the third term in the double integral in (2.8) by

$$\begin{aligned}C&\Big(\int\int t^\delta\Delta u_x^2\Big)^{1/2}\sup\|\Delta v\|\int_{t_1}^{t_2}t^\delta\|\Delta u(\cdot,t)\|_\infty^2\,dt\\ &\le C\left(\int_{t_1}^{t_2}\int t^\delta\Delta u_x^2\right)^{1/2}\sup\|\Delta v\|\end{aligned}$$

by (1.23). Finally, we apply the definition (1.17) of Q_2 to bound the last term in (2.8) by $Q_2(t_1,t_2;U^h)^2$. Assembling these estimates and applying Young's inequality, we thus obtain from (2.8) that

(2.9)

$$t_2^{\delta/2}\|\Delta u(\cdot,t_2)\| + \left(\int_{t_1}^{t_2}\int t^\delta \Delta u_x^2\right)^{1/2} \leq C\Bigg[t_1^{\delta/2}\|\Delta u(\cdot,t_1)\| + \sup_{t_1\leq t\leq t_2}\left(\|\Delta v(\cdot,t)\| + |\Delta u(\cdot,t)|_{-\alpha}\right) + |t_2-t_1|^{(\delta-\beta+1)/2}\sup_{t_1\leq t\leq t_2}\left(t^{\beta/2}\|\Delta e(\cdot,t)\|\right) + Q_2(t_1,t_2;U^h)\Bigg].$$

We now add a small multiple of (2.1) to (2.9), apply a simple Gronwall argument, and then take appropriate sups to obtain the result. □

In the following lemma, we obtain bounds for the terms $\|\Delta v\|$, $|\Delta u|_{-\alpha}$, and $|\Delta e|_{-\beta}$ appearing in the definition of E.

LEMMA 2.4. *There is a positive constant C, as described in the statement of the theorem, and a positive exponent θ, depending only on α, β, δ, and r, such that, for $0\leq t_1\leq t_2\leq \bar{t}$,*

$$\sup_{t_1\leq t\leq t_2}\left(\|\Delta v(\cdot,t)\| + |\Delta u(\cdot,t)|_{-\alpha} + |\Delta e(\cdot,t)|_{-\beta}\right) \leq C\Bigg[E(t_1) + |t_2-t_1|^\theta\left(\sup_{t_1\leq t\leq t_2}E(t) + \left(\int_{t_1}^{t_2}\int t^\delta\Delta u_x^2\right)^{1/2}\right) + Q(t_1,t_2;U^h)\Bigg]. \tag{2.10}$$

Proof. Without loss of generality, $[t_1,t_2]\subseteq(0,\bar{t})$. We subtract equations (1.10)–(1.12) $\mathcal{L}_j(t_1,t_2,\cdot\,;U) = 0$ from the corresponding definitions of the functionals $\mathcal{L}_j(t_1,t_2,\cdot\,;U^h)$ for test functions $\varphi,\psi,\chi\in X([t_1,t_2])$. Writing $\Delta p = a\Delta v + b\Delta e$ and rearranging, we obtain

(2.11)

$$\begin{aligned}\mathcal{L}_1+\mathcal{L}_2+\mathcal{L}_3 &= \int(\Delta v\varphi + \Delta u\psi + \Delta e\chi)\,dx\Big|_{t_1}^{t_2}\\ &+\int_{t_1}^{t_2}\int\Bigg[-\Delta v(\varphi_t + a\psi_x) - \Delta u(\psi_t+\varphi_x) + \Delta u_x\frac{\varepsilon\psi_x}{v^h}\\ &\qquad - \Delta e\chi_t + \Delta T_x\frac{\lambda\chi_x}{v^h}\Bigg]\,dxdt\\ &+\int_{t_1}^{t_2}\int\Bigg[-b\Delta e\psi_x + \varepsilon\left(\frac{1}{v^h}-\frac{1}{v}\right)u_x\psi_x + \lambda\left(\frac{1}{v^h}-\frac{1}{v}\right)T_x\chi_x + \Delta u_x p^h\chi\\ &\qquad + \Delta p u_x\chi + \frac{\varepsilon(u_x^h+u_x)\Delta u_x}{v^h}\chi + \varepsilon\left(\frac{1}{v^h}-\frac{1}{v}\right)u_x^2\chi\Bigg]\,dxdt.\end{aligned}$$

We shall choose the test functions φ, ψ, and χ so as to effectively eliminate the first double integral on the right-hand side in (2.11). First, we again denote the standard space–time mollifier by j_n, and we write $v_\eta = j_n * v$, $v_\eta^h = j_n * v^h$, etc. Then, just as in the proof of Lemma 2.1, we take A and A_η to be the divided differences $A = T[e^h, e]$ and $A_\eta = T[e_\eta^h, e_\eta]$. Next, we define the section p^e by $p^e(v) = p(v,e)$, so that $\Delta p = p^e[v^h,v]\Delta v + p(v^h,e^h) - p(v^h,e)$. We may then take $a = p^e[v^h,v]$ and $a_\eta = p^{e_\eta}[v^h_\eta, v_\eta]$. Now, given $F, G, H \in \mathcal{S}(\mathbb{R})$, we solve the adjoint system

$$\tag{2.12} \begin{cases} \varphi_t + a_\eta \psi_x = 0, \\ \psi_t + \varphi_x + \left(\dfrac{\varepsilon \psi_x}{v^h_\eta}\right)_x = 0, \\ \chi_t + \lambda A_\eta \left(\dfrac{\chi_x}{v^h_\eta}\right)_x = 0 \end{cases}$$

backwards in time, with initial data given at $t = t_2$:

$$\tag{2.13} (\varphi, \psi, \chi)\Big|_{t=t_2} = (F, G, H).$$

Observe that the third equation in (2.12) is identical to (2.3) so that, just as in the proof of Lemma 2.1,

$$\tag{2.14} \begin{aligned} &\int_{t_1}^{t_2} \int \left(-\Delta e \chi_t + \Delta T_x \frac{\lambda \chi_x}{v^h}\right) dxdt \\ &\quad = \lambda \int_{t_1}^{t_2} \int \left[\left(\frac{1}{v^h} - \frac{1}{v^h_\eta}\right) \Delta T_x \chi_x + \Big((A - A_\eta)\Delta e\Big)_x \frac{\chi_x}{v^h_\eta}\right] dxdt. \end{aligned}$$

Substituting (2.12) and (2.14) into (2.11) and rearranging, we thus obtain that

$$\tag{2.15} \begin{aligned} &\left| \int (\Delta v \varphi + \Delta u \psi + \Delta e)(x, t_2)\, dx \right| \\ &\quad \le \left| \int (\Delta v \varphi + \Delta u \psi + \Delta e \chi)(x, t_1)\, dx \right| \\ &\qquad + \left| \int_{t_1}^{t_2} \int \left[(a_\eta - a)\Delta v \psi_x + \varepsilon \Delta u_x \psi_x \left(\frac{1}{v^h} - \frac{1}{v^h_\eta}\right) \right.\right. \\ &\qquad\qquad \left.\left. + \lambda \Delta T_x \chi_x \left(\frac{1}{v^h} - \frac{1}{v^h_\eta}\right) + \lambda \Big((A - A_\eta)\Delta e\Big)_x \frac{\chi_x}{v^h_\eta} \right] dxdt \right| \\ &\qquad + \left| \int_{t_1}^{t_2} \int \left[-b \Delta e \psi_x + \varepsilon \left(\frac{1}{v^h} - \frac{1}{v}\right) u_x \psi_x + \lambda \left(\frac{1}{v^h} - \frac{1}{v}\right) T_x \chi_x + \Delta u_x p^h \chi \right.\right. \\ &\qquad\qquad \left.\left. + \Delta p u_x \chi + \varepsilon \frac{u^h_x + u_x}{v^h} \Delta u_x \chi + \varepsilon \left(\frac{1}{v^h} - \frac{1}{v}\right) u_x^2 \chi \right] dxdt \right| \\ &\qquad + |\mathcal{L}_1| + |\mathcal{L}_2| + |\mathcal{L}_3|. \end{aligned}$$

Before proceeding further, we need to obtain various bounds for the adjoint functions φ, ψ, and χ in terms of their data F, G, and H.

LEMMA 2.5. *There are positive constants C and ζ, as described in the statement of the theorem, and independent of η, such that, if* (1.26) *holds, then the solution (φ, ψ, χ) of the adjoint system* (2.12)–(2.13) *satisfies*

$$\tag{2.16} \begin{aligned} &|||(\varphi, \psi, \chi)|||_{[t_1, t_2]} \equiv \sup_{t_1 \le t \le t_2} \Big[\|\varphi(\cdot,t)\| + |\psi(\cdot,t)|_\alpha + |\chi(\cdot,t)|_\beta \\ &\qquad + (t_2 - t)^{(1-\alpha)/2} \|\psi_x(\cdot,t)\| + (t_2 - t)^{(1-\beta)/2} \|\chi_x(\cdot,t)\| \Big] \\ &\qquad + \left(\int_{t_1}^{t_2} \int [\psi_x^2 + \chi_x^2 + (t_2 - t)^{1-\alpha} \psi_t^2 + (t_2 - t)^{1-\beta} \chi_t^2]\, dxdt \right)^{1/2} \\ &\qquad \le CK, \end{aligned}$$

where $K = \|F\| + |G|_\alpha + |H|_\beta$.

(2.16) follows from standard energy estimates and interpolation results; the proof is deferred to §3.

Applying Lemma 2.5, we can thus bound the first integral on the right-hand side of (2.15) by $CKE(t_1)$, and straightforward arguments similar to those given in the proof of Lemma 2.1 show that the second integral vanishes in the limit as $\eta \to 0$. We shall apply Lemma 2.5 to bound the two most difficult terms in the third integral on the right-hand side of (2.15); similar, but simpler, arguments apply to the other terms. The second term in the integral in question is bounded by

$$C\left[\int_{t_1}^{t_2}\int |\Delta v u_x|\left|\frac{\varepsilon\psi_x}{v_\eta^h}+\varphi\right| dxdt + \int_{t_1}^{t_2}\int |\Delta v u_x \varphi|\,dxdt\right]. \tag{2.17}$$

We apply (1.23) and (2.16) to bound the second term in (2.17) by

$$\begin{aligned}C\sup_{t_1\le t\le t_2}\|\Delta v(\cdot,t)\| &\int_{t_1}^{t_2}\|u_x\|_\infty\|\varphi\|dt\\ &\le CK|t_2-t_1|^\theta \sup_{t_1\le t\le t_2} E(t).\end{aligned}$$

To bound the first term in (2.17), we note that, from the second equation in (2.12),

$$\begin{aligned}\left\|\frac{\varepsilon\psi_x}{v_\eta^h}+\varphi\right\|_\infty^2 &\le 2\left\|\frac{\varepsilon\psi_x}{v_\eta^h}+\varphi\right\|\,\left\|\left(\frac{\varepsilon\psi_x}{v_\eta^h}+\varphi\right)_x\right\|\\ &\le C(\|\psi_x\|+\|\varphi\|)\|\psi_t\|\\ &\le C(t_2-t)^{(\alpha-1)/2}K\|\psi_t\|.\end{aligned}$$

Thus by (1.22),

(2.18)

$$\begin{aligned}\int_{t_1}^{t_2}\int |\Delta v u_x|\left|\frac{\varepsilon\psi_x}{v_x^h}+\varphi\right| dxdt &\le CK^{1/2}\sup_{t_1\le t\le t_2}\|\Delta v(\cdot,t)\|\int_{t_1}^{t_2}(t_2-t)^{(\alpha-1)/4}\|\psi_t\|^{1/2}\|u_x\|dt\\ &\le CK^{1/2}\sup\|\Delta v\|\int_{t_1}^{t_2}(t_2-t)^{(\alpha-1)/2}t^{-1/4}\left[(t_2-t)^{1-\alpha}\int\psi_t^2dx\right]^{1/4}dt\\ &\le CK^{1/2}\sup\|\Delta v\|\left[\int_{t_1}^{t_2}(t_2-t)^{2(\alpha-1)/3}t^{-1/3}dt\right]^{3/4}\left[\int_{t_1}^{t_2}\int(t_2-t)^{1-\alpha}\psi_t^2\,dxdt\right]^{1/4}\\ &\le CK|t_2-t|^\theta\sup_{t_1\le t\le t_2}E(t).\end{aligned}$$

This completes the argument for the term in (2.17). Next, in order to estimate the second-to-last term in the third integral on the right side of (2.15), we shall derive a bound for $\|\chi\|_\infty$ in terms of $|\chi|_\beta$ and $|\chi|_1$. Thus let $m(\xi) = 1 + |\xi|$ and choose

$$q \in [0,1] \cap \left[0, \frac{1}{2(1-\beta)}\right). \tag{2.19}$$

Then

$$\begin{aligned}\|\chi\|_\infty \le \|\hat\chi\|_{L^1} &= \int (m^\beta|\hat\chi|)^q\left(m|\hat\chi|\right)^{1-q}m^{q(1-\beta)-1}dx\\ &\le \left[\int(m^\beta|\hat\chi|)^2\right]^{q/2}\left[\int(m|\hat\chi|)^2\right]^{(1-q)/2}\left[\int m^{2q(1-\beta)-2}\right]^{1/2}\\ &\le C|\chi|_\beta^q|\chi|_1^{1-q}\\ &\le CK(t_2-t)^{(\beta-1)(1-q)/2}.\end{aligned} \tag{2.20}$$

Applying (1.22) and (2.20), we may then bound the second-to-last term in the third integral in (2.15) by

$$
\begin{aligned}
C\left(\int_{t_1}^{t_2}\int t^\delta \Delta u_x^2\right)^{1/2} &\left[\int_{t_1}^{t_2} t^{-\delta}(\|u_x\|^2+\|u_x^h\|^2)\|\chi\|_\infty^2 dt\right]^{1/2} \\
&\leq CK\left(\int_{t_1}^{t_2}\int t^\delta \Delta u_x^2\right)^{1/2}\left[\int_{t_1}^{t_2} t^{-\delta-1/2}(t_2-t)^{(\beta-1)(1-q)}dt\right]^{1/2}.
\end{aligned}
$$

Treating the cases $\beta > 1/2$ and $\beta \leq 1/2$ separately, we find that q can be chosen, consistently with (2.19), so that the integral in this expression is bounded by $|t_2-t_1|^\theta$, and the term in question is bounded by

$$
CK|t_2-t_1|^\theta\left(\int_{t_1}^{t_2}\int t^\delta \Delta u_x^2\right)^{1/2}.
$$

(It is here that we use the hypothesis (1.13) that $\delta < \min\{1/2, \beta\}$). Last, we apply the definition (1.20) of Q and the estimate (2.16) to obtain

$$
\begin{aligned}
|\mathcal{L}_1| + |\mathcal{L}_2| + |\mathcal{L}_3| &\leq \||(\varphi,\psi,\chi)\||_{[t_1,t_2]} Q(t_1,t_2;U^h) \\
&\leq CKQ(t_1,t_2;U^h).
\end{aligned}
$$

Estimating the other terms in (2.15) in a similar way, letting $\eta \to 0$, then taking the sup over (F,G,H), we finally obtain the result (2.10). □

Proof of the theorem. Adding a small multiple of (2.10) to (2.7), we obtain that

$$
\begin{aligned}
\sup_{t_1\leq t\leq t_2} E(t) &+ \left(\int_{t_1}^{t_2}\int t^\delta \Delta u_x^2\right)^{1/2} \\
&\leq C\left[E(t_1) + |t_2-t_1|^\theta \sup_{t_1\leq t\leq t_2} E(t) + Q(t_1,t_2;U^h)\right].
\end{aligned}
$$

A simple Gronwall-type argument then enables us to eliminate the middle term on the right. This proves the main estimate (1.27). □

3. Adjoint equation estimates.

In this section, we prove the estimates (2.5) and (2.16) in Lemmas 2.2 and 2.5 for solutions of the adjoint system (2.12)–(2.13). We begin with a statement of two interpolation results; these will be used to derive the required fractional-Sobolev norm estimates from L^2 and H^1 estimates.

LEMMA 3.1. (a) *Let $T : \mathcal{S}(\mathbb{R}) \to L^2(\mathbb{R})$ be linear ($\mathcal{S}$ is the Schwartz class), and suppose that there are constants C_j such that, for $g \in \mathcal{S}(\mathbb{R})$,*

$$
|Tg|_{b_j} \leq C_j |g|_{a_j}, \qquad j = 1, 2,
$$

where $a_j, b_j \in \mathbb{R}$. Let $s \in [0,1]$ and set

$$
\left\{
\begin{aligned}
a &= sa_1 + (1-s)a_2, \\
b &= sb_1 + (1-s)b_2, \\
C &= C_1^s C_2^{1-s}.
\end{aligned}
\right. \tag{3.1}
$$

Then

$$|Tg|_b \le C|g|_a$$

for all $g \in \mathcal{S}(\mathbb{R})$.

(b) *Let* $S : \mathcal{S}(\mathbb{R}) \to L^2(\mathbb{R}\times[t_1,t_2])$ *be linear and* $w \in C([t_1,t_2])$ *be strictly positive. Suppose that there are constants* C_j *such that, for* $g \in \mathcal{S}(\mathbb{R})$,

$$\left(\int_{t_1}^{t_2}\int w^{b_j}|Sg|^2\,dxdt\right)^{1/2} \le C_j|g|_{a_j}, \qquad j=1,2,$$

where $a_j, b_j \in \mathbb{R}$. *Let* a, b, *and* C *be as in* (3.1); *then*

$$\left(\int_{t_1}^{t_2}\int w^{b}|Sg|^2\,dxdt\right)^{1/2} \le C|g|_a$$

for all $g \in \mathcal{S}(\mathbb{R})$.

These results follow from straightforward variations of the proof of the Riesz–Thorin theorem; we therefore omit their proofs. (However, see [3, Lem. 2.2] for related interpolation results in which $\mathbb{R}$ is replaced by a finite interval.)

Proof of Lemma 2.2. Let χ be the solution of the adjoint system (2.3), and define

$$\mathcal{A}(t) = \int \chi(x,t)^2 dx + \lambda(t_2 - t)\int \chi_x(x,t)^2\,dx.$$

Choosing times $t \le t_1 \le t_2$, we then obtain from the partial differential equation in (2.3) that

$$\frac{1}{2}\int \chi^2 dx\Big|_t^{t_1} + \lambda\int_t^{t_1}\int \frac{A_\eta \chi_x^2}{v_\eta^h}\,dxds = \lambda\int_t^{t_1}\int \frac{\chi\chi_x(A_\eta)_x}{v_\eta^h}\,dxds.$$

The term on the right here is bounded by

$$\begin{aligned}
&C\lambda\|T''\|_\infty\int_t^{t_1}\left(\int\chi^2\right)^{1/4}\left(\int\chi_x^2\right)^{3/4}\left[\left(\int e_x^2\right)^{1/2}+\left(\int e_x^{h2}\right)^{1/2}\right]dt\\
&\quad\le C\lambda^{1/4}\|T''\|_\infty C_0^{1/2}\sup_{t\le s\le t_1}\mathcal{A}(s)\int_t^{t_1}(t_2-s)^{-3/4}s^{-1/4}ds\\
&\quad\le C\lambda^{1/4}\|T''\|_\infty C_0^{1/2}\sup_{t\le s\le t_1}\mathcal{A}(s)
\end{aligned}$$

by (1.22), where C now denotes a generic positive constant depending only on the inf of v in R and the inf of T in $[\underline{e},\overline{e}]$ (see (1.3) and (1.4)) and on an upper bound for λ. Thus

$$\begin{aligned}
&\int\chi(x,t)^2dx + \lambda\int_t^{t_1}\int\chi_x^2\,dxds\\
&\quad\le C\left[\int\chi(x,t_1)^2dx + \lambda^{1/4}\|T''\|_\infty C_0^{1/2}\sup_{t\le s\le t_1}\mathcal{A}(s)\right].
\end{aligned}\tag{3.2}$$

Next, we multiply the partial differential equation in (2.3) by $(t_2 - s)A_\eta^{-1}\chi_t$ and integrate to obtain that

$$\int_t^{t_1}\int \frac{(t_2-s)}{A_\eta}\chi_t^2\,dxds - \lambda(t_2-s)\int \frac{\chi_x^2}{2v_\eta^h}\,dx\Big|_t^{t_1}$$
$$\leq C\lambda\int_t^{t_1}\int\left[\chi_x^2 + (t_2-s)\chi_x^2|(v_\eta^h)_t|\right]dxds.$$

The second term on the right here is bounded by

$$C\lambda\sup_{t\leq s\leq t_1}\left[(t_2-s)\int\chi_x^2 dx\right]\int_t^{t_1}\|(v_\eta^h)_t(\cdot,s)\|_\infty\,ds$$
$$\leq CC_0^{1/2}\sup_{t\leq s\leq t_1}\mathcal{A}(s)|t_1-t|^r$$

by (1.23). Thus

$$\int_t^{t_1}\int(t_2-s)\chi_t^2\,dxds + \lambda(t_2-t)\int\chi_x(x,t)^2\,dx$$
$$\leq C\lambda\left[(t_2-t_1)\int\chi_x(x,t_1)^2\,dx + \int_t^{t_1}\int\chi_x^2\,dxds\right] \tag{3.3}$$
$$+CC_0^{1/2}\sup_{t\leq s\leq t_1}\mathcal{A}(s)|t_1-t|^r.$$

Adding a small multiple of (3.3) to (3.2), we obtain

$$\mathcal{A}(t) + \int_t^{t_1}\int\left[\lambda\chi_x^2 + (t_2-s)\chi_t^2\right]dxds$$
$$\leq C\left[\mathcal{A}(t_1) + \left(\lambda^{1/4}\|T''\|_\infty C_0^{1/2} + |t_1-t|^r\right)\sup_{t\leq s\leq t_1}\mathcal{A}(s)\right].$$

We now choose $t_0 \in [0, t_1]$, take the sup over $t \in [t_0, t_1]$, and apply the hypothesis (1.26) that $C\lambda^{1/4}\|T''\|_\infty C_0^{1/2} < 1/2$; the result is that

$$\sup_{t_0\leq t\leq t_1}\mathcal{A}(t) + \int_{t_0}^{t_1}\int\left[\lambda\chi_x^2 + (t_2-t)\chi_t^2\right]dxdt$$
$$\leq C\left[\mathcal{A}(t_1) + C_0^{1/2}|t_1-t_0|^r\sup_{t_0\leq t\leq t_1}\mathcal{A}(t)\right].$$

A simple Gronwall-type argument then completes the proof of (2.5). The proof of (2.6) is similar. □

Proof of Lemma 2.5. We first observe that the initial-value problem in (2.12) and (2.13) for χ is identical to (2.3), so that the estimates (2.5) and (2.6) hold. We shall apply Lemma 3.1 to obtain bounds for the various fractional norms of χ in (2.16). Thus fix a time $t < t_2$ and define a linear operator $T : \mathcal{S}(\mathbb{R}) \to L^2(\mathbb{R})$ by $TH = \chi(\cdot, t)$, where χ is the solution of the problem (2.3). Then by (2.5) and (2.6), $|TH|_0 \leq C|H|_0$ and $|TH|_1 \leq C|H|_1$. Lemma 3.1(a) therefore applies to show that $|TH|_\beta \leq C|H|_\beta$, that is, that $\|\chi(\cdot,t)\|_\beta \leq C|H|_\beta$. In addition, $|TH|_1 \leq C(t_2-t)^{-1/2}|H|_0$ and $|TH|_1 \leq$

$C|H|_1$, again by (2.5) and (2.6), so that, by Lemma 3.1(a), $\|\chi_x(\cdot,t)\| \le |TH|_1 \le C(t_2-t)^{(\beta-1)/2}|H|_\beta$.

To obtain a bound for the last term on the left side of (2.16), we fix a time $t' < t_2$ and define a linear map $S : \mathcal{S} \to L^2(\mathbb{R} \times [0,t'])$ by $SH = \chi_t$, where again χ is the solution of (2.3). The estimates (2.5) and (2.6) then show that

$$\int_0^{t'}\int (t_2-t)|SH|^2\,dxdt \le C|H|_0^2$$

and

$$\int_0^{t'}\int |SH|^2\,dxdt \le C|H|_1^2.$$

Lemma 3.1(b) therefore applies to show that

$$\int_0^{t'}\int (t_2-t)^{1-\beta}|SH|^2\,dxdt \le C|H|_\beta^2,$$

so that

$$\int_0^{t_2}\int (t_2-t)^{1-\beta}\chi_t^2\,dxdt \le C|H|_\beta^2.$$

Summarizing, we have thus shown that

$$\begin{aligned}\sup_{0\le t\le t_2}&\left(|\chi(\cdot,t)|_\beta + (t_2-t)^{(1-\beta)/2}\|\chi_x(\cdot,t)\|\right)\\ &+\left(\int_0^{t_2}\int\left[\chi_x^2+(t_2-t)^{1-\beta}\chi_t^2\right]dxdt\right)^{1/2}\\ \le{}& C|H|_\beta.\end{aligned}\tag{3.4}$$

This completes the proof of those bounds in (2.16) which involve χ.

Next, we derive bounds for φ and ψ in L^2. We multiply the first two equations in (2.12) by φ and ψ, respectively, and add and integrate. Dropping the modifiers on the coefficients a and v, we obtain that, for $t \le t_2$,

$$\begin{aligned}\frac{1}{2}\int(\varphi^2+\psi^2)(x,t)\,dx &+ \varepsilon\int_t^{t_2}\int\frac{\psi_x^2}{v}\,dxds\\ &= \frac{1}{2}\int(F^2+G^2)dx + \int_t^{t_2}\int(a-1)\varphi\psi_x\,dxds.\end{aligned}$$

It follows easily that

$$\begin{aligned}\sup_{0\le t\le t_2}\left(\|\varphi(\cdot,t)\|+\|\psi(\cdot,t)\|\right) &+ \left(\int_0^{t_2}\int\psi_x^2\,dxdt\right)^{1/2}\\ &\le C(\|F\|+\|G\|).\end{aligned}\tag{3.5}$$

To derive an H^1 bound for ψ, we multiply the second equation in (2.12) by $(t_2-t)\psi_t$ and integrate to obtain

$$\begin{aligned}\int_t^{t_2}\int(t_2-s)\psi_t^2\,dxds &+ \varepsilon(t_2-t)\int\left(\frac{\psi_x^2}{2v}\right)(x,t)dx\\ &= (t_2-t)\int(\psi_x\varphi)(x,t)dx + \int_t^{t_2}\int\frac{\varepsilon\psi_x^2}{2}\left[\frac{1}{v}+\frac{(t_2-s)v_t}{v^2}\right]dxds\\ &\quad+\int_t^{t_2}\int\psi_x[\varphi+(t_2-s)a\psi_x]\,dxds.\end{aligned}$$

Applying (3.5), we then get

$$\begin{aligned}\int_t^{t_2}\int(t_2-s)\psi_t^2+(t_2-t)\int\psi_x(x,t)^2\,dx\\ \leq C\left[\|F\|^2+\|G\|^2+\int_t^{t_2}\int(t_2-s)\psi_x^2|v_t|\right].\end{aligned}$$

The last term here is bounded by

$$C\sup_{t\leq s\leq t_2}\left[(t_2-s)\int\psi_x(x,s)^2\,dx\right]|t_2-t|^r$$

by (1.23). Substituting, we conclude that

$$\begin{aligned}\sup_{0\leq t\leq t_2}\left[(t_2-t)^{1/2}\|\psi_x(\cdot,t)\|\right]\cdot+\left[\int_0^{t_2}\int(t_2-t)\psi_t^2\,dxdt\right]^{1/2}\\ \leq C(\|F\|+\|G\|)\leq C(\|F\|+|G|_\alpha).\end{aligned}\tag{3.6}$$

A simple variation of the above argument would show that

$$\begin{aligned}\sup_{0\leq t\leq t_2}\|\psi_x(\cdot,t)\|+\left[\int_0^{t_2}\int\psi_t^2\,dxdt\right]^{1/2}\\ \leq C(\|F\|+|G|_1).\end{aligned}\tag{3.7}$$

To obtain the fractional-norm estimates in (2.16) for ψ, we fix a time $t<t_2$ and linear maps $Z:\mathcal{S}\times\mathcal{S}\to L^2$ and $T:\mathcal{S}\to L^2$ by $Z(F,G)=\psi(\cdot,t)$ and $TG=Z(0,G)$. (3.5) and (3.7) then show that $|TG|_0\leq C|G|_0$ and $|TG|_1\leq C|G|_1$, so that, by Lemma 3.1(a),

$$|TG|_\alpha\leq C|G|_\alpha.\tag{3.8}$$

In addition, we have from (3.7) that

$$|Z(F,0)|_\alpha\leq|Z(F,0)|_1\leq C\|F\|.\tag{3.9}$$

Combining (3.8) and (3.9), we thus obtain that

$$\begin{aligned}|\psi(\cdot,t)|_\alpha&=|Z(F,G)|_\alpha\\ &\leq|Z(F,0)|_\alpha+|TG|_\alpha\\ &\leq C(\|F\|+|G|_\alpha).\end{aligned}\tag{3.10}$$

The bounds

$$\begin{aligned}\sup_{0\leq t\leq t_2}t^{(1-\alpha)/2}\|\psi_x(\cdot,t)\|+]\left[\int_0^{t_2}\int(t_2-t)^{1-\alpha}\psi_t^2\,dxdt\right]\\ \leq C(\|F\|+|G|_\alpha)\end{aligned}\tag{3.11}$$

follow in a similar way, just as in the derivation of (3.4) above. Combining (3.4), (3.5), (3.10), and (3.11), we thus obtain (2.16). □

REFERENCES

[1] D. Hoff, *Discontinuous solutions of the Navier–Stokes equations for compressible flow*, Arch. Rational Mech. Anal., 114 (1991), pp. 15–46.

[2] ———, *Global well-posedness of the Cauchy problem for nonisentropic gas dynamics with discontinuous initial data*, J. Differential Equations, 95 (1992), pp. 33–73.

[3] D. Hoff and R. Zarnowski, *Continuous dependence in L^2 for discontinuous solutions of the viscous p-system*, Ann. Inst. H. Poincaré Anal. Non Linéare, 11 (1994), pp. 154–187.

[4] J. Zhao and D. Hoff, *A convergent finite difference scheme for the Navier–Stokes equations of one-dimensional, nonisentropic flow*, SIAM J. Numer. Anal., 31 (1994), pp. 1289–1311.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1212–1234, September 1996

A STEFAN PROBLEM FOR MULTIDIMENSIONAL REACTION-DIFFUSION SYSTEMS*

AVNER FRIEDMAN[†] AND BEI HU[‡]

Abstract. This paper deals with a Stefan problem for a system of three weakly coupled semilinear parabolic equations. This system describes the dissolution of a particle in a solution. The dissolved species A reacts chemically with species B already in the solution, thereby forming species C. Species C diffuses in the solution and some of it adsorbs to the particle's boundary and causes either (i) a decrease in the dissolution rate or (ii) an increase in the dissolution rate. It is proved that for the model in case (i) the solution is unstable in any small time interval, whereas for the model in case (ii), the problem has a unique solution in a small time interval.

Key words. Stefan problem, free boundary, nonstability

AMS subject classifications. 35B35, 35R35, 35R25, 35K57

1. Introduction.

Consider a solid particle composed of chemical A with uniform concentration A^*. The particle is in a solution. In the solution there is also another chemical B. As the particle dissolves, the A that enters the solution reacts with B to form C. Then species C diffuses in the solution and some of it reaches the solid particle and adsorbs to its surface. We shall denote the concentrations of species A, B, and C in the solution simply by A, B, and C, respectively.

We consider here the two-dimensional model in polar coordinates (r, θ). Assuming that the solid particle is enclosed by a surface $r = g(\theta, t)$, the reaction-diffusion equations are

$$\frac{\partial A}{\partial t} = D_A \Delta A - KAB, \tag{1}$$

$$\frac{\partial B}{\partial t} = D_B \Delta B - KAB, \tag{2}$$

$$\frac{\partial C}{\partial t} = D_C \Delta C + KAB \tag{3}$$

in $\{r > g(\theta, t)\}$, where K is the reaction rate and D_A, D_B, and D_C are the diffusion coefficients.

Let $\vec{n}$ denote the inward normal to the surface

$$\Gamma_t = \{r = g(\theta,\, t)\}$$

and V_n the velocity of Γ_t in the normal direction. Then

$$\vec{n} = (n_r,\, n_\theta) = \left(\frac{-g}{\sqrt{g^2 + g_\theta^2}},\, \frac{g_\theta}{g\sqrt{g^2 + g_\theta^2}} \right), \tag{4}$$

$$V_n = \frac{-g}{\sqrt{g^2 + g_\theta^2}}\, g_t. \tag{5}$$

* Received by the editors August 15, 1994; accepted for publication (in revised form) March 16, 1995.

[†] Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, MN 55455. The research of this author was partially supported by National Science Foundation grant DMS 94-01251.

[‡] Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556. The research of this author was partially supported by National Science Foundation grant DMS 92-24935.

By conservation of mass, the rate at which the particle's boundary moves is proportional to the flux of species A away from the particle, that is,

$$V_n = \alpha D_A \frac{\partial A}{\partial n} \quad \text{on } \Gamma_t \qquad (\alpha > 0). \tag{6}$$

Since there is no flux of B through or adsorption of B to the particle's surface,

$$\frac{\partial B}{\partial n} = 0 \quad \text{on } \Gamma_t. \tag{7}$$

The adsorption of C to the surface is proportional to the local saturation and is given by the empirical law $D_C \frac{\partial C}{\partial n} = -\gamma C^n$ for some positive constants γ and n. As in [2, Chap. 18], we take $n = 4$, i.e.,

$$D_C \frac{\partial C}{\partial n} = -\gamma C^4; \tag{8}$$

all the results of this paper, however, remain valid for general n.

As in [2, Chap. 18] and [4], we take

$$A(\infty, t) = 0, \qquad B(\infty, t) = B^*, \qquad C(\infty, t) = 0, \tag{9}$$

where "∞" means the limit as r goes to ∞, uniformly in θ, and B^* is a positive constant.

We next define the boundary condition for A. Denote by $\zeta(\theta, t)$ the concentration of C which covers A in a unit area (length) of the free boundary at the point $r = g(\theta, t)$; $\zeta(\theta, t) = 1$ if the point $r = g(\theta, t)$ is fully covered. We take the boundary condition for A on the free boundary, as in [2, Chap. 18], to be

$$D_A \zeta(\theta, t) \frac{\partial A}{\partial n} + (1 - \zeta(\theta, t))^+ (A - A^*) = 0 \quad \text{on } \Gamma_t. \tag{10}$$

Finally, we impose the following initial conditions:

$$g(\theta, 0) = g_0(\theta), \qquad \zeta(\theta, 0) = \zeta_0(\theta), \tag{11}$$

$$A(r, \theta, 0) = A_0(r, \theta), \quad B(r, \theta, 0) = B_0(r, \theta), \quad C(r, \theta, 0) = C_0(r, \theta). \tag{12}$$

For simplicity, we shall henceforth take $\alpha = 1$ and $\gamma = 1$.

Before we can analyze problems (1)–(12), we need to derive an equation for the evolution of ζ. There are two factors affecting the change of $\zeta(\theta, t)$: the flux $-D_C \frac{\partial C}{\partial n}$ and the change in the surface element of Γ_t along the normal.

Denote by $\frac{D}{Dt}$ the total derivative along the normal direction. Then

$$\begin{aligned} \frac{D\zeta(\theta, t)}{Dt} &= \lim_{\Delta t \to 0} \frac{\zeta(\theta + V_n n_\theta \Delta t, t + \Delta t) - \zeta(\theta, t)}{\Delta t} \\ &= \frac{\partial \zeta}{\partial t} + n_\theta V_n \frac{\partial \zeta}{\partial \theta}. \end{aligned} \tag{13}$$

If we introduce the surface element along Γ_t,

$$S(\theta, t) = \Psi(\theta, t) d\theta, \quad \text{where } \Psi(\theta, t) = \sqrt{g^2(\theta, t) + g_\theta^2(\theta, t)},$$

then the evolution of ζ can be described in the form

$$\begin{aligned}\zeta(\theta + V_n n_\theta \Delta t,\, t+\Delta t)S(\theta + V_n n_\theta \Delta t,\, t+\Delta t) - \zeta(\theta,\, t)S(\theta,\, t)\\ = -D_C\frac{\partial C}{\partial n}\cdot S(\theta,\, t)\Delta t + O\left((\Delta t)^2\right).\end{aligned}$$

From this equation, we easily derive

$$\frac{D\zeta}{Dt} + Q\zeta = -D_C\frac{\partial C}{\partial n}, \tag{14}$$

where

$$\begin{aligned}Q &= \lim_{\Delta t\to 0}\frac{S(\theta + V_n n_\theta \Delta t,\, t+\Delta t) - S(\theta,\, t)}{\Delta t\cdot S(\theta,\, t)}\\
&= \lim_{\Delta t\to 0}\frac{\Psi(\theta + V_n n_\theta \Delta t,\, t+\Delta t)d(\theta + V_n n_\theta\Delta t) - \Psi(\theta,\, t)d\theta}{\Delta t\cdot \Psi(\theta,\, t)d\theta}\\
&= \lim_{\Delta t\to 0}\frac{\Psi(\theta + V_n n_\theta \Delta t,\, t+\Delta t)\left[1 + \Delta t\cdot(\partial(V_n n_\theta)/\partial\theta)\right] - \Psi(\theta,\, t)}{\Delta t\cdot \Psi(\theta,\, t)}\\
&= \frac{\Psi_t + V_n n_\theta \Psi_\theta + \Psi(V_n n_\theta)_\theta}{\Psi}\\
&= \frac{1}{\Psi}\left\{\frac{gg_t + g_\theta g_{\theta t}}{\sqrt{g^2+g_\theta^2}} + (\Psi V_n n_\theta)_\theta\right\} \qquad \text{(by (5))}\\
&= \frac{1}{\Psi}\left\{-V_n - \frac{g_\theta}{\Psi}\left(\frac{\Psi V_n}{g}\right)_\theta + \left(\frac{g_\theta}{g}V_n\right)_\theta\right\} \qquad \text{(by (4) and (5))}\\
&= \frac{1}{\Psi}\left\{-V_n + \frac{\Psi V_n}{g}\left(\frac{g_\theta}{\Psi}\right)_\theta\right\} = -\frac{V_n}{\Psi} + \frac{V_n}{g}\left(\frac{g_\theta}{\Psi}\right)_\theta.\end{aligned}$$

Substituting this into (14) and using (13), we get

$$\frac{\partial\zeta}{\partial t} + n_\theta V_n\frac{\partial\zeta}{\partial\theta} + V_n\left\{-\frac{1}{\sqrt{g^2+g_\theta^2}} + \frac{1}{g}\left(\frac{g_\theta}{\sqrt{g^2+g_\theta^2}}\right)_\theta\right\}\zeta = -D_C\frac{\partial C}{\partial n}. \tag{15}$$

It will be convenient to work with the variable

$$\xi(\theta,\, t) = \frac{1}{\zeta(\theta,\, t)}. \tag{16}$$

Then equations (6), (10), and (15) become

$$V_n = \frac{-gg_t}{\sqrt{g^2+g_\theta^2}} = D_A\frac{\partial A}{\partial n} = (\xi - 1)^+(A^* - A) \quad \text{on } \Gamma_t, \tag{17}$$

$$\begin{aligned}\frac{\partial\xi}{\partial t} + \frac{V_n}{g}\frac{g_\theta}{\sqrt{g^2+g_\theta^2}}\frac{\partial\xi}{\partial\theta} - \frac{V_n}{g}\left(\frac{g_\theta}{\sqrt{g^2+g_\theta^2}}\right)_\theta\xi + \frac{V_n}{\sqrt{g^2+g_\theta^2}}\xi\\ = -\xi^2C^4 \quad \text{on } \Gamma_t.\end{aligned} \tag{18}$$

When the initial data are independent of θ, it is shown [4] that the system has a unique classical solution. However, the situation becomes more complicated when we allow the initial data to depend on θ. In fact, it will be shown that the problem is thus not well posed for classical solutions. In §2, we linearize the problem about a radial solution. We then show that the solution to the linearized problem is unique (§4) but may blow up in arbitrarily small time no matter how smooth the initial data are (§3). Furthermore, the full nonlinear problem is not stable near the radial solution (§5).

There are some similarities between our problem and the Stefan problem with supercooled water. The Mullin–Sekerka instabilities for the latter problem are not as bad as in our case; the linearized problem (for the supercooled water model) is well posed for all time provided the data are smooth, and the instability is only in the sense that the absolute value of the solution goes to ∞ as $t \to \infty$. To explain the origin of the instabilities for our problem, consider a nonradial particle as in Figure 1.1.

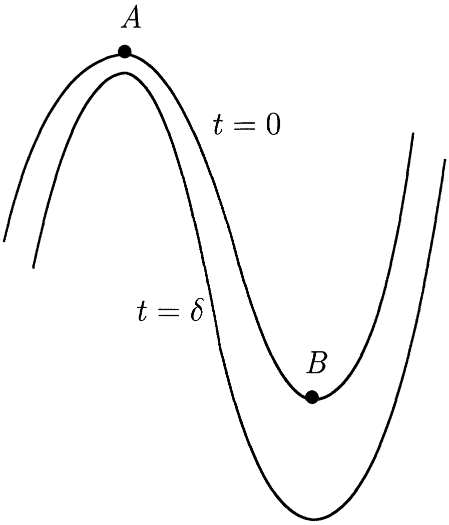


FIG. 1.1.

At a convex point A on the free boundary, as the particle dissolves, the local area shrinks. This increases the concentration ζ, which will then slow down the dissolution near A. The reverse situation occurs at a point B: the local area increases, ζ decreases, and the dissolution increases. This process tends to accentuate the ripples in the free boundary and will generally result in blow-up of the $C^{1+\beta}$ norm of Γ_t, at a very short time. This situation does not preclude the existence of a "weak solution," but the construction of an appropriate weak solution remains an open problem, even for much simpler Stefan problems such as in [3].

In the final section (§6), we shall consider the model where the adsorbed C increases the dissolution rather than inhibits it. (A situation like this arises, for instance, for an oil drop in water when soap is added to the water.) We shall establish in this case the existence and uniqueness of solutions for some small time.

Remark 1.1. It will be shown that $\zeta(r, \theta)$ satisfies a nonlinear second-order partial differential equation which is elliptic in case (i), where the adsorption of C slows the dissolution rate, and hyperbolic in case (ii), where the adsorption of C increases the dissolution rate; this equation is coupled, of course, to the equations for A, B, and C. In both cases, the data for ζ are the Cauchy data. Since the Cauchy problem is well posed for hyperbolic equations and ill posed for elliptic equations, this would explain mathematically why we get instability in case (i) and stability in case (ii).

Remark 1.2. The results of the paper extend to three-dimensional particles; the formulas are more complicated but the methods are the same.

2. The linearized problem. When the initial conditions in (11) and (12) are independent of θ and satisfy some regularity and compatibility assumptions, it is proved in [4] that the nonlinear system has a unique global radial solution (A^0, B^0, C^0, ζ^0, R^0) and that there is a finite shutdown time T^*, that is, $\zeta(T^*) = 1$ (and $\zeta(t) > 1$ if $t > T^*$). The radial solution satisfies

$$\left\{\begin{aligned} &R^0,\ \zeta^0 \in C^2[0, T^*), \\ &R^0(0) = R_0 > 0, \qquad \zeta^0(0) = \delta \in (0, 1), \\ &\frac{dR^0}{dt} < 0, \quad \frac{d\zeta^0}{dt} > 0 \quad \text{for } 0 \le t < T^* \end{aligned}\right. \tag{19}$$

and

$$\left\{\begin{aligned} &A^0, B^0, C^0 \in C^{2+\nu, 1+\frac{\nu}{2}}\{r \ge R^0(t), 0 \le t < T^*\}, \\ &\frac{\partial A^0}{\partial r} \le 0, \qquad \frac{\partial B^0}{\partial r} \ge 0, \\ &A^0 \ge 0, \quad B^0 \ge 0, \quad C^0 \ge 0, \end{aligned}\right. \tag{20}$$

where $0 < \nu < 1$; it is assumed that initially (20) holds, that

$$A^0(r, 0) = 0, \quad C^0(r, 0) = 0, \quad B^0(r, 0) = B^* \quad \text{if } r \ge R_0 + \delta_0 \tag{21}$$

for some $\delta_0 > 0$, and that the compatibility conditions

$$\frac{\partial^2 A^0(R_0, 0)}{\partial r^2} = \frac{\partial^2 B^0(R_0, 0)}{\partial r^2} = \frac{\partial^2 C^0(R_0, 0)}{\partial r^2} = 0, \quad A^0(R_0, 0) < A^*, \tag{22}$$

hold.

We linearize the system (1)–(3), (17), (18) with boundary conditions (7) and (8) about a radially symmetric solution by setting

$$\begin{gathered} A = A^0 + \varepsilon A^1, \qquad B = B^0 + \varepsilon B^1, \qquad C = C^0 + \varepsilon C^1, \\ g = R^0 + \varepsilon h, \qquad \xi = \xi^0 + \varepsilon \eta, \end{gathered} \tag{23}$$

where $\xi^0 = 1/\zeta^0$. Substituting (23) into (1)–(3) and dropping $O(\varepsilon^2)$ terms, we obtain

$$\begin{aligned} \left(\frac{\partial}{\partial t} - D_A \Delta\right) A^1 &= -K(A^1 B^0 + A^0 B^1), \qquad && r > R^0(t), \\ \left(\frac{\partial}{\partial t} - D_B \Delta\right) B^1 &= -K(A^1 B^0 + A^0 B^1), \qquad && r > R^0(t), \\ \left(\frac{\partial}{\partial t} - D_C \Delta\right) C^1 &= K(A^1 B^0 + A^0 B^1), \qquad && r > R^0(t). \end{aligned} \tag{24}$$

Clearly,

$$g_\theta^2 = \varepsilon^2 h_\theta^2.$$

Therefore, by (17),

$$D_A \frac{\partial A}{\partial n} = -g_t + O(\varepsilon^2).$$

Using (23), we get

$$D_A \left\{ \frac{\partial A^0}{\partial n}\bigg|_{r=g} - \frac{\partial A^0}{\partial n}\bigg|_{r=R^0} \right\} + \varepsilon D_A \frac{\partial A^1}{\partial n}\bigg|_{r=g} + D_A \frac{\partial A^0}{\partial n}\bigg|_{r=R^0}$$
$$= -\frac{dR^0}{dt} - \varepsilon \frac{\partial h}{\partial t} + O(\varepsilon^2)$$

so that, after dropping $O(\varepsilon^2)$ terms,

$$D_A \frac{\partial A^1}{\partial n} - D_A \frac{\partial^2 A^0}{\partial r^2} h = -\frac{\partial h}{\partial t} \quad \text{on} \quad r = R^0(t). \tag{25}$$

Similarly,

$$\frac{\partial B^1}{\partial n} - \frac{\partial^2 B^0}{\partial r^2} h = 0 \quad \text{on} \quad r = R^0(t), \tag{26}$$

$$\frac{\partial C^1}{\partial n} - \frac{\partial^2 C^0}{\partial r^2} h = 4(C^0)^3 C^1 + 4(C^0)^3 \frac{\partial C^0}{\partial r} h \quad \text{on} \quad r = R^0(t). \tag{27}$$

We next substitute g and ξ from (23) into (17) and (18) to obtain linearized equations for h and η. From (17),

$$V_n = -g_t + O(\varepsilon^2), \tag{28}$$

$$-g_t = (\xi - 1)(A^* - A) + O(\varepsilon^2). \tag{29}$$

Substituting g and ξ from (23) into (29), we get

$$\begin{aligned} -\frac{dR^0}{dt} - \varepsilon \frac{\partial h}{\partial t} &= \varepsilon \eta (A^* - A^0) - \varepsilon A^1 (\xi^0 - 1) \\ &\quad + (\xi^0 - 1) \left\{ (A^* - A^0)\Big|_{r=g} - (A^* - A^0)\Big|_{r=R^0} \right\} \\ &\quad + (\xi^0 - 1)(A^* - A^0)\Big|_{r=R^0} + O(\varepsilon^2). \end{aligned}$$

Dropping $O(\varepsilon^2)$ terms, we obtain

$$-\frac{\partial h}{\partial t} = \eta (A^* - A^0) - (\xi^0 - 1) A^1 + \frac{1}{D_A} (\xi^0 - 1)^2 (A^* - A^0) h. \tag{30}$$

Next, substituting (28) into (18) and recalling that $g_\theta = O(\varepsilon)$, we get

$$\xi_t - \frac{g_t g_\theta}{g^2} \xi_\theta + \frac{g_t}{g} \left(\frac{g_\theta}{g} \right)_\theta \xi - \frac{g_t}{g} \xi = -\xi^2 C^4 + O(\varepsilon^2).$$

Substituting g, ξ, and C from (23) and dropping $O(\varepsilon^2)$, we find that

$$\begin{aligned}&\eta_t + \frac{1}{(R^0)^2}\frac{dR^0}{dt}\xi^0 h_{\theta\theta} - \frac{1}{R^0}\frac{dR^0}{dt}\eta - \frac{1}{R^0}\xi^0 h_t + \frac{1}{(R^0)^2}\frac{dR^0}{dt}\xi^0 h\\ &= -2\xi^0(C^0)^4\eta - 4(C^0)^3(\xi^0)^2C^1 - 4(C^0)^3(\xi^0)^2\frac{\partial C^0}{\partial r}h \quad \text{on } r = R^0(t).\end{aligned}\tag{31}$$

We next express η in terms of h and h_t from (30) and substitute this into (31). This results in the elliptic equation

$$\begin{aligned}\mathcal{L}h \equiv\ &\frac{\partial}{\partial t}\left(a_1(t)\frac{\partial h}{\partial t}\right) + a_2(t)\frac{\partial^2 h}{\partial\theta^2} + b(t)\frac{\partial h}{\partial t} + c(t)h\\ &= f_1(t)C^1(R^0(t), \theta, t) + f_2(t)A^1(R^0(t), \theta, t)\\ &\quad + \frac{\partial}{\partial t}\left[f_3(t)A^1(R^0(t), \theta, t)\right],\end{aligned}\tag{32}$$

where

$$a_1(t) = -\frac{1}{A^* - A^0(R^0(t), t)}, \qquad (a_1(t) < 0),\tag{33}$$

$$a_2(t) = \frac{1}{(R^0(t))^2}\frac{dR^0(t)}{dt}\xi^0(t), \qquad (a_2(t) < 0),\tag{34}$$

$$\begin{aligned}b(t) = &-\frac{(\xi^0(t)-1)^2}{D_A} - \frac{\xi^0(t)}{R^0(t)}\\ &-\left\{2\xi^0(t)\Big(C^0\big(R^0(t), t\big)\Big)^4 - \frac{1}{R^0(t)}\frac{dR^0(t)}{dt}\right\}\frac{1}{A^* - A^0(R^0(t), t)},\end{aligned}\tag{35}$$

$$\begin{aligned}c(t) = &-\frac{2}{D_A}(\xi^0(t)-1)\frac{d\xi^0(t)}{dt} + \frac{1}{(R^0(t))^2}\frac{dR^0(t)}{dt}\xi^0(t)\\ &+4\Big(C^0\big(R^0(t), t\big)\Big)^3\big(\xi^0(t)\big)^2\frac{\partial C^0}{\partial r}\big(R^0(t), t\big)\\ &-\left\{2\xi^0(t)\Big(C^0\big(R^0(t), t\big)\Big)^4 - \frac{1}{R^0(t)}\frac{dR^0(t)}{dt}\right\}\frac{(\xi^0(t)-1)^2}{D_A},\end{aligned}\tag{36}$$

$$f_1(t) = -4\Big(C^0\big(R^0(t), t\big)\Big)^3\big(\xi^0(t)\big)^2,\tag{37}$$

$$f_2(t) = \left\{2\xi^0(t)\Big(C^0\big(R^0(t), t\big)\Big)^4 - \frac{1}{R^0(t)}\frac{dR^0(t)}{dt}\right\}\frac{\xi^0(t)-1}{A^* - A^0(R^0(t), t)},\tag{38}$$

$$f_3(t) = -\frac{\xi^0(t)-1}{A^* - A^0(R^0(t), t)}.\tag{39}$$

Notice that

$$\frac{\partial C^0}{\partial r} = \frac{1}{D_C}(C^0)^4 \quad \text{at } (R^0(t), t).$$

Therefore, all the coefficients a_1, a_2, b, c, f_1, f_2, and f_3 are Lipschitz continuous under the assumptions (20) and (21).

3. The linearized problem is unstable. In this section, we prove that the linearized problem is unstable. More specifically, we show that for any given small $T > 0$ and $\beta > 0$ and for any large positive integer m, there exist initial data for h and η whose first m derivatives are bounded by 1 such that a smooth solution exists for $0 < t < T$ but the $C^{1,\beta}$ norm of h becomes infinite at $t = T$.

THEOREM 3.1. *Consider the linearized problem* (24)–(27), (30), *and* (31) *under the assumptions* (19)–(22). *For any small* $T > 0$, $0 < \beta < \frac{1}{2}$, *and positive integer* m, *there exists a* C^m *solution* (A^1, B^1, C^1, h, η) *for* $0 < t < T$ *such that*

$$\left|D^j_\theta \eta(\theta, 0)\right| \leq 1, \quad \left|D^j_\theta h(\theta, 0)\right| \leq 1 \quad \textit{for} \quad 0 \leq j \leq m, \tag{40}$$

$$A^1(r, \theta, 0) = B^1(r, \theta, 0) = C^1(r, \theta, 0) \equiv 0 \quad \textit{for} \quad r \geq R^0(0), \tag{41}$$

but

$$\|h(\cdot, T - 0)\|_{C^{1+\beta}([0,2\pi])} = +\infty. \tag{42}$$

Proof. Let

$$\begin{aligned} G_T &= \{0 \leq \theta \leq 2\pi, \ 0 \leq t \leq T\}, \\ X &= \{h(\theta, t); \ h \ \text{and} \ h_t \ \text{belong to} \ C(G_T) \\ &\qquad \text{and are } 2\pi\text{-periodic in } \theta\}. \end{aligned}$$

Introduce the norm

$$\|h\|_X = \|h\|_{C(G_T)} + \|h_t\|_{C(G_T)}.$$

For each $h \in X$, we solve (24) with the boundary conditions (25)–(27), zero initial conditions, and zero boundary conditions at $r = \infty$. Clearly,

$$\begin{aligned} \|A^1\|_{L^\infty} &\leq C\|h\|_X + CT\|B^1\|_{L^\infty}, \\ \|B^1\|_{L^\infty} &\leq C\|h\|_X + CT\|A^1\|_{L^\infty}, \\ \|C^1\|_{L^\infty} &\leq C\|h\|_X + CT\left(\|A^1\|_{L^\infty} + \|B^1\|_{L^\infty}\right), \end{aligned}$$

so that

$$\|A^1\|_{L^\infty} + \|B^1\|_{L^\infty} + \|C^1\|_{L^\infty} \leq C\|h\|_X \tag{43}$$

if T is small enough.

Next, we use (43) and Hölder estimates for parabolic equations [7] to obtain

$$\begin{aligned} \|A^1\|_{C^{\alpha,\alpha/2}(\Omega_T)} + \|B^1\|_{C^{\alpha,\alpha/2}(\Omega_T)} + \|C^1\|_{C^{\alpha,\alpha/2}(\Omega_T)} \\ \leq C\|h\|_X \end{aligned} \tag{44}$$

for some $0 < \alpha < 1$, where

$$\Omega_T = \left\{(x_1, x_2, t); \ \sqrt{x_1^2 + x_2^2} \geq R^0(t), 0 \leq t \leq T\right\}.$$

The constant C is independent of T.

In what follows, we assume without loss of generality that $\frac{1}{2}\beta < \alpha < \beta$.

Since A^1, B^1, and C^1 have zero initial values, (44) implies that

$$\|A^1\|_{L^\infty} + \|B^1\|_{L^\infty} + \|C^1\|_{L^\infty} \le CT^{\alpha/2}\|h\|_X. \tag{45}$$

We now pick up any periodic function $k(\theta)$ such that

$$\|k\|_{C^{1+\beta}} = \infty, \qquad \|k\|_{C^{1+\beta/2}} < \infty \tag{46}$$

and denote by $\bar{h}$ the solution to the elliptic problem

$$\mathcal{L}\bar{h} = f_1(t)C^1(R^0(t),\, \theta,\, t) + f_2(t)A^1(R^0(t),\, \theta,\, t) + \frac{\partial}{\partial t}\left[f_3(t)A^1(R^0(t),\, \theta,\, t)\right], \tag{47}$$

$$\bar{h} \text{ is } 2\pi\text{-periodic in } \theta, \tag{48}$$

$$\bar{h}_t(\theta,\, 0) = 0, \qquad \bar{h}(\theta,\, T) = k(\theta). \tag{49}$$

We define the map W by

$$(Wh)(\theta,\, t) = \bar{h}(\theta,\, t).$$

We shall show that W has a fixed point, and this will give us the solution asserted in Theorem 3.1 (with η defined by (30)).

We first need to derive some estimates on h and its derivatives which will depend on the small parameter T in an appropriate way. It is convenient to scale variables by introducing

$$s = \frac{t}{T}, \qquad \varphi = \frac{\theta}{T}, \qquad \tilde{h}(s,\, \varphi) = \bar{h}(\theta,\, t).$$

Setting

$$\tilde{L}\tilde{h} \equiv \frac{\partial}{\partial s}\left(a_1(Ts)\frac{\partial \tilde{h}}{\partial s}\right) + a_2(Ts)\frac{\partial^2 \tilde{h}}{\partial \varphi^2} + Tb(Ts)\frac{\partial \tilde{h}}{\partial s} + T^2 c(Ts)\tilde{h},$$

we have

$$\begin{aligned} \tilde{L}\tilde{h} &= T^2 f_1(Ts)C^1\left(R^0(Ts),\, T\varphi,\, Ts\right) + T^2 f_2(Ts)A^1\left(R^0(Ts),\, T\varphi,\, Ts\right) \\ &\quad + T\frac{\partial}{\partial s}\left[f_3(Ts)A^1(R^0(Ts),\, T\varphi,\, Ts)\right] \end{aligned} \tag{50}$$

and

$$\begin{gathered} \left.\frac{\partial \tilde{h}}{\partial s}\right|_{s=0} = 0, \qquad \left.\tilde{h}\right|_{s=1} = k(T\varphi), \\ \tilde{h}(\varphi,\, s) \text{ is } \frac{2\pi}{T}\text{-periodic in } \varphi. \end{gathered} \tag{51}$$

To analyze the solution $\tilde{h}$, we introduce the auxiliary problem

$$
\begin{gathered}
a_1(Ts)\frac{\partial^2 v}{\partial s^2} + a_2(Ts)\frac{\partial^2 v}{\partial \varphi^2} = Tf_3(Ts)A^1(R^0(Ts), T\varphi, Ts), \\
v\Big|_{s=0} = 0, \qquad \frac{\partial v}{\partial s}\Big|_{s=1} = 0, \\
v \text{ is } \frac{2\pi}{T}\text{-periodic in } \varphi.
\end{gathered} \tag{52}
$$

Then

$$|||v|||_{L^\infty(\tilde{G})} \le CT|||A^1|||_{L^\infty},$$

where

$$\tilde{G} = \left[0, \frac{2\pi}{T}\right] \times [0, 1]$$

and $|||\cdot|||$ denotes norms in the variable (s, φ).

By Schauder's interior-boundary estimates [5],

$$
\begin{aligned}
|||v|||_{C^{2+\alpha/2}(\tilde{G})} &\le C\left(T|||A^1|||_{C^{\alpha/2}} + |||v|||_{L^\infty(\tilde{G})}\right) \\
&\le CT|||A^1|||_{C^{\alpha/2}},
\end{aligned} \tag{53}
$$

where $C^{\alpha/2}$ denotes Hölder space with exponent $\frac{\alpha}{2}$ in both space and time variables. Differentiating (52) in s and subtracting from equation (50), we find that

$$
\begin{aligned}
\tilde{L}_0(\tilde{h} - v_s) \;=\; & T^2 f_1(Ts)C^1\left(R^0(Ts), T\varphi, Ts\right) \\
& + T^2 f_2(Ts)A^1\left(R^0(Ts), T\varphi, Ts\right) \\
& - Tb(Ts)\frac{\partial^2 v}{\partial s^2} + T\frac{\partial a_2(Ts)}{\partial t}\frac{\partial^2 v}{\partial \varphi^2} \\
& - T^2 c(Ts)(\tilde{h} - v_s) - T^2 c(Ts)v_s, \\
\frac{\partial}{\partial s}(\tilde{h} - v_s)\Big|_{s=0} = 0, & \qquad (\tilde{h} - v_s)\Big|_{s=1} = k(Ts),
\end{aligned} \tag{54}
$$

where

$$\tilde{L}_0 = \tilde{L} - T^2 c(Ts).$$

By the maximum principle,

$$|||\tilde{h} - v_s|||_{L^\infty(\tilde{G})} \le C\Big(|||k|||_{L^\infty} + |||\text{the right-hand side of (54)}|||_{L^\infty}\Big).$$

Therefore, if we apply elliptic $C^{1+\alpha}$ estimates [9] to (54), we get

$$|||\tilde{h} - v_s|||_{C^{1+\alpha/2}(\tilde{G})} \le C\Big(|||k|||_{C^{1+\alpha/2}} + T^2|||A^1|||_{C^{\alpha/2}} + |||C^1|||_{C^{\alpha/2}} + T^2|||\tilde{h} - v_s|||_{L^\infty}\Big).$$

Recalling (53), we conclude that

$$|||\tilde{h}|||_{C^{1+\alpha/2}(\tilde{G})} \le C\left[|||k|||_{C^{1+\alpha/2}} + T\left(|||A^1|||_{C^{\alpha/2}} + |||C^1|||_{C^{\alpha/2}}\right)\right]$$

provided T is small enough. The above estimate written in terms of the variable $(t,\,\theta)$ reads

$$\begin{aligned}
&\|\bar h\|_{L^\infty} + T\Big(\|\bar h_t\|_{L^\infty} + \|\bar h_\theta\|_{L^\infty}\Big) + T^{1+\alpha/2}\Big([\bar h]_{C^{\alpha/2}} + [\bar h_\theta]_{C^{\alpha/2}}\Big)\\
&\qquad \le C\Big(\|k\|_{L^\infty} + T\|k_\theta\|_{L^\infty} + T^{1+\alpha/2}[k_\theta]_{C^{\alpha/2}}\Big)\\
&\qquad\quad + CT\Big(\|A^1\|_{L^\infty} + \|C^1\|_{L^\infty}\Big) + CT^{1+\alpha/2}\Big([A^1]_{C^{\alpha/2}} + [C^1]_{C^{\alpha/2}}\Big).
\end{aligned}\tag{55}$$

Using (44) and (45), we conclude that

$$\begin{aligned}
&\|\bar h\|_{L^\infty} + T\Big(\|\bar h_t\|_{L^\infty} + \|\bar h_\theta\|_{L^\infty}\Big) + T^{1+\alpha/2}\Big([\bar h_t]_{C^{\alpha/2}} + [\bar h_\theta]_{C^{\alpha/2}}\Big)\\
&\qquad \le C^*\Big(\|k\|_{L^\infty} + T\|k_\theta\|_{L^\infty} + T^{1+\alpha/2}[k_\theta]_{C^{\alpha/2}}\Big)\\
&\qquad\quad + CT^{1+\alpha/2}\Big(\|h\|_{L^\infty} + \|h_t\|_{L^\infty}\Big).
\end{aligned}\tag{56}$$

This estimate shows that if

$$X_0 = \Big\{h \in X;\ \|h\|_{L^\infty} + T\|h_t\|_{L^\infty} \le C^*\|k\|_{C^{1+\alpha/2}} + 1\Big\},$$

then W maps X_0 into itself provided T is small enough. It is also clear that WX_0 is a precompact subset of X_0 and that W is continuous. The Schauder fixed point theorem can then be applied to conclude that W has a fixed point h, and this gives a solution to the linearized problem with the same h.

Next, we show that h and η are C^m smooth.

The estimate (56) implies that

$$\|h\|_{C^{1+\alpha/2}(G_T)} \le C. \tag{57}$$

Recall by (49) that $h_t|_{t=0} = 0$. Since we have assumed (22), the compatibility condition (for parabolic equations) is satisfied for A^1, B^1, and C^1. We can therefore apply to equations (24) with boundary conditions (25)–(27), $C^{1+\alpha}$ parabolic estimates [8, Thm. 1.2], and Schauder estimates [7] to get

$$\begin{aligned}
&\|A^1\|_{C^{1+\alpha/2,\,1/2+\alpha/4}(\Omega_T)} \le C,\\
&\|B^1\|_{C^{2+\alpha/2,\,1+\alpha/4}(\Omega_T)} + \|C^1\|_{C^{2+\alpha/2,\,1+\alpha/4}(\Omega_T)} \le C.
\end{aligned}\tag{58}$$

We now differential (32) in θ and obtain the same equation for h_θ but with A^1 and C^1 replaced by A^1_θ and C^1_θ. If we use interior-boundary Schauder estimates (away from $t = T$), the procedure that led to (55)–(57) gives

$$\|h_\theta\|_{C^{1+\alpha/4}(G_{T-\varepsilon})} \le C_\varepsilon \tag{59}$$

for any $\varepsilon > 0$. By iterating this process m times, we arrive at the estimate

$$\left\|\frac{\partial^j h}{\partial\theta^j}\right\|_{C^{1+2^{-j-1}\alpha}(G_{T-\varepsilon j})} \le C_{\varepsilon,j} \quad (0 \le j \le m).$$

Since η is given by (31), we get similar estimates for η. Noting that ε is arbitrary, it follows that the solution has m derivatives (which are actually continuous) in θ.

Next, using (58) and interior-boundary Schauder estimates for h (away from $t = T$), we obtain

$$\|h\|_{C^{1+(1/2+\alpha/4)}(G_{T-\varepsilon})} \leq C. \tag{60}$$

The estimates (59) and (60) imply that

$$\|h_t\|_{C^{1+\alpha/2,1/2+\alpha/4}(G_{T-\varepsilon})} \leq C,$$

where $C^{1+\alpha/2,1/2+\alpha/4}$ means $C^{1+\alpha/2}$ in θ and $C^{1/2+\alpha/4}$ in t. Thus, by parabolic Schauder estimates,

$$\|A^1\|_{C^{2+\alpha/2,\,1+\alpha/4}(\Omega_{T-\varepsilon})} \leq C.$$

Notice that the coefficients in (33)–(39) are C^∞ for $t > 0$ (see [4]). Therefore, we can iterate the above process to conclude that h is C^m in (θ, t) away from $t = 0$ and $t = T$.

Finally, by multiplying the solution by a small enough constant δ, we find that δh and $\delta\eta$ have all their first m derivatives (in θ) bounded by 1 at $t = 0$ and that the solution of the linearized problem has (continuous) m derivatives for all $0 < t < T$ since (by (49)) $\delta h(\theta,\, T-0) = \delta k(\theta)$ and $\|\delta k\|_{C^{1+\beta}} = \infty$; this completes the proof of the theorem. □

4. Uniqueness for the linearized problem.

THEOREM 4.1. *Suppose that* $(A^1,\, B^1,\, C^1,\, h,\, \eta)$ *is a solution of the linearized problem for* $0 \leq t < T_0$ *such that*

$$A^1 = B^1 = C^1 = 0 \quad \text{at} \quad t = 0, \tag{61}$$

$$h = \eta = 0 \quad \text{at} \quad t = 0, \tag{62}$$

and

$$\begin{gathered} A^1,\, B^1, \text{ and } C^1 \quad \text{belong to} \quad C^{1+\alpha,(1+\alpha)/2}(\Omega_{T_0}), \\ h \in C^{1+\alpha,1+\alpha}(G_{T_0}), \qquad \eta \in C^{\alpha,\alpha}(G_{T_0}). \end{gathered} \tag{63}$$

Then

$$A^1 \equiv B^1 \equiv C^1 \equiv 0 \quad \text{in} \quad \Omega_{T_0}$$

and

$$h \equiv \eta \equiv 0 \quad \text{in} \quad G_{T_0}.$$

Proof. From (61), (62), and (30), it follows that

$$h_t(\theta,\, 0) \equiv 0. \tag{64}$$

Let $\varphi(\theta)$ be any 2π-periodic function such that

$$\varphi''(\theta) = -k^2\varphi(\theta),$$

where k is a nonnegative integer. Let

$$\tilde{A}(r, t) = \int_0^{2\pi} \varphi(\theta) A^1(r, \theta, t) d\theta$$

and define $\tilde{B}(r, t)$, $\tilde{C}(r, t)$, and $\tilde{h}(t)$ in a similar way. Then $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{h}$ satisfy the same boundary conditions (25)–(27) as A, B, C, and h. Multiplying the equations for A^1, B^1, and C^1 by $\varphi(\theta)$ and integrating with respect to θ, we find that

$$\begin{aligned} \tilde{A}_t &= D_A \left(\tilde{A}_{rr} + \frac{1}{r} \tilde{A}_r - \frac{k^2}{r^2} \tilde{A} \right) - K(\tilde{A}B^0 + A^0\tilde{B}), \\ \tilde{B}_t &= D_B \left(\tilde{B}_{rr} + \frac{1}{r} \tilde{B}_r - \frac{k^2}{r^2} \tilde{B} \right) - K(\tilde{A}B^0 + A^0\tilde{B}), \\ \tilde{C}_t &= D_C \left(\tilde{C}_{rr} + \frac{1}{r} \tilde{C}_r - \frac{k^2}{r^2} \tilde{C} \right) + K(\tilde{A}B^0 + A^0\tilde{B}) \end{aligned}$$

if $r > R^0(t)$. The same argument as in (43)–(45) shows that

(65) $$\|\tilde{A}\|_{L^\infty} + \|\tilde{B}\|_{L^\infty} + \|\tilde{C}\|_{L^\infty} \le CT^{\alpha/2} \left(\|\tilde{h}\|_{L^\infty} + \|\tilde{h}_t\|_{L^\infty} \right), \quad L^\infty = L^\infty(\Omega_T).$$

Next, multiplying (32) by $\varphi(\theta)$ and integrating with respect to θ, we obtain the differential equation

$$\begin{aligned} &\frac{d}{dt} \left(a_1 \frac{d\tilde{h}}{dt} \right) - k^2 a_2 \tilde{h} + b\tilde{h}_t + c\tilde{h} \\ &\quad = f_2(t)\tilde{C}\left(R^0(t), t\right) + f_2(t)\tilde{A}\left(R^0(t), t\right) + \frac{\partial}{\partial t}\left(f_3(t)\tilde{A}\left(R^0(t), t\right) \right). \end{aligned}$$

Since

$$\tilde{h} = \tilde{h}_t = 0 \quad \text{and} \quad \tilde{A} = 0 \quad \text{at} \quad t = 0,$$

we get by integration that

(66) $$\|\tilde{h}_t\|_{L^\infty} \le C\left(\|\tilde{A}\|_{L^\infty} + \|\tilde{C}\|_{L^\infty} \right) + CT\left((k^2+1)\|\tilde{h}\|_{L^\infty} + \|\tilde{h}_t\|_{L^\infty} \right).$$

Clearly,

(67) $$\|\tilde{h}\|_{L^\infty} \le T\|\tilde{h}_t\|_{L^\infty}.$$

Substituting (67) and (65) into (66) we find that if T is small enough (depending on k), then

$$\tilde{h} \equiv 0.$$

It then also follows that $\tilde{A} \equiv \tilde{B} \equiv \tilde{C} \equiv 0$ and $\eta \equiv 0$ if $t < T$. We can continue the process step by step to deduce that $\tilde{h} \equiv 0$ if $0 \le t < T_0$ (T_0 is independent of k).

Taking in particular $\varphi(\theta) = \sin kx$ or $\cos kx$ for any integers k, we see that the Fourier series of the continuously differentiable function $\theta \to h(\theta, t)$ vanishes identically. Therefore, $h \equiv 0$ and then also $A \equiv B \equiv C \equiv 0$ and $\eta \equiv 0$. □

5. The nonlinear problem is unstable. Theorem 3.1 can be extended to the full nonlinear problem. To be more precise, we say that a radial solution (A^0, B^0, C^0, ξ^0, R^0) satisfying (19)–(22) is *locally Lipschitz stable* if there exist $T > 0$ and a positive integer $m \geq 3$ such that for any solution (A, B, C, ξ, g) of the nonlinear problem, if

$$\begin{gathered}\|(A - A^0)_{t=0}\|_{C^m} + \|(B - B^0)_{t=0}\|_{C^m} + \|(C - C^0)_{t=0}\|_{C^m} \leq \varepsilon, \\ \|(\xi - \xi^0)_{t=0}\|_{C^m} + \|(g - R^0)_{t=0}\|_{C^m} \leq \varepsilon \quad \forall\varepsilon > 0\end{gathered} \tag{68}$$

and if the compatibility conditions hold (at $r = g(\theta,\, 0)$), then

$$\|A - A^0\|_{C^{2,1}} + \|B - B^0\|_{C^{2,1}} + \|C - C^0\|_{C^{2,1}} \leq C^* \varepsilon, \tag{69}$$

$$\|\xi - \xi^0\|_{C^{1,1}} \leq C^* \varepsilon, \qquad \|g - R^0\|_{C^{2,1}} \leq C^* \varepsilon \tag{70}$$

for some constant C^* independent of ε.

In (69), A^0, B^0, and C^0 have been extended as $C^{2+\nu,\, 1+\nu/2}$ functions in the domain

$$\Omega_{T,0} = \left\{\sqrt{x_1^2 + x_2^2} \geq R_0(t) - \delta_0, \quad 0 \leq t \leq T\right\}$$

for some $\delta_0 > 0$, and $C^{2,1}$ means $C^{2,1}(\Omega_T)$, where

$$\Omega_T = \{r \geq g(\theta,\, t), \quad 0 \leq t \leq T\}.$$

For T small enough, Ω_T is contained in $\Omega_{T,0}$. In (70), $C^{j,1}$ means $C^{j,1}(G_T)$.

THEOREM 5.1. *Each radial solution is not locally Lipschitz stable.*

Proof. Suppose the assertion is not true for a particular radial solution (A^0, B^0, C^0, ξ^0, R^0), that is, this solution is Lipschitz stable up to some fixed time $T > 0$.

Take any solution of the nonlinear problem with

$$\begin{gathered}A = A^0 + \varepsilon A^1_\varepsilon, \qquad B = B^0 + \varepsilon B^1_\varepsilon, \qquad C = C^0 + \varepsilon C^1_\varepsilon, \\ g = R^0 + \varepsilon h_\varepsilon, \qquad \xi = \xi^0 + \varepsilon \eta_\varepsilon,\end{gathered} \tag{71}$$

where

$$\begin{gathered}\|A^1_\varepsilon|_{t=0}\|_{C^m} + \|B^1_\varepsilon|_{t=0}\|_{C^m} + \|C^1|_{t=0}\|_{C^m} \leq 1, \\ \|h_\varepsilon|_{t=0}\|_{C^m} + \|\eta_\varepsilon|_{t=0}\|_{C^m} \leq 1\,.\end{gathered} \tag{72}$$

The Lipschitz stability implies that

$$\|A^1_\varepsilon\|_{C^{2,1}(\Omega_T)} + \|B^1_\varepsilon\|_{C^{2,1}(\Omega_T)} + \|C^1_\varepsilon\|_{C^{2,1}(\Omega_T)} \leq C, \tag{73}$$

$$\|\eta_\varepsilon\|_{C^{1,1}(G_T)} \leq C, \tag{74}$$

$$\|h_\varepsilon\|_{C^{2,1}(G_T)} \leq C, \tag{75}$$

where C is a constant independent of ε.

Differentiating the equation

$$g_t = -\frac{1}{g}\sqrt{g^2 + g_\theta^2}\,(\xi - 1)(A^* - A) \tag{76}$$

in θ and applying the estimates (73)–(75), we derive the bound

$$\|(h_\varepsilon)_{t\theta}\|_{L^\infty} \leq C. \tag{77}$$

Next, we differentiate (76) in t and use (73)–(75) and (77). We obtain

$$\|(h_\varepsilon)_{tt}\|_{L^\infty} \le C. \tag{78}$$

The estimates (75), (77), and (78) combined mean that

$$\|h_\varepsilon\|_{C^{2,2}(G_T)} \le C. \tag{79}$$

We now take a sequence of initial values as above with $\varepsilon \to 0$. For a subsequence $\varepsilon = \varepsilon_j \to 0$,

$$A^1_\varepsilon \to A^1, \quad B^1_\varepsilon \to B^1, \quad C^1_\varepsilon \to C^1 \quad \text{in } C^{1+\alpha,\,(1+\alpha)/2},$$
$$\eta_\varepsilon \to \eta \quad \text{in } C^{\alpha,\,\alpha},$$
$$h_\varepsilon \to h \quad \text{in } C^{1+\alpha,\,1+\alpha}$$

for any $0 < \alpha < 1$, and

$$\|\eta_t\|_{L^\infty} + \|\eta_\theta\|_{L^\infty} + \|h_{\theta\theta}\|_{L^\infty} + \|h_{\theta t}\|_{L^\infty} + \|h_{tt}\|_{L^\infty} < \infty. \tag{80}$$

We can then proceed as in §2 (but rigorously!) to prove that $(A^1_\varepsilon,\, B^1_\varepsilon,\, C^1_\varepsilon,\, \eta_\varepsilon,\, h_\varepsilon)$ satisfies the linearized equations and boundary conditions (on $r = R^0 + \varepsilon h_\varepsilon$) with an error term of order $O(\varepsilon)$. Letting $\varepsilon = \varepsilon_j \to 0$, we deduce that $(A^1\, B^1,\, C^1,\, \eta,\, h)$ is a solution to the linearized problem. Choose the initial values for the linearized problem to be as in Theorem 3.1. By uniqueness (Theorem 4.1), $(A^1,\, B^1,\, C^1,\, \eta,\, h)$ must coincide with the linearized solution established in Theorem 3.1 and, consequently,

$$\|h(\cdot,\, T)\|_{C^{1+\beta}} = \infty,$$

which is a contradiction to (80). □

6. A model with accelerated dissolution. In this section, we consider the case where the adsorbed C increases the dissolution of the grain:

$$D_A \frac{\partial A}{\partial n} + P(\zeta)(A - A^*) = 0 \quad \text{on } \Gamma_t, \tag{81}$$

where $P(s)$ is a smooth function and

$$P(s) > 0, \quad P'(s) > 0 \quad \text{for } s \ge 0. \tag{82}$$

Setting

$$k = g_\theta, \qquad F = A^* - A\,(g(\theta,\, t),\, \theta,\, t) \tag{83}$$

we then have, by (6),

$$V_n = P(\zeta)F$$

and, by (13)–(15),

$$\zeta_t = -\frac{k}{g\sqrt{g^2+k^2}} P(\zeta)F\zeta_\theta - \frac{P(\zeta)\zeta F}{g}\left(\frac{k}{\sqrt{g^2+k^2}}\right)_\theta + \frac{P(\zeta)\zeta F}{\sqrt{g^2+k^2}} + C^4$$

or

$$
\begin{aligned}
\zeta_t &= -\frac{k}{g\sqrt{g^2+k^2}}\,P(\zeta)F\zeta_\theta - \frac{P(\zeta)\zeta F g}{(g^2+k^2)^{3/2}}\,k_\theta \\
&\quad + P(\zeta)\zeta F\left[\frac{k^2}{(g^2+k^2)^{3/2}} + \frac{1}{\sqrt{g^2+k^2}}\right] + C^4.
\end{aligned}
\tag{84}
$$

Differentiating equation (5), written in the form

$$
g_t = -\frac{\sqrt{g^2+k^2}}{g}\,P(\zeta)F,
\tag{85}
$$

in θ, we obtain

$$
\begin{aligned}
k_t &= -\frac{k}{g\sqrt{g^2+k^2}}\,P(\zeta)Fk_\theta - \frac{\sqrt{g^2+k^2}}{g}\,P'(\zeta)F\zeta_\theta \\
&\quad - \frac{\sqrt{g^2+k^2}}{g}\,P(\zeta)F_\theta + P(\zeta)F\,\frac{k^3}{g^2\sqrt{g^2+k^2}}.
\end{aligned}
\tag{86}
$$

The system (84)–(86) is a nonlinear hyperbolic system. To transform it into a diagonalized form, introduce

$$
\begin{aligned}
\ell(\theta, t) &= \int_0^{k(\theta,t)} \frac{g(\theta,t)}{g^2(\theta,t)+s^2}\,ds + \int_{c_0}^{\zeta(\theta,t)} \left(\frac{P'(s)}{sP(s)}\right)^{1/2} ds, \\
m(\theta, t) &= \int_0^{k(\theta,t)} \frac{g(\theta,\, t)}{g^2(\theta,\, t)+s^2}\,ds - \int_{c_0}^{\zeta(\theta,t)} \left(\frac{P'(s)}{sP(s)}\right)^{1/2} ds, \qquad c_0 > 0.
\end{aligned}
\tag{87}
$$

Then

$$
\ell_t = \frac{g}{g^2+k^2}\,k_t + \left(\frac{P'(\zeta)}{\zeta P(\zeta)}\right)^{1/2} \zeta_t + g_t(\theta,\, t)\int_0^{k(\theta,t)} \frac{s^2 - g^2(\theta,\, t)}{(g^2(\theta,\, t)+s^2)^2}\,ds,
$$

$$
\ell_\theta = \frac{g}{g^2+k^2}\,k_\theta + \left(\frac{P'(\zeta)}{\zeta P(\zeta)}\right)^{1/2} \zeta_\theta + k(\theta,\, t)\int_0^{k(\theta,\, t)} \frac{s^2 - g^2(\theta,\, t)}{(g^2(\theta,\, t)+s^2)^2}\,ds,
$$

and similar formulas hold for m_t and m_θ.

A direct computation shows that

$$
\begin{aligned}
\ell_t &= -\left(\frac{k}{g\sqrt{g^2+k^2}}\,P(\zeta)F + \sqrt{\frac{P'(\zeta)\zeta P(\zeta)}{g^2+k^2}}\,F\right)\ell_\theta \\
&\quad + \left(\frac{k}{g\sqrt{g^2+k^2}}\,P(\zeta)F + \sqrt{\frac{P'(\zeta)\zeta P(\zeta)}{g^2+k^2}}\,F\right) k\int_0^k \frac{s^2-g^2}{(s^2+g^2)^2}\,ds \\
&\quad + g_t\int_0^k \frac{s^2-g^2}{(s^2+g^2)^2}\,ds - \frac{P(\zeta)}{\sqrt{g^2+k^2}}\,F_\theta + \frac{P(\zeta)Fk^3}{g(g^2+k^2)^{3/2}}
\end{aligned}
\tag{88}
$$

$$+ \sqrt{P'(\zeta)\zeta P(\zeta)}\, F \left[\frac{k^2}{(g^2+k^2)^{3/2}} + \frac{1}{\sqrt{g^2+k^2}} \right] + C^4 \sqrt{\frac{P'(\zeta)}{\zeta P(\zeta)}},$$

$$\begin{aligned} (89) \qquad m_t = &- \left(\frac{k}{g\sqrt{g^2+k^2}} P(\zeta) F - \sqrt{\frac{P'(\zeta)\zeta P(\zeta)}{g^2+k^2}}\, F \right) m_\theta \\ &+ \left(\frac{k}{g\sqrt{g^2+k^2}} P(\zeta) F - \sqrt{\frac{P'(\zeta)\zeta P(\zeta)}{g^2+k^2}}\, F \right) k \int_0^k \frac{s^2-g^2}{(s^2+g^2)^2}\, ds \\ &+ g_t \int_0^k \frac{s^2-g^2}{(s^2+g^2)^2}\, ds - \frac{P(\zeta)}{\sqrt{g^2+k^2}} F_\theta + \frac{P(\zeta) F k^3}{g(g^2+k^2)^{3/2}} \\ &- \sqrt{P'(\zeta)\zeta P(\zeta)}\, F \left[\frac{k^2}{(g^2+k^2)^{3/2}} + \frac{1}{\sqrt{g^2+k^2}} \right] - C^4 \sqrt{\frac{P'(\zeta)}{\zeta P(\zeta)}}. \end{aligned}$$

Substituting (85) into (88) and (89), we get

$$\ell_t + a\ell_\theta = \varphi(g,\, h,\, \zeta,\, F,\, F_\theta,\, E), \tag{90}$$

$$m_t + b m_\theta = \psi(g,\, h,\, \zeta,\, F,\, F_\theta,\, E), \tag{91}$$

where

$$E = C^4\,(g(\theta,\, t),\, \theta,\, t)\,,$$

φ and ψ are smooth functions as long as $g \ge c_0 > 0$, $\zeta \ge c_0$, and

$$a = \frac{k}{g\sqrt{g^2+k^2}} P(\zeta) F + \left(\frac{P'(\zeta)\zeta P(\zeta)}{g^2+k^2} \right)^{1/2} F, \tag{92}$$

$$b = \frac{k}{g\sqrt{g^2+k^2}} P(\zeta) F - \left(\frac{P'(\zeta)\zeta P(\zeta)}{g^2+k^2} \right)^{1/2} F. \tag{93}$$

We impose initial conditions

$$g|_{t=0} = g_0(\theta), \quad \zeta|_{t=0} = \zeta_0(\theta), \tag{94}$$

where $g_0(\theta)$ and $\zeta_0(\theta)$ are periodic functions such that

$$\begin{aligned} &g_0(\theta) \ge 2c_0, \quad \zeta_0(\theta) \ge 2c_0 \qquad (c_0 > 0), \\ &\|g_0\|_{C^{2+\alpha}} \le M, \quad \|\zeta_0\|_{C^{1+\alpha}} \le M, \end{aligned} \tag{95}$$

where $M < \infty$.

The system (85), (90), (91) is hyperbolic with given initial data

$$g = g_0(\theta), \qquad \ell|_{t=0} = \ell_0(\theta), \qquad m|_{t=0} = m_0(\theta) \tag{96}$$

and

$$g_0(\theta) \ge 2c_0 > 0,$$

$$\begin{aligned} &\ell_0(\theta) - m_0(\theta) \geq 2 \int_{c_0}^{2c_0} \left(\frac{P'(s)}{sP(s)} \right)^{1/2} ds \equiv 4c_1 > 0, \\ &\ell_0(\theta) + m_0(\theta) \leq \pi - 2c_2 \quad \text{for some } \ c_2 > 0, \\ &\|g_0\|_{C^{2+\alpha}} \leq M, \quad \|\ell_0\|_{C^{1+\alpha}} \leq C^*(M, c_0), \\ &\|m_0\|_{C^{1+\alpha}} \leq C^*(M, c_0). \end{aligned} \tag{97}$$

In what follows, we denote by

$$\| \quad \|_{C_\theta^{m+\alpha}(G_T)}$$

the $C^{m+\alpha}$ Hölder norm in θ, taken uniformly in t as (θ, t) varies in G_T, and by

$$\| \quad \|_{C_t^\alpha(G_T)}$$

the C^α norm in t, uniformly in θ as (θ, t) varies in G_T. Finally, by

$$\|u\|_{C^{k+\alpha,\, k+\alpha}(G_T)}$$

we denote the sum of the C^α Hölder norms (in $(\theta, t) \in G_T$) of all the derivatives $\partial_\theta^i \partial_t^j u$ with $0 \leq i + j \leq k$.

LEMMA 6.1. *Given F and E continuous in G_T and 2π-periodic in θ such that*

$$\|F\|_{C_\theta^{2+\alpha}(G_T)} \leq M_1, \quad \|E\|_{C_\theta^{1+\alpha}(G_T)} \leq M_1, \tag{98}$$

there exists a unique solution (ζ, g, k) of the system (84)–(86) *with initial conditions* (94) *and with $g_\theta = k$ for $0 \leq t \leq T$ provided T is small enough, and*

$$\|\zeta\|_{C_\theta^{1+\alpha}(G_T)} + \|g\|_{C_\theta^{2+\alpha}(G_T)} \leq \tilde{C}(M, c_0), \tag{99}$$

$$\|\zeta\|_{C^{1+\alpha,\, 1+\alpha}(G_T)} + \|g\|_{C^{2+\alpha,\, 2+\alpha}(G_T)} \leq C(M, M_1, c_0), \tag{100}$$

$$g(\theta, t) \geq c_0, \quad \zeta(\theta, t) \geq c_0 \quad in \ \Omega_T; \tag{101}$$

here T depends on M, M_1, and c_0.

LEMMA 6.2. *If $(\hat{F}, \hat{E})$ is another pair satisfying* (98), *then the corresponding solution $\hat{\zeta}$, $\hat{g}$ satisfies*

$$\begin{aligned} &\|\zeta - \hat{\zeta}\|_{C_\theta^\alpha(G_t)} + \|g - \hat{g}\|_{C_\theta^{1+\alpha}(G_t)} \\ &\qquad \leq C(M, M_1, c_0)t \left[\|F - \hat{F}\|_{C_\theta^{1+\alpha}(G_t)} + \|E - \hat{E}\|_{C_\theta^\alpha(G_t)} \right] \end{aligned} \tag{102}$$

for $0 \leq t \leq T$.

Proof of Lemma 6.1. From (87), we see that

$$k = g \tan\left(\frac{\ell + m}{2} \right) \quad \text{if } \ |\ell + m| < \pi \tag{103}$$

and

$$\int_{c_0}^{\zeta} \left(\frac{P'(s)}{sP(s)} \right)^{1/2} ds = \frac{\ell - m}{2} \quad \text{if } \ \ell - m > 0. \tag{104}$$

Since we are interested only in solutions for small time, without loss of generality, we may take $P(s) = s$ for s large. Then (104) uniquely defines ζ as a function of $(\ell - m)/2$,

$$\zeta = Y\left(\frac{\ell - m}{2}\right), \tag{105}$$

and $Y(s)$ is smooth for $s \geq c_1$, c_1 as in (97).

Substituting (103), (105) into the right-hand sides of (90)–(93) and (85), we find that (85) and (90), (91) form a quasilinear hyperbolic system for (g, ℓ, m). By standard results [1], [6], this system with the initial conditions (96) has a unique solution such that

$$\|g\|_{C^{1+\alpha,\,1+\alpha}(G_T)} + \|\ell\|_{C^{1+\alpha,\,1+\alpha}(G_T)} + \|m\|_{C^{1+\alpha,\,1+\alpha}(G_T)} \leq C(M, M_1, c_0) \tag{106}$$

provided T is small enough; the smallness of T is required also to ensure that

$$g(\theta, t) \geq c_0, \qquad |\ell + m| \leq \pi - c_2, \qquad \ell - m \geq c_1$$

in Ω_T (so that (103) and (105) determine k and ζ as smooth functions of ℓ and m); (101) is then also satisfied.

We next show that $k = g_\theta$. Differentiating (85) formally in θ and comparing with (86), we easily get

$$(g_\theta - k)_t = P(\zeta)F\frac{k^2}{g^2\sqrt{g^2 + k^2}}(g_\theta - k)$$

and, since $(g_\theta - k) = 0$ at $t = 0$, we conclude that $g_\theta \equiv k$. The differentiation in θ in (85) can be justified by first integrating in t and then differentiating in θ.

Since $g_\theta = k$, the estimates (106) imply (100). To prove (99), we consider the characteristics for ℓ given by

$$\frac{d\xi(\theta, t)}{dt} = a(\xi, t), \qquad \xi(\theta, t) = \theta. \tag{107}$$

It follows that

$$\frac{d\xi_\theta}{dt} = a_\xi(\xi, t)\xi_\theta, \qquad \xi_\theta(\theta, 0) = 1,$$

and therefore

$$\exp\left[-C(M, M_1, c_0)t\right] \leq \xi_\theta \leq \exp\left[C(M, M_1, c_0)t\right],$$

and so

$$\frac{1}{2} \leq \xi_\theta(\theta, t) \leq 2 \tag{108}$$

if T is small enough. This allows us to establish the C^α estimate of ℓ_θ:

$$\|\ell\|_{C_\theta^{1+\alpha}(G_T)} \leq C(M, c_0);$$

the constant $C(M, c_0)$ is independent of M_1. A similar estimate can be established for m, and together they yield the assertion (99). □

Remark 6.1. The fact that the constant $\tilde{C}(M, c_0)$ is independent of M_1 is crucial for establishing existence for the full nonlinear problem.

Proof of Lemma 6.2. The proof follows by standard arguments (cf. [1]), integrating along characteristics and applying the Gronwall inequality. □

Remark 6.2. From (100), we have

$$\begin{aligned}[g_t]_{C_t^{\alpha/2}(G_T)} + [D_\theta^2 g]_{C_t^{\alpha/2}(G_T)} + [\zeta_t]_{C_t^{\alpha/2}(G_T)} + [D_\theta \zeta]_{C_t^{\alpha/2}(G_T)} \\ \le C(M, M_1, c_0)T^{\alpha/2} \le 1\end{aligned} \tag{109}$$

if T is small enough.

Since $0 \le F \le A^*$, (84), (85), and (99) imply that

$$\|g_t\|_{C(G_T)} + \|\zeta_t\|_{C(G_T)} \le C(M, c_0, A^*), \tag{110}$$

$$\|g_t\|_{C_\theta^\alpha(G_T)} \le C(M, c_0)\left(1 + \|F\|_{C_\theta^\alpha(G_T)}\right). \tag{111}$$

Combining estimates (109)–(111), we then have

$$\|\zeta\|_{C^{1+\alpha, (1+\alpha)/2}(G_T)} \le \hat{C}(M, c_0, A^*), \tag{112}$$

$$\|g\|_{C^{2+\alpha, 1+\alpha/2}(G_T)} \le \hat{C}(M, c_0)\left[1 + \|F\|_{C_\theta^\alpha(G_T)}\right]. \tag{113}$$

We now consider the full nonlinear problem with initial conditions (94) satisfying (95) and with

$$A|_{t=0} = A_0(r, \theta), \qquad B|_{t=0} = B_0(r, \theta), \qquad C|_{t=0} = C_0(r, \theta) \tag{114}$$

for $r \ge g_0(\theta)$, where

$$A_0 \equiv 0, \quad B \equiv B^*, \quad C_0 \equiv 0 \quad \text{for} \quad r > R^* \tag{115}$$

for some large $R^* > R_0$. We further assume that

$$0 \le A_0 < A^*, \qquad 0 \le B_0 \le B^*, \qquad C_0 \ge 0, \tag{116}$$

$$\|A_0\|_{C^{2+\alpha}} + \|B_0\|_{C^{2+\alpha}} + \|C_0\|_{C^{2+\alpha}} \le M \tag{117}$$

in the region $\sqrt{x_1^2 + x_2^2} \ge g_0(\theta)$, and

$$\begin{aligned} D_A \frac{\partial A_0}{\partial n} + P(\zeta_0)(A_0 - A^*) = 0, \\ \frac{\partial B_0}{\partial n} = 0, \qquad D_C \frac{\partial C_0}{\partial n} = -C_0^4 \quad \text{on} \quad r = g_0(\theta). \end{aligned} \tag{118}$$

For simplicity, we have taken the M in (117) to be the same as (95).

THEOREM 6.3. *The full nonlinear problem with initial data satisfying* (95), (97), *and* (115)–(118) *has a unique classical solution for* $0 \le t \le T$ *for some small enough* $T > 0$; T *depends only on the constants appearing in conditions* (95) *and* (97).

The solution is such that g and ζ belong to $C_\theta^{1+\alpha}(G_T)$.

Proof. We shall prove existence by a fixed-point argument for a mapping W defined in the class

$$\begin{aligned} K = \Big\{(g, \zeta)\,;\ \|g\|_{C_\theta^{2+\alpha}(G_T)} \le \tilde{C}(M, c_0), \quad \|\zeta\|_{C^{1+\alpha, (1+\alpha)/2}(G_T)} \le \hat{C}(M, c_0, A^*), \\ \|g\|_{C^{2+\alpha, 1+\alpha/2}(G_T)} \le \hat{C}(M, c_0) 2\pi \left[1 + A^* + (1 + M)(1 + \tilde{C}(M, c_0))\right], \\ g|_{t=0} = g_0(\theta), \quad \zeta|_{t=0} = \zeta_0(\theta), \quad g(\theta, t) \ge c_0, \quad \zeta(\theta, t) \ge c_0\Big\}, \end{aligned}$$

where the constants $\hat{C}$ and $\tilde{C}$ are taken from (112), (113), and (99).

Given $(g, \zeta) \in K$, we substitute it into the boundary condition for A and then solve for A, B, and C. By the maximum principle,

$$0 \le A(r, \theta, t) \le A^*, \qquad 0 \le B(r, \theta, t) \le B^*$$

and hence

$$0 \le C(r, \theta, t) \le C(M, c_0),$$

where the constant is independent of T.

By the Schauder estimates [7]

$$\|A\|_{C^{2+\alpha,\,1+\alpha/2}(\Omega_T)} + \|B\|_{C^{2+\alpha,\,1+\alpha/2}(\Omega_T)} + \|C\|_{C^{2+\alpha,\,1+\alpha/2}(\Omega_T)} \le C(M, c_0) \tag{119}$$

with another constant $C(M, c_0)$ independent of T.

Consider the functions

$$F(\theta, t) = A^* - A(g(\theta, t), \theta, t), \qquad E(\theta, t) = C^4(g(\theta, t), \theta, t).$$

Then $0 \le F \le A^*$ and

$$\begin{aligned}[F]_{C^\alpha_\theta(G_T)} &\le 2\pi \|F_\theta\|_{C(G_T)} \\ &\le 2\pi \left[\|A_r\|_{C(\Omega_T)} \|g_\theta\|_{C(G_T)} + \|A_\theta\|_{C(G_T)}\right]\end{aligned}$$

and, by (119), (117), and the definition of K, we easily get

$$[F]_{C^\alpha_\theta(G_T)} \le 2\pi(M+1)(1+\tilde{C}(M, c_0)). \tag{120}$$

It is also clear that

$$\|F\|_{C^{2+\alpha}_\theta(G_T)} + \|E\|_{C^{1+\alpha}_\theta(G_T)} \le C_1(M, c_0) \equiv M_1. \tag{121}$$

We now solve (84) and (85) with (F, E) as above and denote the solution by $(\bar{g}, \bar{\zeta})$, and we then define a mapping W by

$$W(g, \zeta) = (\bar{g}, \bar{\zeta}).$$

By Lemma 6.1 and the estimates (112), (113), and (120), W maps K into itself.

Introduce a topology on K by the norm

$$\begin{aligned}\|(g, \zeta)\| = \|g\|_{C(G_T)} + \|g_\theta\|_{C(G_T)} + \|g_{\theta\theta}\|_{C(G_T)} \\ + \|\zeta\|_{C(G_T)} + \|\zeta_\theta\|_{C(G_T)}.\end{aligned}$$

Then K is a closed convex set and, by (100), $W : K \to K$ is compact. Since the mapping W is uniquely defined, it is then also continuous, and so, by Schauder's fixed-point theorem, W has a fixed point. This gives a solution to the full nonlinear problem.

To prove uniqueness suppose $(\hat{A}, \hat{B}, \hat{C}, \hat{g}, \hat{\zeta})$ is another solution. Then

$$\|\hat{\ell}_\theta\|_{C^\alpha_\theta(G_T)} \le C, \qquad \|\hat{m}_\theta\|_{C^\alpha_\theta(G_T)} \le C. \tag{122}$$

Using the differential equations for g, ℓ, and m, (122), and (108), we easily find that

$$\begin{aligned}&\|\zeta-\hat{\zeta}\|_{C_t^{\alpha}(G_T)}+\|g_\theta-\hat{g}_\theta\|_{C_t^{\alpha}(G_T)}+\|g_t-\hat{g}_t\|_{C(G_T)}\\ &\quad\le C\left[\|F-\hat{F}\|_{C_\theta^{1+\alpha}(G_T)}+\|E-\hat{E}\|_{C_\theta^{\alpha}(G_T)}\right]\equiv J.\end{aligned}\tag{123}$$

Hence

$$\|\zeta-\hat{\zeta}\|_{C_t^{\alpha/2}(G_T)}+\|g_\theta-\hat{g}_\theta\|_{C_t^{\alpha/2}(G_T)}\le CT^{\alpha/2}J.$$

Using also Lemma 6.2, we get

$$\|\zeta-\hat{\zeta}\|_{C^{\alpha,\,\alpha/2}(G_T)}+\|g-\hat{g}\|_{C^{1+\alpha,\,(1+\alpha)/2}(G_T)}\le CT^{\alpha/2}J.\tag{124}$$

We now proceed as in [3, §8]. Let

$$V(T)=\|g-\hat{g}\|_{L^\infty(G_T)}$$

and consider A and $\hat{A}$ in the common domain

$$\tilde{\Omega}_T=\{r>\hat{g}(r,\theta)+V(T)\}.$$

As in [3],

$$\begin{aligned}&\left\|\left[D_A\frac{\partial(A-\hat{A})}{\partial n}+P(\zeta)(A-\hat{A})\right]_{r=\hat{g}(\theta,t)+V(T)}\right\|_{C^{\alpha,\,\alpha/2}(G_T)}\\ &\quad\le C\left\{\|g-\hat{g}\|_{C^{1+\alpha,\,(1+\alpha)/2}(G_T)}+\|\zeta-\hat{\zeta}\|_{C^{\alpha,\,\alpha/2}(G_T)}\right\}.\end{aligned}$$

Similar estimates hold for B and C. By parabolic $C^{1+\alpha,\,(1+\alpha)/2}$ estimates [8], we then have

$$\begin{aligned}&\|A-\hat{A}\|_{C_\theta^{1+\alpha}(\tilde{\Omega}_T)}+\|B-\hat{B}\|_{C_\theta^{1+\alpha}(\tilde{\Omega}_T)}+\|C-\hat{C}\|_{C_\theta^{1+\alpha}(\tilde{\Omega}_T)}\\ &\quad\le C\left[\|g-\hat{g}\|_{C^{1+\alpha,\,(1+\alpha)/2}(G_T)}+\|\zeta-\hat{\zeta}\|_{C^{\alpha,\,\alpha/2}(G_T)}\right]\equiv L(T).\end{aligned}$$

Using this and the $C_\theta^{2+\alpha}$ regularity of A and $\hat{A}$, we find that

$$\|F-\hat{F}\|_{C_\theta^{1+\alpha}(\Omega_T)}\le CL(T).$$

A similar estimate holds for $E-\hat{E}$. Substituting these two estimates into (124) (recall the definition of J in (123)), we see that

$$L(T)\le CT^{\alpha/2}L(T)$$

and, therefore, $L(T)=0$ if T is small enough. Similarly, $L(t)\equiv 0$ if $0\le t\le T$ and then also $A=\hat{A}$, $B=\hat{B}$, and $C=\hat{C}$ if $0\le t\le T$.

The uniqueness proof can be extended to any time interval for which the solutions exist. □

Acknowledgment. The authors thank David Ross from Eastman Kodak for suggesting the problem studied in this paper and for stimulating discussions.

REFERENCES

[1] A. Douglas, *Existence theorems for hyperbolic systems*, Comm. Pure Appl. Math., 5 (1952), pp. 119–154.

[2] A. Friedman, *Mathematics in Industrial Problems, Part* 6, IMA Math. Appl., vol. 57, Springer-Verlag, New York, 1993.

[3] A. Friedman and B. Hu, *The Stefan problem with kinetic condition at the free boundary*, Ann. Scuola. Norm. Sup. Pisa Cl. Sci. (4), 19 (1992), pp. 87–111.

[4] A. Friedman, J. Zhang and D. S. Ross, *A Stefan problem for a reaction-diffusion system*, SIAM J. Math. Anal., 26 (1995), pp. 1089–1112.

[5] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[6] P. Hartman and A. Wintner, *On hyperbolic partial differential equations*, Amer. J. Math., 74 (1952), pp. 834–864.

[7] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[8] G. M. Lieberman, *Hölder continuity of the gradient of solutions of uniformly parabolic equations with conormal boundary conditions*, Ann. Mat. Pura Appl. (4), 148 (1987), pp. 77–99.

[9] K.-O. Widman, *Inequalities for the Green function and boundary continuity of the gradient of solutions of elliptic differential equations*, Math. Scand., 21 (1967), pp. 17–37.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1235–1260, September 1996

003

ON THE EXISTENCE OF SOLUTIONS OF THE CAUCHY PROBLEM FOR A DOUBLY NONLINEAR PARABOLIC EQUATION*

KAZUHIRO ISHIGE[†]

Abstract. For the existence of weak solutions of

$$\frac{\partial}{\partial t}(|u|^{\beta-1}u) = \operatorname{div}(|\nabla u|^{p-2}\nabla u) \qquad \text{with} \quad |u|^{\beta-1}u(\cdot,0) = \mu(\cdot),$$

we give a sufficient condition for the growth order of the initial data $\mu(x)$ as $|x| \to \infty$

Key words. Cauchy problem, doubly nonlinear parabolic equation

AMS subject classifications. 35K55, 35K57

1. Introduction.

We investigate the Cauchy problem for the following nonlinear diffusion equation:

$$(1.1) \qquad \frac{\partial}{\partial t}(|u|^{\beta-1}u) = \operatorname{div}(|\nabla u|^{p-2}\nabla u) \qquad \text{in} \quad \mathbf{S}_T, \quad \beta > 0, \ p > 1,$$

$$(1.2) \qquad |u|^{\beta-1}u(\cdot,0) = \mu(\cdot) \qquad \text{in} \quad \mathbf{R}^N.$$

Here $\mathbf{S}_T = \mathbf{R}^N \times (0,T)$, $0 < T < \infty$, and μ is an $L^1_{\text{loc}}(\mathbf{R}^N)$ function or a σ-finite Borel measure.

Equation (1.1) is called a doubly nonlinear parabolic equation, which contains the heat equation (i.e., $\beta = 1$, $p = 2$), the porous-medium equation (i.e., $\beta > 0$, $p = 2$), and the p-Laplacian equation (i.e., $\beta = 1$, $p > 1$). Equation (1.1) has been studied by several authors; for example, see [9], [11]–[13], [15], [16], [18], and [19]. We divide the Cauchy problem (1.1) with (1.2) into three cases,

$$\text{(I)} \quad (p-1)/\beta > 1, \qquad \text{(II)} \quad (p-1)/\beta = 1, \qquad \text{(III)} \quad 0 < (p-1)/\beta < 1,$$

and study the existence of the solution, respectively. For cases (I), (II), and (III), the behavior of solutions of (1.1) is completely different, and so we need this separation. Case (I) contains the so-called degenerate cases of the porous-medium and p-Laplacian equations, and case (III) contains the singular cases of these. In what follows, we call (I) the degenerate case and (III) the singular case, respectively.

A classical result of A. N. Tychonov [17] states that the Cauchy problem for the heat equation, $u_t = \Delta u$, has a unique classical solution in the strip $\mathbf{S}_T$ for continuous initial data $\mu(x)$ satisfying

$$(1.3) \qquad |\mu(x)| \le C \exp(|x|^2/4T) \qquad \text{as} \quad |x| \to \infty.$$

Moreover, D. G. Aronson [1] generalized the result of A. N. Tychonov for a parabolic operator with variable coefficients:

$$(1.4) \qquad \mathcal{L}u \equiv \frac{\partial}{\partial t}u - \frac{\partial}{\partial x_j}\left\{A_{ij}(x,t)\frac{\partial}{\partial x_i}u + A_j(x,t)u\right\}$$

*Received by the editors June 29, 1994; accepted for publication (in revised form) May 2, 1995..

[†]Department of Mathematics, Faculty of Science, Tokyo Institute of Technology, Oh-Okayama, Meguro-ku, Tokyo 152, Japan. Current address: Graduate School of Information Sciences, Tohoku University, Aoba-ku, Sendai 980–77, Japan.

with suitable conditions imposed on coefficients $A_{ij}(x,t)$ and $A_j(x,t)$. For the Cauchy problem for the equation $\mathcal{L}u = 0$ with initial data $\mu(x)$ satisfying

$$\int_{\mathbf{R}^N} \mu(x)\exp(-\Lambda|x|^2)dx < \infty, \qquad \Lambda > 0, \tag{1.5}$$

he proved that it has a unique classical solution in some strip $\mathbf{S}_{T'}$, where T' is a constant dependent on Λ. Furthermore, he proved that the solution u is written in the strip $\mathbf{S}_{T'}$ in the form

$$u(x,t) = \int_{\mathbf{R}^N} \Gamma(x,t;\xi,0)u_0(\xi)d\xi,$$

where Γ is the fundamental solution of $\mathcal{L}u = 0$. See also [2], [14], and [20].

For the degenerate case of the porous-medium equation,

$$u_t = \Delta(u^m), \quad m > 1, \tag{1.6}$$

P. Benilan, M. G. Crandall, and M. Pierre [4] proved that the Cauchy problem is uniquely solvable in the sense of weak solutions for initial data satisfying

$$\limsup_{\rho\to\infty} \rho^{-N-2/(m-1)} \int_{B_\rho} d|\mu| < \infty, \tag{1.7}$$

where B_ρ is a ball of radius $\rho > 0$ with center 0. On the other hand, for the degenerate case of the p-Laplacian equation,

$$u_t = \operatorname{div}(|\nabla u|^{p-2}\nabla u), \quad p > 2, \tag{1.8}$$

E. DiBenedetto and M. A. Herrero [6] proved similar results for initial data satisfying

$$\limsup_{\rho\to\infty} \rho^{-N-p/(p-2)} \int_{B_\rho} d|\mu| < \infty. \tag{1.9}$$

Furthermore, E. DiBenedetto and M. A. Herrero [7] and E. DiBenedetto and T. C. Kwong [8] studied the Cauchy problem for singular cases of the porous-medium equation ($(N-2)^+/2 < m < 1$) and the p-Laplacian equation ($2N/(N+1) < p < 2$), respectively, and obtained $L^\infty_{\mathrm{loc}}(\mathbf{R}^N)$ estimates of the solution for $L^1_{\mathrm{loc}}(\mathbf{R}^N)$ initial data.

Our purpose in this paper is to extend earlier results on the existence of solutions of the heat, porous-medium, and p-Laplacian equations to the doubly nonlinear parabolic equation (1.1). The main point of this paper is to treat case (II).

For case (II), if the initial data μ satisfies

$$\int_{\mathbf{R}^N} \exp(-\Lambda|x|^{p/(p-1)})d|\mu| < \infty \tag{1.10}$$

for some constant $\Lambda > 0$, then we prove that there exists a weak solution of (1.1) with (1.2) in the strip $\mathbf{S}_{T(\Lambda)}$, where $T(\Lambda) = (p-1)^{2(p-1)}\Lambda^{1-p}/p^p$. Here we remark that there exists a solution of (1.1) with initial data satisfying (1.10) which cannot be extended to a strip larger than $\mathbf{S}_{T(\Lambda)}$. Case (II) contains the heat equation, but our proof is completely different from that of [1] and [17]. In fact, in the proof of case (II), we do not use the fundamental solution of the heat equation, and the proof depends only on the structure conditions of the equation. Therefore our proof is applicable to more general equations than that of [1].

For the proof of case (II), we essentially use the techniques given in [5]–[8]. However, it seems difficult to apply them directly to case (II). To overcome this difficulty, we introduce a new function with weight ϕ_Λ (see equation (2.8) below) and obtain an $L^1(\mathbf{R}^N)$ estimate of ϕ_Λ (see (2.10)). By the estimate of $\|\phi_\Lambda(\cdot,t)\|_{L^1(\mathbf{R}^N)}$, we can estimate $\|u(\cdot,t)\|_{L^\infty(\bar{B}_\rho)}$ and $\|\nabla u(\cdot,t)\|_{L^{p-1}(B_\rho)}$ and prove that the optimal growth order of the initial data for this case is exponential.

For case (I), we will prove that there exists a weak solution of (1.1) under initial data satisfying

$$\limsup_{\rho\to\infty} \rho^{-N-p/d} \int_{B_\rho} d|\mu| < \infty, \tag{1.11}$$

with $d = (p-1)/\beta - 1$. Furthermore, for case (III) with $Nd+p>0$, we will give $L^\infty_{\text{loc}}(\mathbf{R}^N)$ estimates of the solution for $L^1_{\text{loc}}(\mathbf{R}^N)$ initial data. For cases (I) and (III), our proof depends heavily on an approach used in [5]–[8]. Recently, for case (III) of equation (1.1), V. Vespri [19] proved several inequalities through which the existence of solutions is proved.

Finally, we remark that—to our knowledge—there are no results for the uniqueness of weak solutions of (1.1), though the uniqueness of strong solutions of (1.1) is given in [9] and [16].

2. The main results.

In this section, we give the definition of a weak solution of (1.1) with (1.2) and state our results.

DEFINITION 1.1. *A measurable function $u(x,t)$ defined in $\mathbf{R}^N \times (0,T)$ is a weak solution of* (1.1) *and* (1.2) *if for any $\epsilon \in (0,T)$ and any bounded open set $\Omega \subset \mathbf{R}^N$, $|\nabla u|^{p-1} \in L^1(\Omega\times(0,T_\epsilon))$, $|u|^{\beta-1}u \in C(0,T_\epsilon;L^1(\Omega))$, and*

$$\int_\Omega |u|^{\beta-1}u\varphi(x,t)dx + \int_0^t\int_\Omega \{-|u|^{\beta-1}u\varphi_t + |\nabla u|^{p-2}\nabla u\cdot\nabla\varphi\}dxd\tau = \int_\Omega \varphi(x,0)d\mu \tag{2.1}$$

for all $0<t<T_\epsilon$ and all testing functions

$$\varphi \in W^{1,\infty}(0,T_\epsilon;L^\infty(\Omega)) \cap L^\infty(0,T_\epsilon;W_0^{1,\infty}(\Omega)). \tag{2.2}$$

Here $T_\epsilon = T-\epsilon$.

Throughout this paper, we set

$$m = 1/\beta, \qquad d = (p-1)/\beta - 1 = m(p-1)-1, \qquad \kappa_r = Nd+rp.$$

In particular, we set

$$\kappa = \kappa_1 = Nd+p, \qquad \kappa^* = \kappa_{mp} = Nd+mp^2$$

for simplicity. Furthermore, by $C = C(A_1,A_2,\dots)$ we denote a positive constant which depends only $\beta,p,A_1,A_2,\dots$.

Case (I): *The degenerate case* $(d > 0)$. In order to represent the growth order of the initial data $\mu(x)$, we define $|||f|||_r$ by

$$|||f|||_r = \sup_{\rho \geq r} \rho^{-\kappa/d} \int_{B_\rho} |f| dx \tag{2.3}$$

for $f \in L^1_{\mathrm{loc}}(\mathbf{R}^N)$. This norm is a modification of the one introduced in [6].

THEOREM 1.2. *Let $d > 0$ and μ be a σ-finite Borel measure in $\mathbf{R}^N$ satisfying*

$$|||\mu|||_r < \infty \qquad \textit{for some} \quad r > 0.$$

Then there exists a weak solution u of (1.1) *and* (1.2) *in the strip $\mathbf{S}_{T(\mu)}$, where*

$$T(\mu) = \begin{cases} C_1[\lim_{r\to\infty} |||\mu|||_r]^{-d} & \textit{if} \quad \lim_{r\to\infty} |||\mu|||_r > 0, \\ +\infty & \textit{if} \quad \lim_{r\to\infty} |||\mu|||_r = 0, \end{cases} \tag{2.4}$$

and $C_1 = C_1(N, \beta, p)$.

Let $T_r(\mu) = C_1 |||\mu|||_r^{-d}$. Then for any $t \in (0, T_r(\mu))$ and $\rho > 0$,

$$|||\, |u|^\beta(\cdot, t)\, |||_r \leq C_2 |||\mu|||_r, \tag{2.5}$$

$$\|u(\cdot, t)\|_{L^\infty(B_\rho)} \leq C_3 t^{-N/\beta\kappa} \rho^{p/\beta d} |||\mu|||_r^{p/\beta\kappa}, \tag{2.6}$$

and

$$\int_0^t \int_{B_\rho} |\nabla u|^{p-1} dx d\tau \leq C_4 t^{1/\kappa} \rho^{1+\kappa/d} |||\mu|||_r^{1+d/\kappa}, \tag{2.7}$$

where $C_i = C_i(N, \beta, p)$, $i = 2, 3, 4$.

Case (II): *The critical case* $(d = 0)$. For any $\lambda > 0$ and $\delta > 0$, let $\phi_\lambda(t)$ be a function defined by

$$\phi_\lambda(t) = \sup_{\tau \in (0,t)} \int_{\mathbf{R}^N} \eta(|x|) F(e^{g_\lambda} u^{p-1}) dx, \tag{2.8}$$

where

$$F(s) = \begin{cases} |s|^{1+\delta}/(1+\delta) & \text{if} \quad |s| \leq 1, \\ |s| - \delta/(1+\delta) & \text{if} \quad |s| \geq 1, \end{cases} \qquad \eta(s) = \begin{cases} 1 & \text{if} \quad s \leq 1, \\ s^{(N+p)/(p-1)} & \text{if} \quad s \geq 1, \end{cases}$$

and

$$g_\lambda(x, t) = -\lambda \left(\frac{|x|^p}{1-t} \right)^{1/(p-1)} (1 + t^l), \qquad 0 < l < 1/2. \tag{2.9}$$

Then our result for Case (II) is as follows.

THEOREM 1.3. *Let $d = 0$ and μ be a σ-finite Borel measure satisfying*

$$\int_{\mathbf{R}^N} \exp(-\Lambda |x|^{p/(p-1)}) d|\mu(x)| < \infty \tag{2.10}$$

for some $\Lambda > 0$. *Then there exists a weak solution* u *of* (1.1) *and* (1.2) *in the strip* $\mathbf{S}_{T(\Lambda)}$, *where* $T(\Lambda)$ *is a constant such that*

$$T(\Lambda) = \frac{(p-1)^{2(p-1)}}{p^p}\Lambda^{1-p}. \tag{2.11}$$

Furthermore, let $\delta > 0$ *be a sufficiently small constant. Then for any* $\lambda > \Lambda$, *there exists a constant* $T_0 > 0$ *such that* u *satisfies the following inequalities:*

$$\phi_\lambda(t) \le C_1\phi_\lambda(0) \le C\int_{\mathbf{R}^N} \exp(-\Lambda|x|^{p/(p-1)})d|\mu(x)|, \tag{2.12}$$

$$\|e^{g_\lambda(\cdot,t)}|u|^{p-1}\|_{L^\infty(\mathbf{R}^N)} \le C_2(1+t^{-N/p})\phi_\lambda(t) \tag{2.13}$$

for all $t \in (0,T_0)$, *where* $C = C(p,N,\Lambda,\lambda,\delta)$ *and* $C_i = C_i(p,N,\delta)$, $i = 1,2$.

The estimate of $T(\Lambda)$ in (2.11) is optimal in the sense that there exists a solution blowing up at $T(\Lambda)$. In fact, let $u(x,t)$ be a function such that

$$u(x,t) = (1-\sigma t)^{-N/p(p-1)}\exp\left[\gamma\left(\frac{|x|^p}{1-\sigma t}\right)^{1/(p-1)}\right], \qquad \gamma = \frac{p-1}{p}\left(\frac{\sigma}{p}\right)^{1/(p-1)},$$

where σ is any positive constant. Then $u(x,t)$ is a solution of (1.1) in the strip $\mathbf{S}_{1/\sigma}$, and

$$u^{p-1}(x,0) = \exp((p-1)\gamma|x|^{p/(p-1)}).$$

Then $T((p-1)\Lambda) = 1/\sigma$ and $u(x,t)$ blows up at $t = 1/\sigma$.

Case (III): *The singular case* ($d < 0$). We treat only the result for the case where $\kappa > 0$.

THEOREM 1.4. *Let* $d < 0$ *such that* $\kappa > 0$ *and let* μ *be a* σ*-finite Borel measure in* $\mathbf{R}^N$. *Then there exists a weak solution* u *of* (1.1) *and* (1.2) *in* $\mathbf{R}^N \times (0,\infty)$. *Furthermore, the solution* $u(x,t)$ *satisfies the following inequalities:*

$$\|\,|u|^\beta(\cdot,t)\,\|_{L^\infty(B_\rho)} \le C_1 t^{-N/\kappa}\left(\sup_{0<\tau<t}\int_{B_{2\rho}} |u|^\beta(x,\tau)dx\right)^{p/\kappa} + C_2\left(\frac{t}{\rho^p}\right)^{p/\kappa}$$

and

$$\sup_{0<\tau<t}\int_{B_\rho} |u|^\beta(x,\tau)dx \le C_3\int_{B_{2\rho}} d|\mu(x)| + C_4\left(\frac{t}{\rho^\kappa}\right)^{1/-d}$$

for any $t > 0$ *and* $\rho > 1$, *where* $C_i = C_i(N,m,p)$, $i = 1,2,3,4$.

The essential part of Theorem 1.4 was proved by V. Vespri. See Theorems 2–1 and 2–2 in [19].

We remark that the estimates of solutions given in Theorems 1.2–1.4 may be extended to nonnegative strong subsolutions of

$$\frac{\partial}{\partial t}(|u|^{\beta-1}u) - \mathrm{div}A(x,t,u,\nabla u) \le B(x,t,u,\nabla u). \tag{2.14}$$

Here the structure conditions below are satisfied:

$$\begin{cases} C_1|q|^p - g_1(x,t) \le A(x,t,u,q)\cdot q \le C_2|q|^p + g_2(x,t), \\ [A(x,t,u,q) - A(x,t,u,\tilde{q})]\cdot(q-\tilde{q}) \ge 0, \\ |B(x,t,u,q)| \le C_3|q|^{p-1} + g_3(x,t) \end{cases} \tag{2.15}$$

for any $(x,t,u) \in \mathbf{R}^N \times \mathbf{R}_+ \times \mathbf{R}$ and $q, \tilde{q} \in \mathbf{R}^N$, where C_i, $i = 1,2,3$, are given constants and g_i, $i = 1,2$, are given bounded functions in $\mathbf{R}^{N+1}$.

Furthermore, we remark that it is not restrictive to treat nonnegative continuous strong solutions of (1.1). In fact, by the results of [12] and [18] and the method of construction of the solutions in this paper, we can take a solution u as a continuous function. Moreover, the solution u is approximated by solutions u_n of suitable approximate equations (see (3.4)), which satisfy the structure conditions in (2.15). Then by Lemma 1–2 in [5], $u_{n+} = \max\{u_n, 0\}$ and $u_{n-} = -\min\{u_n, 0\}$ are continuous nonnegative strong subsolutions of the approximate equations, and the estimates of Theorems 1.2–1.4 hold for nonnegative strong subsolutions of (2.14). Therefore, throughout this paper, we treat only nonnegative continuous solutions of (1.1). See §3.

The organization of this paper is as follows. In §§3 and 4, we prove Theorem 1.3. In §3, we prove (2.11) for the $L^\infty_{\mathrm{loc}}(\mathbf{R}^N)$ initial data satisfying (2.10). In §4, we study the behavior of $\phi_\lambda(t)$ and complete the proof of Theorem 1.3. In §5, we prove Theorem 1.2 by arguments similar to those of [5]–[8]. In §6, via [7] and [19], we present some lemmas to prove Theorem 1.4.

3. Existence of solutions for Case (II). In this section, we consider the Cauchy problem of

$$\begin{cases} \dfrac{\partial}{\partial t}(|u|^{p-2}u) = \operatorname{div}(|\nabla u|^{p-2}\nabla u) & \text{in} \quad \mathbf{R}^N \times (0,T), \quad p > 1, \\ |u|^{p-2}u(x,0) = \mu(x) \in L^\infty_{\mathrm{loc}}(\mathbf{R}^N) & \text{in} \quad \mathbf{R}^N \end{cases} \tag{3.1}$$

under the condition in (2.10) and prove the following proposition. To prove Proposition 3.1, we need several lemmas.

PROPOSITION 3.1. *Let μ be a $L^\infty_{\mathrm{loc}}(\mathbf{R}^N)$ function such that*

$$\mu(x)\exp(-\Lambda|x|^{p/(p-1)}) \in L^\infty(\mathbf{R}^N).$$

Then there exists a weak solution $u(x,t)$ of (3.1) *in $\mathbf{S}_{T(\Lambda)}$, where $T(\Lambda)$ is a constant given in* (2.11). *Furthermore, for any $\epsilon \in (0, T(\Lambda))$, there exist constants C_1^ϵ and C_2^ϵ such that the solution $u(x,t)$ satisfies*

$$\begin{aligned} \sup_{0<\tau<T_\epsilon} \int_{\mathbf{R}^N} \exp(-C_1^\epsilon|x|^{p/(p-1)})u^p(x,\tau)dx + \iint_{\mathbf{S}_{T_\epsilon}} \exp(-C_1^\epsilon|x|^{p/(p-1)})|\nabla u|^p dx d\tau \\ \le C_2^\epsilon \, \|\mu\exp(-\Lambda|\cdot|^{p/(p-1)})\|_{L^\infty(\mathbf{R}^N)}, \end{aligned}$$

where $T_\epsilon = T(\Lambda) - \epsilon$.

To simplify notation, we introduce a functional space with weight $L^q_\lambda(\mathbf{R}^N)$. For $f \in L^q_{\mathrm{loc}}(\mathbf{R}^N)$ with $q \in [1,\infty]$, we define

$$\|f\|_{q,\lambda} \equiv \|\exp(-\lambda|\cdot|^{p/(p-1)})f(\cdot)\|_{L^q(\mathbf{R}^N)}, \qquad \lambda > 0,$$

and we say that $f \in L^q_\lambda(\mathbf{R}^N)$ if $\|f\|_{q,\lambda} < \infty$.

We begin by proving the existence of solutions of (3.1) for $\mu \in C_0^\infty(\mathbf{R}^N)$. The following lemma is an extension of Theorem 3 in [1].

LEMMA 3.2. *Let $\mu \in C_0^\infty(\mathbf{R}^N)$. Then there exists a solution $u(x,t)$ of* (3.1) *in the strip $\mathbf{S}_\infty$ which satisfies $0 \le \|u\|_{L^\infty(\mathbf{R}^N\times(0,\infty))} \le \|\mu\|_{L^\infty(\mathbf{R}^N)}$.*

Let

$$h_{\gamma,T}(x,t) = -\gamma\left(\frac{|x|^p}{T-t}\right)^{1/(p-1)}, \qquad \gamma > 0. \tag{3.2}$$

Then for any $\lambda > 0$, *there exist constants* γ_0 *and* C *dependent only on* p *such that*

$$\sup_{0<\tau<T_\lambda} \int_{\mathbf{R}^N} e^{ph_{\gamma_0,T_\lambda}(\cdot,t)/(p-1)} u^p dx + \iint_{\mathbf{S}_{T_\lambda}} |e^{h_{\gamma_0,T_\lambda}(x,\tau)/(p-1)} \nabla u|^p dx d\tau \le C\|\mu\|_{\frac{p}{p-1},\lambda}, \tag{3.3}$$

where $T_\lambda = \lambda^{1-p}\gamma_0^{p-1}$.

Proof. By $\theta_n(s)$ we denote a sequence of $C^1(\mathbf{R})$ functions such that $\theta(s) \to |s|^{\beta-1}s$ and $\theta'_n(s) \to \beta|s|^{\beta-1}$ uniformly on any compact set of $\mathbf{R}\setminus\{0\}$, where $\beta = p-1$. Then we consider the Cauchy–Dirichlet problem,

$$\begin{cases} \dfrac{\partial}{\partial t}\theta_n(u) = \operatorname{div}\{(|\nabla u|^2 + 1/n)^{(p-2)/2}\nabla u\} & \text{in} \quad B_n(0) \times (0,T), \\ u(x,t) = 0 & \text{on} \quad \partial B_n(0) \times (0,T), \\ \theta_n(u)(x,0) = \mu(x) & \text{in} \quad B_n(0), \end{cases} \tag{3.4}$$

where $p > 1$ and $\mu \in C_0^\infty(\mathbf{R}^N)$. Let n be a sufficiently large integer n such that $\operatorname{supp}(\mu) \subset B_n$. By [12], [16], and [18], there exists a classical solution u_n of (3.4) in $B_n \times (0,\infty)$ such that $0 \le |u_n(x,t)| \le \|\mu\|_{L^\infty(\mathbf{R}^N)}$ and $|u_n(x,t) - u_n(y,s)| \le M(|x-y|^\alpha + |t-s|^{\alpha/p})$ for all $(x,t),(y,s) \in B_n(0) \times (0,\infty)$. Here M and α are constants which are independent of n and depend only on $\|u_n\|_{L^\infty(\mathbf{R}^N\times(0,\infty))}$.

Let $h = h_{\gamma_0,T_\lambda}$ for simplicity, and set

$$\varphi_n(x,t) = \exp\left(\frac{p}{p-1}h(x,t)\right)u_n(x,t).$$

Then we have the following two inequalities:

$$\begin{aligned} &\int_0^{T_\lambda}\int_{B_n} \frac{\partial}{\partial t}\theta_n(u_n)\varphi_n dx d\tau \\ &\quad \ge \int_{B_n} e^{ph/(p-1)}\Theta_n(u_n)dx\Bigg|_{\tau=0}^{\tau=T_\lambda} - \frac{p-1}{p}\int_0^{T_\lambda}\int_{B_n} e^{ph/(p-1)}\Theta_n(u_n)\frac{\partial}{\partial t}h dx d\tau \end{aligned}$$

and

$$\begin{aligned} &\int_0^{T_\lambda}\int_{B_n} (|\nabla u|^2 + 1/n)^{(p-2)/2}\nabla u \cdot \nabla\varphi_n dx d\tau \\ &\quad \ge \int_0^{T_\lambda}\int_{B_n} e^{ph/(p-1)}(|\nabla u_n|^2 + 1/n)^{p-2}|\nabla u|^2 dx d\tau \\ &\qquad - \frac{p}{p-1}\int_0^{T_\lambda}\int_{B_n} e^{ph/(p-1)}u_n(|\nabla u_n|^2 + 1/n)^{(p-2)/2}|\nabla u||\nabla h| dx d\tau \\ &\quad \ge \frac{1}{2}\int_0^{T_\lambda}\int_{B_n} e^{ph/(p-1)}|\nabla u_n|^p dx d\tau - \frac{2^{p-1}}{p-1}\int_0^{T_\lambda}\int_{B_n} e^{ph/(p-1)}u_n^p|\nabla h|^p dx d\tau, \end{aligned}$$

where $\Theta_n(s) = \int_0^{\theta_n(s)} \theta_n^{-1}(t)dt$. We multiply φ_n by (3.4) and integrate over $B_n \times (0, T_\lambda)$. Since $\lim_{n\to\infty} \Theta_n(s) = (p-1)|s|^p/p$ and the function h_{γ_0,T_λ} has the relation

$$-\frac{\partial}{\partial t} h_{\gamma_0,T_\lambda} = \frac{(p-1)^{p-1}}{p^p}\gamma_0^{1-p}|\nabla h_{\gamma_0,T_\lambda}|^p, \tag{3.5}$$

we take a sufficiently small γ_0 such that

$$(p-1)^{p-1}\gamma_0^{1-p}/p^p > 2^{p-1}/(p-1)$$

and have

$$\begin{aligned}\sup_{0<\tau<T_\lambda} \int_{B_n} e^{ph/(p-1)}\Theta_n(u_n)dx + \int_0^{T_\lambda}\int_{B_n} e^{ph/(p-1)}|\nabla u_n|^p dx d\tau \\ \le C(p)\int_{B_n} e^{ph(x,0)/(p-1)}\Theta_n(\mu_n)dx\end{aligned} \tag{3.6}$$

for sufficiently large integers n.

Taking the limit as $n \to \infty$, by inequality (3.6) and the definition of Θ_n, there exists a function $u \in L^p_{\mathrm{loc}}(0, T_\lambda : W^{1,p}_{\mathrm{loc}}(\mathbf{R}^N))$ such that

$$u_n \rightharpoonup u \qquad \text{in} \quad L^p_{\mathrm{loc}}(0, T_\lambda : W^{1,p}_{\mathrm{loc}}(\mathbf{R}^N)) \quad \text{weakly},$$

and u satisfies inequality (3.3) and $0 \le \|u\|_{L^\infty(\mathbf{R}^N \times (0,T_\lambda))} \le \|\mu\|_{L^\infty(\mathbf{R}^N)}$. Furthermore, by the Minty lemma (see [3]), u is a weak solution of (3.1) in $\mathbf{S}_{T_\lambda}$. By $\lim_{\lambda\to 0} T_\lambda = \infty$, the solution $u(x,t)$ exists in $\mathbf{S}_\infty$. □

By Lemma 3.2, for any $\mu \in L^\infty_\lambda(\mathbf{R}^N)$, there exists a solution of (3.1) in the strip $\mathbf{S}_{T_{\gamma_0}}$. In order to expand the strip $\mathbf{S}_{T_{\gamma_0}}$ to a larger strip, we need two lemmas.

LEMMA 3.3. *Let μ be a function in $C_0^\infty(\mathbf{R}^N)$ and $u(x,t)$ be a solution of* (3.1) *constructed in Lemma* 3.2. *For any $\delta > 0$, let γ_δ be a constant such that*

$$\gamma_\delta = (p-1)\delta/(1+\delta)^{p/(p-1)}. \tag{3.7}$$

Then for any $T > 0$, there exists a constant $C = C(p, \delta)$ independent of T such that

$$\sup_{0<t<T} \| |u|^{p-1}(\cdot,t)e^{h_{\gamma_\delta,T}(\cdot,t)}\|_{L^{1+\delta}(\mathbf{R}^N)} \le C\|\mu(\cdot)e^{h_{\gamma_\delta,T}(\cdot,0)}\|_{L^{1+\delta}(\mathbf{R}^N)}. \tag{3.8}$$

Furthermore, we have $\max_{\delta>0}\gamma_\delta = \gamma_{p-1}$.

Proof. Let u_n be a solution of the Cauchy–Dirichlet problem (3.5). For any $\epsilon > 0$, set

$$\varphi_n(x,t) = (e^{h_{\gamma_\delta,T}}(u_n+\epsilon)^{p-1})^\delta e^{h_{\gamma_\delta,T}(x,t)} - \epsilon^{\delta(p-1)} \exp\left(-(1+\delta)\gamma\left(\frac{n^p}{T-t}\right)^{1/(p-1)}\right).$$

Then for any $0 < t < T$, we have

$$\begin{aligned}&\int_0^t\int_{B_n} (|\nabla u_n|^2 + 1/n)^{(p-2)/2}\nabla u_n \cdot \nabla\varphi_n dx d\tau \\ &\ge \delta(p-1)\int_0^t\int_{B_n} e^{(1+\delta)h_{\gamma_\delta,T}}(u_n+\epsilon)^{\delta(p-1)-1}(|\nabla u_n|^2+1/n)^{(p-2)/2}|\nabla u_n|^2 dx d\tau \\ &\quad - \int_0^t\int_{B_n} e^{(1+\delta)h_{\gamma_\delta,T}}(u_n+\epsilon)^{\delta(p-1)}(|\nabla u_n|^2+1/n)^{(p-2)/2}|\nabla u_n||\nabla h_{\gamma_0,T}|dx d\tau \\ &\ge -\frac{(1+\delta)^p}{\delta^{p-1}p^p}\int_0^t\int_{B_n}(e^{h_{\gamma_\delta,T}}(u_n+\epsilon)^{p-1})^{1+\delta}|\nabla h_{\gamma_0,T}|^p dx d\tau.\end{aligned}$$

Here we use the Young inequality $ab^{p-1} \leq (\tilde{\epsilon}^{1-p}/p)a^p + ((p-1)/p)\tilde{\epsilon}b^p$ with $\tilde{\epsilon} = \delta p/(1+\delta)$. Thus we obtain the following inequalities:

$$\begin{aligned}\liminf_{\epsilon\to 0} \int_0^t \int_{B_n} \frac{\partial}{\partial t}\theta_n(u_n)\varphi_n dx d\tau &\geq \int_{B_n} e^{(1+\delta)h_{\gamma_\delta,T}(x,\tau)}\tilde{\Theta}_n(u_n)dx\Big|_{\tau=0}^{\tau=t} \\ &- (1+\delta)\int_0^t \int_{B_n} e^{(1+\delta)h_{\gamma_\delta,T}}\tilde{\Theta}_n(u_n)\frac{\partial}{\partial t}h_{\gamma_\delta,T}dx d\tau\end{aligned}$$

and

$$\begin{aligned}&\liminf_{\epsilon\to 0} \int_0^t \int_{B_n} (|\nabla u_n|^2 + 1/n)^{(p-2)/2}\nabla u_n \cdot \nabla\varphi_n dx d\tau \\ &\quad\geq -\frac{(1+\delta)^p}{\delta^{p-1}p^p}\int_0^t \int_{B_n} (e^{h_{\gamma_\delta,T}}u_n^{p-1})^{1+\delta}|\nabla h_{\gamma_0,T}|^p dx d\tau,\end{aligned}$$

where $\tilde{\Theta}_n(s) = \int_0^{\theta_n(s)}[\theta^{-1}(t)]^{\delta(p-1)}dt$. Taking the limit as $n \to \infty$, by (3.5) and (3.7), we see that there exists a constant $C(\delta)$ such that

$$\int_{B_n} (e^{h_{\gamma_\delta,T}}|u|^{p-1})^{1+\delta}dx\Big|_{\tau=t} \leq C(\delta)\int_{B_n} (e^{h_{\gamma_\delta,T}}(x,0)\mu(x))^{1+\delta}dx,$$

and so the proof of Lemma 3.3 is complete. □

The following lemma is proved by the arguments similar to those of [5]–[8].

LEMMA 3.4. *Let u be a solution of* (3.1) *constructed in Lemma* 3.2. *For any $\lambda > 0$ and $r > 1$, there exist constants $C = C(r,\lambda)$ and $T_\lambda^* < 1$ such that*

$$\begin{aligned}&\|\eta(|x|)e^{g_\lambda}u^{p-1}\|_{L^\infty(B_{R_1}\times(t_1,t))} \\ &\quad\leq CM^{(N+p)/rp}\left(\int_{t_2}^t \int_{B_{R_2}} [\eta(|x|)e^{g_\lambda}u^{p-1}]^r dx d\tau\right)^{1/r}\end{aligned} \tag{3.9}$$

for any R_1 and R_2 with $0 < R_1 < R_2$ and any t, t_1, and t_2 with $0 < t_2 < t_1 < t \leq T_\lambda^*$, where

$$M = |\nabla g_\lambda|^p(R_2,t) + (R_2 - R_1)^{-p} + (t_1 - t_2)^{-1}. \tag{3.10}$$

Here η and g_λ are functions in (2.8).

Proof. We first assume that u is a nonnegative strong solution of (3.1). Let $t \in (0,1)$, $\sigma \in (0,1]$, and R_1 and R_2 with $0 < R_1 < R_2$ be fixed and consider the sequences

$$r_n = R_1 + \sigma(R_2 - R_1)2^{-n}, \qquad s_n = t_1 - \sigma(t_1 - t_2)2^{-n}.$$

Set $B_n = B_{r_n}$ and $Q_n = B_n \times (s_n, t)$, $n = 0, 1, 2, \ldots$, and denote by ζ_n a nonnegative piecewise-smooth function in Q_n such that $\zeta_n = 1$ on Q_{n+1}, $\mathrm{supp}\zeta_n \subset Q_n$,

$$|\nabla\zeta_n| \leq 2^{n+1}/\sigma(R_2 - R_1), \quad \text{and} \quad 0 \leq \frac{\partial}{\partial t}\zeta_n \leq 2^{n+1}/\sigma(t_1 - t_2).$$

Moreover, for any $k > 0$, set $k_n = k(1 - 2^{-(n+1)})$.

Set $\varphi_n = \eta(|x|)e^{g_\lambda}U_{k_n}^q\zeta_n^p$ and $U_k = [(\eta(|x|)e^{g_\lambda})^{1/(p-1)}u - k]_+$, where q is a constant to be chosen later. Then we have

$$\begin{aligned}(3.11)\quad \iint_{Q_n} |\nabla u|^{p-2}\nabla u \cdot \nabla\varphi_n dxd\tau &\geq \frac{q}{2}\iint_{Q_n} (\eta(|x|)e^{g_\lambda})^{p/(p-1)}U_{k_n}^{q-1}|\nabla u|^p\zeta_n^p dxd\tau \\ &- C\iint_{Q_n} \{\eta(|x|)e^{g_\lambda}u^{p-1}U_{k_n}^q\zeta_n^p|\nabla g_\lambda|^p + U_0^pU_{k_n}^{q-1}[|\nabla g_\lambda|^p + |\nabla\zeta_n|^p]\}dxd\tau\end{aligned}$$

and

$$\begin{aligned}(3.12)\quad \iint_{Q_n} \frac{\partial}{\partial t}(u^{p-1})\varphi_n dxd\tau &\geq \int_{B_n} \Lambda_n(x,t)\zeta_n^p dx\Big|_{t_n}^{t} - \iint_{Q_n} \Lambda_n(x,\tau)\frac{\partial}{\partial t}\zeta_n^p dxd\tau \\ &- \iint_{Q_n} \eta(|x|)e^{g_\lambda}u^{p-1}U_{k_n}^q\zeta^p\frac{\partial}{\partial t}g_\lambda dxd\tau,\end{aligned}$$

where

$$\Lambda_n(x,t) = \int_{k_n}^{\eta(|x|)e^{g_\lambda}u^{p-1}} (s^{1/(p-1)} - k_n)_+^q ds.$$

We divide the proof into two cases—(A) $p \geq 2$ and (B) $1 < p < 2$—and proceed the proof of (3.9), respectively.

Case (A): ($p \geq 2$). Since the function g_λ has the relation

$$(3.13)\qquad -\frac{\partial}{\partial t}g_\lambda(x,t) \geq C\lambda^{1-p}t^{l-1}|\nabla g_\lambda(x,t)|^p,$$

by (3.11) and (3.12), there exists a constant T_λ^* dependent only on λ such that for any $t \in (0, T_\lambda^*)$,

$$k_n^{p-2}\sup_{\tau\in(0,t_{n+1})}\int_{B_{n+1}} U_{k_n}^{q+1}dx + \iint_{Q_{n+1}} U_{k_n}^{q-1}|\nabla U_{k_n}|^p dxd\tau \leq K\iint_{Q_n} U_{k_n}^{q-1}dxd\tau,$$

where

$$K \equiv \frac{\|U_0\|_{L^\infty(Q_0)}^p}{\sigma^p}M = \frac{\|U_0\|_{L^\infty(Q_0)}^p}{\sigma^p}(|\nabla g_\lambda(R_2,t)|^p + (R_2 - R_1)^{-p} + (t_1 - t_2)^{-1}).$$

Now we set $\omega_n = U_{k_n}^{(p+q-1)/p}$, $s = p(q+1)/(p+q-1)$, and $s' = p(q-1)/(p+q-1)$ and obtain the following inequality:

$$(3.14)\qquad k_n^{p-2}\sup_{\tau\in(0,t_{n+1})}\int_{B_{n+1}} \omega_n^s dx + \iint_{Q_n} |\nabla\omega_n|^p dxd\tau \leq K\iint_{Q_n} \omega_n^{s'} dxd\tau.$$

By (3.14), the Gagliardo–Nirenberg inequality yields

$$(3.15)\qquad \iint_{Q_{n+1}} \omega_{n+1}^{q'}dxd\tau \leq Ck_n^{p(2-p)/N}K^{(N+p)/N}\left(\iint_{Q_n} \omega_n^{s'} dxd\tau\right)^{(N+p)/N},$$

where $q' = p(1 + s/N)$.

On the other hand by $s' < q'$ and the Hölder inequality, we have

$$\iint_{Q_{n+1}} \omega_{n+1}^{s'} dx d\tau \le \left(\iint_{Q_{n+1}} \omega_{n+1}^{q'} dx d\tau \right)^{s'/q'} (\text{meas } A_{n+1})^{1-s'/q'},$$

where $A_n = \{(x,t) \in Q_n \,|\, e^{g_\lambda/(p-1)} u > k_n\}$. Since

$$\iint_{Q_n} \omega_n^{s'} dx d\tau \ge |k_{n+1} - k_n|^{q-1} \text{meas } A_{n+1} = \frac{k^{q-1}}{2^{(n+1)(q-1)}} \text{meas } A_{n+1},$$

we set

$$y_n = \iint_{Q_n} U_{k_n}^{q-1} dx d\tau$$

and have

$$\left(\frac{2^{(n+1)(p-1)}}{k^{p-1}} y_n \right)^{(s'/q')-1} y_{n+1} \le \left(\iint_{Q_{n+1}} \omega_{n+1}^{q'} dx d\tau \right)^{s'/q'}. \tag{3.16}$$

By (3.15) and (3.16), we obtain

$$y_{n+1} \le C b^n k^c K^{\frac{N+p}{N} \frac{s'}{q'}} y_n^{1+\frac{s'p}{q'N}}, \tag{3.17}$$

where

$$c = \frac{p(2-p)}{N} \frac{s'}{q'} - (q-1) \frac{q'-s'}{q'}.$$

By Lemma 5–6 in [10, p. 95], it follows that $\|U_0\|_{L^\infty(Q_\infty)} \le k$ provided

$$\iint_{Q_0} U_0^{q-1} dx d\tau \le C k^{N+p+q-1} K^{-(N+p)/p}. \tag{3.18}$$

Therefore, we have

$$\begin{aligned} \|U_0\|_{L^\infty(Q_\infty)} \le{}& C \|U_0\|_{L^\infty(Q_0^\sigma)}^{(N+p)/(N+p+q-1)} \\ &\times \left[\frac{M^{(N+p)/p}}{\sigma^{N+p}} \iint_{Q_0^\sigma} U_0^{q-1} dx d\tau \right]^{1/(N+p+q-1)}. \end{aligned} \tag{3.19}$$

Here $Q_0^\sigma \equiv Q_0 = B_{R_1+\sigma(R_2-R_1)} \times (t_1 - \sigma(t_1 - t_2), t)$ and $Q_\infty = B_{R_1} \times (t_1, t)$.

Next, we use the method of iteration with respect to σ. Set $Q_s = Q_0^\sigma$ with $\sigma = \sum_{i=1}^s 2^{-i-1}$, and define $X_s = \|\eta e^{g_\lambda} u^{p-1}\|_{L^\infty(Q_s)}$. Applying (3.19) to the pair of cylinders $Q_s \subset Q_{s+1}$, we obtain

$$X_s \le C X_{s+1}^{(N+p)/(N+p+q-1)} \left[2^{s(N+p)} M^{(N+p)/p} \iint_{Q_0} U_0^{q-1} dx d\tau \right]^{1/(N+p+q-1)}.$$

By the Young inequality, for any $\nu > 0$, there exists a constant $C(\nu)$ such that

$$X_s \le \nu X_{s+1} + C(\nu) (2^{\frac{N+p}{q-1}})^s \left[M^{(N+p)/p} \iint_{Q_0} U_0^{q-1} dx d\tau \right]^{1/(q-1)}.$$

Iteration of these inequalities yields

$$X_0 \leq \nu^s X_\infty + C\left[M^{(N+p)/p} \iint_{Q_0} U_0^{q-1} dx d\tau\right]^{1/(q-1)} \left(\sum_{i=1}^{s} (\nu 2^{\frac{N+p}{q-1}})^i\right).$$

Choosing $\nu = 2^{-\frac{N+p}{q-1}-1}$ and taking the limit as $s \to \infty$, we obtain

$$\|U_0\|_{L^\infty(Q_{R_1})} \leq CM^{(N+p)/p(q-1)} \left(\int_{t_1}^{t} \int_{B_{R_2}} U_0^{q-1} dx d\tau\right)^{1/(q-1)}. \tag{3.20}$$

Therefore, we set $q-1 = r(p-1)$ and obtain inequality (3.9) for Case (A).

Case (B): $(1 < p < 2)$. By (3.11)–(3.13), instead of (3.14), we have

$$\|U_0\|_{L^\infty(Q_0)}^{p-2} \sup_{\tau \in (0,t)} \int_{B_{n+1}} \omega_n^s dx + \iint_{Q_n} |\nabla \omega_n|^p dx d\tau \leq k^{p-2} \tilde{K} \iint_{Q_n} \omega^{s'} dx d\tau$$

where

$$\tilde{K} = \|U_0\|_{L^\infty(Q_0)}^2 M.$$

By an argument similar to that of Case (A), instead of (3.17), we obtain

$$y_{n+1} \leq Cd^n \|U_0\|_{L^\infty(Q_0)}^{\frac{p(2-p)s'}{Nq'}} k^{-(q-1)\frac{q'-s'}{q'}+(p-2)\frac{(N+p)s'}{Nq'}} \tilde{K}^{\frac{(N+p)s'}{Nq'}} y_n^{1+\frac{s'p}{q'N}}.$$

Then by Lemma 5–6 in [8], it follows that $\|U_0\|_{L^\infty(Q_\infty)} \leq k$ provided

$$\iint_{Q_0} U_0^{q-1} dx d\tau \leq C\|U_0\|_{L^\infty(Q_0)}^{p-2} k^{q-p+1+\frac{2(N+p)}{p}} \tilde{K}^{-\frac{N+p}{p}}.$$

Thus we obtain

$$\|U_0\|_{L^\infty(Q_\infty)}^{q-p+1+\frac{2(N+p)}{p}} \leq C\|U_0\|_{L^\infty(Q_0)}^{2-p+\frac{2(N+p)}{p}} \frac{M^{(N+p)/p}}{\sigma^{N+p}} \iint_{Q_0} U_0^{q-1} dx d\tau.$$

We can obtain (3.20) for Case (B) by the calculations similar to those of (3.19), and therefore we complete the proof of Lemma 3.4 for nonnegative strong solutions.

Let u be a solution of (3.1) constructed in Lemma 3.2. Then we can approximate the solution u by u_n, where u_n is a solution of (3.4). By Lemma 1–2 in [5], the functions $\max\{u_n, 0\}$ and $-\min\{u, 0\}$ are strong subsolutions of (3.4). Since the argument above holds for not only nonnegative strong solution of (3.1) but also subsolution, $\max\{u_n, 0\}$ and $-\min\{u_n, 0\}$ satisfy the inequality (3.9). Consequently, by the Hölder inequality of u_n and the Ascoli–Arzelá theorem, the function u satisfies the inequality (3.4). □

Using Lemmas 3.2–3.4, we shall prove Proposition 3.1.

Proof of Proposition 3.1. Let $\mu \in C_0^\infty(\mathbf{R}^N)$. By Lemma 3.2, there exists a continuous solution $u(x,t)$ in $\mathbf{S}_\infty$, which satisfies (3.3). To simplify the following arguments, we assume $T_\Lambda = \gamma_0^{p-1}\Lambda^{p-1} > 1$. Then for any $\lambda > \Lambda$ with $T_\lambda > 1$, by Lemma 3.2, we have

$$\sup_{0<t<1} \|u(\cdot,t)\|_{p,\lambda_1}^p + \int_0^1 \|\nabla u(\cdot,t)\|_{p,\lambda_1}^p dt \leq C\|\mu\|_{\frac{p}{p-1},\lambda}, \tag{3.21}$$

where $\lambda_1 = (p-1)^{-1}\gamma_0(T_\lambda - 1)^{-1/(p-1)} = -h_{\gamma_0, T_\lambda}(\overline{x}, 1)$ with $|\overline{x}| = 1$.

Set $T(\lambda) = (p-1)^{2(p-1)}\lambda^{1-p}/p^p$. Then by Lemma 3.3, we obtain the following inequality:

$$\sup_{0<t<T(\lambda)} \|u^{p-1}(\cdot,t)e^{h_{\gamma_{p-1},T(\lambda)}(\cdot,t)}\|_{L^p(\mathbf{R}^N)} \le C\|\mu e^{h_{\gamma_{p-1},T(\lambda)}(\cdot,0)}\|_{L^p(\mathbf{R}^N)} = C\|\mu\|_{p,\lambda}$$

for any $\lambda > \Lambda$. In particular, for any $\epsilon > 0$, we have

$$\sup_{0<t<T_\epsilon(\lambda)} \|u^{p-1}(\cdot,t)\|_{p,\lambda_\epsilon} \le C\|\mu\|_{p,\lambda}, \tag{3.22}$$

where $T_\epsilon(\lambda) = T(\lambda) - \epsilon$ and $\lambda_\epsilon = \gamma_{p-1}\epsilon^{1/(p-1)} = -h_{\gamma_{p-1},T(\lambda)}(\overline{x}, T_\epsilon(\lambda))$ with $|\overline{x}| = 1$.

For any fixed $t_0 \in (0, T_\epsilon(\lambda))$, we apply Lemma 3.4 to the function $u(x, t+t_0)$. Then by (3.22), there exists a constant $T^*_{\lambda_\epsilon} (< 1)$ independent of t_0 such that

$$\begin{aligned}\|e^{g_{\lambda_\epsilon}(\cdot,t)}u^{p-1}\|_{\infty,B_\rho}(t+t_0) &\le C(\rho,t)\left(\int_{t_0+t/4}^{t_0+t}\int_{B_{2\rho}} [e^{g_{\lambda_\epsilon}}u^{p-1}]^p dx d\tau\right)^{1/p} \\ &\le C(\rho,t)\left(\int_{t_0+t/4}^{t_0+t} \|u^{p-1}(\cdot,\tau)\|^p_{p,\lambda_\epsilon} d\tau\right)^{1/p}\end{aligned}$$

for any $t \in (0, \min\{t_0 + T^*_{\lambda_\epsilon}, T_\epsilon(\lambda)\})$. Taking $t_\epsilon = T^*_{\lambda_\epsilon}/2$, we have

$$\sup_{1<\tau<T_\epsilon(\lambda)} \|e^{g_{\lambda_\epsilon}(\cdot,t_\epsilon)}u^{p-1}(\cdot,\tau)\|_{L^\infty(B_\rho)} \le C(\rho)\|\mu\|_{p,\lambda}.$$

Since $C(\rho)$ has at most polynomial growth order in ρ, we get

$$\sup_{1<\tau<T_\epsilon(\lambda)} \|u^{p-1}\|_{p/(p-1),\lambda_\epsilon{}'} \le C\|\mu\|_{p,\lambda}$$

for any $\lambda_\epsilon{}' > -g_{\lambda_\epsilon}(\overline{x}, t_\epsilon) > 0$ with $|\overline{x}| = 1$. Therefore, by Lemma 3.2, there exists a constant $\lambda_2 > \lambda_\epsilon{}'$ such that

$$\sup_{1<\tau<T_\epsilon(\lambda)} \|u(\cdot,\tau)\|^p_{p,\lambda_2} + \int_1^{T_\epsilon(\lambda)} \|\nabla u\|^p_{p,\lambda_2} d\tau \le C(\epsilon)\|\mu\|_{p,\lambda}. \tag{3.23}$$

By (3.21) and (3.23), for any $\epsilon > 0$ and any $\lambda > \Lambda$, there exist constants C_1^ϵ and C_2^ϵ such that

$$\begin{aligned}\sup_{0<\tau<T_\epsilon(\lambda)} \|u(\cdot,\tau)\|_{p,C_1^\epsilon} &+ \int_0^{T_\epsilon(\lambda)} \|\nabla u(\cdot,\tau)\|^p_{p,C_1^\epsilon} d\tau \\ &\le C_2^\epsilon\,[\|\mu\|_{p,\lambda} + \|\mu\|_{p/(p-1),\lambda}] \le C_2^\epsilon\,\|\mu\|_{\infty,\Lambda}.\end{aligned} \tag{3.24}$$

By the arbitrariness $\lambda > \Lambda$, we get the inequality in Proposition 3.1.

For any $\mu \in L^\infty_\Lambda(\mathbf{R}^N)$, there exists a sequence $\{\mu_n\}_{n=1}^\infty \subset C_0^\infty(\mathbf{R}^N)$ such that $\lim_{n\to\infty} \|\mu - \mu_n\|_{\infty,\Lambda} = 0$. Let $u_n(x,t)$ be a solution of (3.1) for initial condition μ_n. Then by (3.24) and an argument similar to that of Lemma 3.2, there exists a solution of (3.1) for the initial condition μ in the strip $\mathbf{S}_{T_\epsilon(\lambda)}$. By the arbitrariness of ϵ and

λ and the results of [12] and [18], there exists a solution u in the strip $\mathbf{S}_{T(\Lambda)}$, and therefore we complete the proof of Proposition 3.1. □

4. $L^\infty_{\rm loc}(\mathbf{R}^N)$ estimate for Case (II). In this section, we give $L^\infty_{\rm loc}(\mathbf{R}^N)$ estimates of the solution of (1.1) for Case (II) and complete the proof of Theorem 1.3. By the arguments in §3, we have only to treat continuous nonnegative strong solution of (1.1).

By Lemma 3.4, we have the following lemma.

LEMMA 4.1. *Let u be a solution of* (3.1) *constructed in* §3. *Then for* $0 < T_0 < T^*_\lambda$, *there exists a constant* $C = C(N, p, \delta, T_0)$ *such that*

$$\|e^{g_\lambda} u^{p-1}\|_{L^\infty(B_\rho)}(t) \le C + Ct^{-N/p}\phi_\lambda(t) \tag{4.1}$$

for $0 < t < T_0$.

Proof. For any $\rho > 0$ and $t > 0$, let $Q_s = B_{\rho_s} \times (t_s, t)$, where $\rho_s = \sum_{i=1}^s (2^{-i-1})\rho$ and $t_s = (1 - \sum_i^s (2^{-i-1}))t$. Applying Lemma 3.4 to the pair of cylinders $Q_s \subset Q_{s+1}$, we have

$$\|\eta e^{g_\lambda} u^{p-1}\|_{L^\infty(Q_s)} \le C\frac{M^{(N+p)/p(1+\delta)}}{b^{s/(1+\delta)}}\left(\iint_{Q_{s+1}} [\eta e^{g_\lambda} u^{p-1}]^{1+\delta} dx d\tau\right)^{1/(1+\delta)}, \tag{4.2}$$

where $M = |\nabla g_\lambda|^p(2\rho, t) + \rho^{-p} + t^{-1}$ and $b = 2^{-N-p}$.

For any measurable set E in $\mathbf{R}^N$, let χ_E be the characteristic function of E. Set

$$\chi_1(x) = \chi_{\{e^{g_\lambda} u^{p-1} \le 1\}}(x), \qquad \chi_2(x) = \chi_{\{e^{g_\lambda} u^{p-1} \ge 1\}}(x). \tag{4.3}$$

By (2.8),

$$\begin{aligned}
&\iint_{Q_{s+1}} [\eta(|x|) e^{g_\lambda} u^{p-1}]^{1+\delta} dx d\tau \\
&\quad \le \eta^\delta(2\rho) \iint_{Q_{s+1}} \eta(|x|)[e^{g_\lambda} u^{p-1}]^{1+\delta} \chi_1(x) dx d\tau \\
&\qquad\qquad + \|\eta e^{g_\lambda} u^{p-1}\|^\delta_{L^\infty(Q_{s+1})} \iint_{Q_{s+1}} \eta(|x|) e^{g_\lambda} u^{p-1} dx d\tau \\
&\quad \le C[\eta^\delta(2\rho) + \|\eta e^{g_\lambda} u^{p-1}\|^\delta_{L^\infty(Q_{s+1})}] \int_0^t \phi_\lambda(\tau) d\tau.
\end{aligned}$$

By (4.2), we have

$$\begin{aligned}
\|\eta e^{g_\lambda} u^{p-1}\|_{\infty, Q_s} \le & C\frac{M^{(N+p)/p(1+\delta)}}{b^{s/(1+\delta)}} \\
& \times \left([\eta^\delta(2\rho) + \|\eta e^{g_\lambda} u^{p-1}\|^\delta_{L^\infty(Q_{s+1})}] \int_0^t \phi_\lambda(\tau) d\tau\right)^{1/(1+\delta)}
\end{aligned}$$

Then by the Young inequality, for any $\nu > 0$, there exists a constant $C = C(\nu, \delta)$ such that

$$\begin{aligned}
\|\eta e^{g_\lambda} u^{p-1}\|_{L^\infty(Q_s)} \le & \nu \|\eta e^{g_\lambda} u^{p-1}\|_{L^\infty(Q_{s+1})} \\
& + C(\nu, \delta) 2^{(N+p)s} \left[\eta(2\rho) + M^{(N+p)/p} \int_0^t \phi_\lambda(\tau) d\tau\right].
\end{aligned}$$

Iteration of these inequalities yields

$$\|\eta e^{g_\lambda}u^{p-1}\|_{L^\infty(Q_0)} \le \nu^s\|\eta e^{g_\lambda}u^{p-1}\|_{L^\infty(Q_s)} + C(\nu,\delta)[\eta(2\rho)+M^{(N+p)/p}t\phi_\lambda(t)]\sum_{i=1}^{s}(\nu 2^{N+p})^i.$$

Setting $\nu = 2^{-(N+p+1)}$ and taking the limit as $s\to\infty$, we obtain

$$\|\eta e^{g_\lambda}u^{p-1}\|_{L^\infty(B_\rho)} \le C(\delta)\eta(2\rho)+C(\delta)[|\nabla g_\lambda|^p+t^{-1}]^{(N+p)/p}t\phi_\lambda(t). \tag{4.4}$$

Therefore, by (4.4), we have inequality (4.1) for the case where $\rho\le 1$.

For any $0<T_0<1$, there exists a constant $C=C(T_0)$ such that $|\nabla g_\lambda|^p(x,t) \le C|x|^{p/(p-1)}$. For the case where $\rho\ge 1$, by (4.4) and the definition of η, we obtain

$$\|e^{g_\lambda}u^{p-1}\|_{L^\infty(B_\rho\setminus B_{\rho/2})} \le C+Ct^{-N/p}\phi_\lambda(t),$$

where $C=C(p,N,\delta,T_0)$. Therefore, we obtain inequality (4.1) for the case where $\rho\ge 1$, and thus the proof of Lemma 4.1 is complete. □

LEMMA 4.2. *Let $\zeta(x)$ be a piecewise-smooth cutoff function such that $\zeta\equiv 1$ on B_ρ, $|\nabla\zeta|\le 1/\rho$, and* supp $\zeta\subset B_{2\rho}$. *For any $\beta>0$, there exists a constant $T_0\in(0,1)$ dependent only on p such that*

$$\limsup_{\epsilon\to 0}\int_0^t\int_{B_{2\rho}}\tau^\beta\eta(|x|)e^{g_\lambda}\frac{|\nabla u|^p}{u+\epsilon}[e^{g_\lambda}(u+\epsilon)^{p-1}]^\delta\zeta^p(x)dxd\tau \le C\left[\int_0^t\tau^{\beta-1}\phi_\lambda(\tau)d\tau+\int_0^t\tau^{\sigma-1}\phi_\lambda^{1+\delta}(\tau)d\tau\right] \tag{4.5}$$

for $0<t<T_0$ and $\rho\ge 1$, where $C=C(N,p,\delta)$ and $\sigma=\beta-\delta N/p$.

Proof. Let

$$\varphi_\epsilon(x,t)=t^\beta\eta(|x|)e^{g_\lambda}[e^{g_\lambda}(u+\epsilon)^{p-1}]^\delta\zeta^p(x).$$

Then we have

$$\liminf_{\epsilon\to 0}\int_0^t\int_{B_{2\rho}}\frac{\partial}{\partial t}(u^{p-1})\varphi_\epsilon dxd\tau \ge -\frac{\beta}{1+\delta}\int_0^t\int_{B_{2\rho}}\tau^{\beta-1}\eta[e^{g_\lambda}u^{p-1}]^{1+\delta}dxd\tau - \frac{1}{1+\delta}\int_0^t\int_{B_{2\rho}}\tau^\beta\eta[e^{g_\lambda}u^{p-1}]^{1+\delta}\frac{\partial}{\partial t}g_\lambda dxd\tau$$

and

$$\int_0^t\int_{B_{2\rho}}|\nabla u|^{p-2}\nabla u\cdot\nabla\varphi_\epsilon dxd\tau \ge \frac{(p-1)\delta}{2}\int_0^t\int_{B_{2\rho}}\tau^\beta e^{g_\lambda}\frac{|\nabla u|^p}{u+\epsilon}[e^{g_\lambda}(u+\epsilon)^{p-1}]^\delta dxd\tau - C(p,\delta)\int_0^t\int_{B_{2\rho}}\tau^\beta\eta[e^{g_\lambda}(u+\epsilon)^{p-1}]^{1+\delta}|\nabla g_\lambda|^p dxd\tau - \frac{C(p,\delta)}{\rho^p}\int_0^t\int_{B_{2\rho}}\tau^\beta\eta[e^{g_\lambda}(u+\epsilon)^{p-1}]^{1+\delta}dxd\tau.$$

Taking a sufficiently small T_0, by (3.13), we obtain

$$\limsup_{\epsilon\to 0}\int_0^t\int_{B_{2\rho}}\tau^\beta\eta(|x|)e^{g_\lambda}\frac{|\nabla u|^p}{u+\epsilon}[e^{g_\lambda}(u+\epsilon)^{p-1}]^\delta\zeta^p dxd\tau \leq C\int_0^t\int_{B_{2\rho}}\tau^{\beta-1}\eta(|x|)(e^{g_\lambda}u^{p-1})^{\delta+1}dxd\tau. \tag{4.6}$$

Here we used the relation $t^\beta/\rho^p \leq t^{\beta-1}$ for $0<t<T_0$. Thus we have

$$\int_0^t\int_{B_{2\rho}}\tau^{\beta-1}\eta(|x|)(e^{g_\lambda}u^{p-1})^{\delta+1}dxd\tau \leq \int_0^t\tau^{\beta-1}\phi_\lambda(\tau)d\tau + \int_0^t\int_{B_{2\rho}}\tau^{\beta-1}\eta(|x|)e^{g_\lambda}u^{p-1}\|e^{g_\lambda}u^{p-1}\|^\delta_{L^\infty(B_{2\rho})}(\tau)\chi_2 dxd\tau. \tag{4.7}$$

Furthermore, by Lemma 4.1, we have

$$\begin{aligned}\int_0^t\int_{B_{2\rho}}&\tau^{\beta-1}\eta(|x|)e^{g_\lambda}u^{p-1}\chi_2\|e^{g_\lambda}u^{p-1}\|^\delta_{L^\infty(B_{2\rho})}(\tau)dxd\tau\\ &\leq\int_0^t\int_{B_{2\rho}}\tau^{\beta-1}\eta(|x|)e^{g_\lambda}u^{p-1}\chi_2[1+\tau^{-\delta N/p}\phi_\lambda^\delta(\tau)]dxd\tau\\ &\leq C\int_0^t\tau^{\beta-1}\phi_\lambda(\tau)d\tau+C\int_0^t\tau^{\beta-1-\delta N/p}\phi_\lambda^{1+\delta}(\tau)d\tau\end{aligned} \tag{4.8}$$

for any $0<t<T_0$. Therefore, by (4.6)–(4.8), we obtain inequality (4.5). □

PROPOSITION 4.3. *For a sufficiently small constant $\delta>0$, there exists a constant $T_0=T_0(p,\delta,l,\|\mu\|_{1,\lambda})$ such that*

$$\phi_\lambda(t)\leq C\phi_\lambda(0) \tag{4.9}$$

for $0<t<T_0$, where $C=C(p,\delta)$. Furthermore,

$$\|e^{g_\lambda(\cdot,t)}u^{p-1}\|_{L^\infty(B_\rho)}\leq C(1+t^{-N/p})\phi_\lambda(0) \tag{4.10}$$

for any $\rho>0$ and $0<t<T_0$.

Proof. Set

$$\varphi_\epsilon(x,t)=\eta(|x|)e^{g_\lambda}F'(e^{g_\lambda}(u+\epsilon)^{p-1})\zeta^p(x),$$

where ζ is a function given in Lemma 4.2. Then we have

$$\lim_{\epsilon\to 0}\int_0^t\int_{B_{2\rho}}\frac{\partial}{\partial t}(u^{p-1})\varphi_\epsilon dxd\tau=\int_{B_{2\rho}}\eta(|x|)F(e^{g_\lambda}u^{p-1})\zeta^p dx\Big|_{\tau=0}^{\tau=t} - \int_0^t\int_{B_{2\rho}}\eta(|x|)[(e^{g_\lambda}u^{p-1})^{1+\delta}\chi_1+e^{g_\lambda}u^{p-1}\chi_2]\zeta^p\frac{\partial}{\partial t}g_\lambda dxd\tau$$

and

$$\begin{aligned}\int_0^t\int_{B_{2\rho}}&|\nabla u|^{p-2}\nabla u\cdot\nabla\varphi_\epsilon dxd\tau\\ &\geq\delta(p-1)\int_0^t\int_{B_{2\rho}}\eta e^{g_\lambda}\frac{|\nabla u|^p}{u+\epsilon}[e^{g_\lambda}(u+\epsilon)^{p-1}]^\delta\zeta^p\chi_1 dxd\tau\\ &-C(p,\delta)\int_0^t\int_{B_{2\rho}}\eta e^{g_\lambda}|\nabla u|^{p-1}(|\nabla g_\lambda|\zeta^p+\zeta^{p-1}|\nabla\zeta|)[e^{g_\lambda}(u+\epsilon)^{p-1}]^\delta\chi_1 dxd\tau\\ &-\int_0^t\int_{B_{2\rho}}\eta e^{g_\lambda}|\nabla u|^{p-1}(|\nabla g_\lambda|\zeta^p+\zeta^{p-1}|\nabla\zeta|)\chi_2 dxd\tau,\end{aligned}$$

where χ_i, $i=1,2$, is given in (4.3). The Young inequality yields

$$\begin{aligned}\int_0^t\int_{B_{2\rho}} &\eta e^{g_\lambda}|\nabla u|^{p-1}|\nabla g_\lambda|\zeta^p[e^{g_\lambda}(u+\epsilon)^{p-1}]^\delta\chi_1 dxd\tau\\ &\le \int_0^t\int_{B_{2\rho}}\tau^{1/2(p-1)}\eta e^{g_\lambda}\frac{|\nabla u|^p}{u+\epsilon}[e^{g_\lambda}(u+\epsilon)^{p-1}]^\delta\zeta^p\chi_1 dxd\tau\\ &\qquad+\int_0^t\int_{B_{2\rho}}\eta[e^{g_\lambda}(u+\epsilon)^{p-1}]^{1+\delta}\tau^{-1/2}|\nabla g_\lambda|^p\zeta^p\chi_1 dxd\tau\end{aligned}$$

and

$$\begin{aligned}\int_0^t\int_{B_{2\rho}}\eta e^{g_\lambda}|\nabla u|^{p-1}|\nabla g_\lambda|\zeta^p\chi_2 dxd\tau &\le \int_0^t\int_{B_{2\rho}}\tau^{1/2(p-1)}\eta e^{g_\lambda}\frac{|\nabla u|^p}{u+\epsilon}\zeta^p\chi_2 dxd\tau\\ &\qquad+\int_0^t\int_{B_{2\rho}}\eta e^{g_\lambda}(u+\epsilon)^{p-1}\tau^{-1/2}|\nabla g_\lambda|^p\zeta^p\chi_2 dxd\tau.\end{aligned}$$

Therefore, by (3.13) and the Young inequality, there exists a constant $C=C(N,p,\delta,l)$ such that

(4.11)
$$\begin{aligned}\int_{B_\rho}\eta(|x|)F(e^{g_\lambda}u^{p-1})dx\Big|_{\tau=t} &\le \int_{B_{2\rho}}\eta(|x|)F(e^{g_\lambda}u^{p-1})dx\Big|_{\tau=0}\\ &+\limsup_{\epsilon\to0}\int_0^t\int_{B_{2\rho}}\tau^{1/2(p-1)}\eta(|x|)e^{g_\lambda}\frac{|\nabla u|^p}{u+\epsilon}[e^{g_\lambda}(u+\epsilon)^{p-1}]^\delta\zeta^p dxd\tau\\ &+\frac{C}{\rho^p}\int_0^t\int_{B_{2\rho}}[\eta(|x|)(e^{g_\lambda}u^{p-1})^{\delta+1}\chi_1+\tau^{-1/2}\eta(|x|)e^{g_\lambda}u^{p-1}\chi_2]dxd\tau.\end{aligned}$$

Taking the limit as $\rho\to\infty$, by Lemma 4.2, we have

$$\begin{aligned}\phi_\lambda(t)&\le\phi_\lambda(0)+C\int_0^t[(\tau^{-1/2}+\tau^{\beta-1})\phi_\lambda(\tau)+\tau^{\sigma-1}\phi_\lambda^{1+\delta}(\tau)]d\tau\\ &\le\phi_\lambda(0)+C(t^{1/2}+t^\beta)\phi_\lambda(t)+C\int_0^t\tau^{\sigma-1}\phi_\lambda^{1+\delta}(\tau)d\tau,\end{aligned}$$

where σ is a constant given in Lemma 4.2. Here we take a sufficiently small $\delta>0$ such that $\sigma>0$ and fix δ. Taking a sufficiently small t such that $C(t^{1/2}+t^\beta)\le1/2$, we have

$$\phi_\lambda(t)\le2\phi_\lambda(0)+C\int_0^t\tau^{\sigma-1}\phi_\lambda^{1+\delta}(\tau)d\tau. \tag{4.12}$$

It follows by (4.12) that $\phi_\lambda(t)$ is maximized by the solution of

$$H'(t)=Ct^{\sigma-1}H^{1+\delta}(t),\qquad H(0)=2\phi_\lambda(0),$$

and so we have

$$\phi_\lambda(t)\le\frac{H(t)}{2}=\left[1-\frac{C\delta2^\delta}{\sigma}t^\sigma\phi_\lambda^\delta(0)\right]^{-1/\delta}\phi_\lambda(0)$$

provided the bracket is positive for any $0 < t < T_0$. In view of $\phi_\lambda(0) \le \|\mu\|_{1,\lambda}$, there exist constants $T_0 > 0$ and $C > 0$ such that $\phi_\lambda(t) \le C\phi_\lambda(0)$ for all $t \in (0, T_0)$, and so we obtain inequality (4.9). Furthermore, by (4.1) and (4.9), we have inequality (4.10), and the proof of Proposition 4.3 is complete. □

PROPOSITION 4.4. *By Proposition* 4.3, *it holds that*

$$\int_0^t \int_{B_\rho} \eta(|x|)e^{g_\lambda}|\nabla u|^{p-1}dxd\tau \le Ct^\sigma(\eta(\rho)\rho^N + \phi_\lambda(0) + \phi_\lambda^\delta(0)) \tag{4.13}$$

for any $\rho \ge 1$ *and* $0 < t < T_0 \le 1$.

Proof. By the Young inequality,

$$\int_0^t \int_{B_\rho} \eta(|x|)e^{g_\lambda}|\nabla u|^{p-1}dxd\tau \le \int_0^t \int_{B_\rho} \tau^\beta \eta e^{g_\lambda}\frac{|\nabla u|^p}{u+\epsilon}[e^{g_\lambda}(u+\epsilon)^{p-1}]^\delta dxd\tau$$
$$+ \int_0^t \int_{B_\rho} \tau^{-\beta(p-1)}\eta[e^{g_\lambda}(u+\epsilon)^{p-1}]^{1-\delta(p-1)}dxd\tau \equiv I_1(\epsilon) + I_2(\epsilon),$$

where $\beta = 1/2(p-1)$. By Lemma 4.2 and Proposition 4.3,

$$\limsup_{\epsilon\to 0} I_1(\epsilon) \le Ct^\sigma[\phi_\lambda(0) + \phi_\lambda^\delta(0)].$$

For the second term $I_2(\epsilon)$, we have

$$\lim_{\epsilon\to 0} I_2(\epsilon) \le \int_0^t \int_{B_\rho} \tau^{-\beta(p-1)}\eta(e^{g_\lambda}u^{p-1})^{1-\delta(p-1)}[\chi_1 + \chi_2]dxd\tau$$
$$\le Ct^{1/2}(\eta(\rho)\rho^N + \phi_\lambda(0)),$$

and so we have (4.13). □

We now can prove Theorem 1.3.

Proof of Theorem 1.3. By the definition of ϕ_λ,

$$\phi_\lambda(0) \le \int_{\mathbf{R}^N} \eta(|x|)\exp(-\lambda|x|^{p/(p-1)})\mu(x)dx.$$

Therefore, by Proposition 4.3 and $\lambda > \Lambda$, we obtain inequalities (2.12) and (2.13) for initial data $\mu \in C_0^\infty(\mathbf{R}^N)$.

Let μ_n be a function in $C_0^\infty(\mathbf{R}^N)$ such that

$$\lim_{n\to\infty} \int_{\mathbf{R}^N} \mu_n \exp(-\lambda|x|^{p/(p-1)})dx = \int_{\mathbf{R}^N} \exp(-\lambda|x|^{p/(p-1)})d\mu(x).$$

By Lemma 3.2, there exists a solution u_n of (3.1) for $u(x,0) = \mu_n(x)$ in $\mathbf{S}_\infty \equiv \mathbf{R}^N \times (0,\infty)$. By the argument of §3 and (4.10), we have

$$\int_\tau^{T_0} \int_{B_\rho} |\nabla u_n|^p dxdt \le C(\rho,\tau) \tag{4.14}$$

for any $0 < \tau < T_0$ and $\rho \ge 1$, where $C(\rho,\tau)$ is a constant independent of n. Consequently, by (4.14), the Minty lemma, and the result of [18], there exists a continuous function u satisfying $|\nabla u| \in L^p_{\text{loc}}(0, T_0 : L^p_{\text{loc}}(\mathbf{R}^N))$ such that

$$\int_\tau^{T_0} \int_{B_\rho} |\nabla u_n|^{p-2}\nabla u_n \cdot \nabla\varphi dxdt \to \int_\tau^{T_0} \int_{B_\rho} |\nabla u|^{p-2}\nabla u \cdot \nabla\varphi dxdt, \tag{4.15}$$

as $n \to \infty$, for all $0 < \tau < T_0$ and all $\varphi \in C^\infty(B_\rho \times (0, T_0)$ with $\operatorname{supp}(\varphi(\cdot, t)) \subset B_\rho$, $0 < t < T_0$. Furthermore, by Proposition 4.4, there exists a Radon measure $V = (V_1, V_2, \dots, V_N)$ such that

$$\lim_{n\to\infty} \int_0^{T_0} \int_{B_\rho} |\nabla u_n|^{p-2} \nabla u_n \cdot \nabla \varphi dx dt = \sum_{i=1}^N \int_0^{T_0} \int_{B_\rho} \frac{\partial}{\partial x_i} \varphi dV_i \tag{4.16}$$

and $\lim_{t\to 0} V_i(B_\rho \times (0,t)) = 0$, $i = 1, 2, \dots, N$. By (4.15) and (4.16), we see $U = |\nabla u|^{p-2}\nabla u$, and so we have $|\nabla u|^{p-1} \in L^1(0, T_0 : L^1_{\mathrm{loc}}\mathbf{R}^N)$. Consequently, the function u is a solution of (1.1) and (1.2) in $\mathbf{S}_{T_0}$.

On the other hand, by (4.10), we have

$$\| \exp(-\lambda(t)|x|^{p/(p-1)}) u(x,t) \|_{L^\infty(\mathbf{R}^N)} < \infty$$

for $0 < t < T_0$, where $\lambda(t) = \lambda(1 + t^{1/2}) = -g_\lambda(\overline{x}, t) =$ with $|\overline{x}| = 1$. Therefore, by Proposition 3.1, we can see that there exists a solution of (1.1) and (1.2) in $\mathbf{S}_{T(\Lambda)}$, and the proof of Theorem 1.3 is completed. □

5. $L^\infty_{\mathrm{loc}}(\mathbf{R}^N)$ estimate for Case (I). In this section, for Case (I), we give $L^\infty(\mathbf{R}^N)$ estimates of the solutions of (1.1). We set $v = |u|^{\beta-1}u$ and $d = m(p-1)-1 > 0$ and consider the following problem:

$$\begin{cases} \dfrac{\partial}{\partial t} v = \operatorname{div}(|\nabla v^m|^{p-2} \nabla v^m) & \text{in } \mathbf{R}^N \times (0, \infty), \\ v = \mu(x) \in C_0^\infty(\mathbf{R}^N) & \text{in } \mathbf{R}^N, \end{cases} \tag{5.1}$$

where $m = 1/\beta$. By arguments similar to those of §3, we have only to consider nonnegative solutions, and so we assume that $v \geq 0$ and $\mu \geq 0$.

LEMMA 5.1. *There exists a weak solution v of* (5.1) *in $\mathbf{S}_\infty$. The solution v satisfies the following inequalities:*

$$\|v\|_{L^\infty(\mathbf{R}^N)} \leq \|\mu\|_{L^\infty(\mathbf{R}^N)} \tag{5.2}$$

and

$$\int_\tau^t \int_{B_\rho} |\nabla v^m|^p dx d\tau \leq C[\rho^N \|v(\cdot, \tau)\|^{m+1}_{L^\infty(B_{2\rho})} + (t-\tau)\rho^{N-p} \|v\|^{mp}_{L^\infty(B_{2\rho} \times (\tau, t))}] \tag{5.3}$$

for any $0 < \tau < t$ and $\rho > 0$, where $C = C(p, m, N)$.

Proof. By the results of [16] and [18], for any sufficiently large n, there exists a weak solution v_n of the problem

$$\begin{cases} \dfrac{\partial}{\partial t} v = \operatorname{div}(|\nabla v^m|^{p-2} \nabla v^m) & \text{in } B_n \times (0, \infty), \\ v(x,t) = 0 & \text{on } \partial B_n \times (0, \infty), \\ v = \mu(x) & \text{in } B_n \end{cases} \tag{5.4}$$

such that $\|v_n\|_{L^\infty(B_n)} \leq \|\mu\|_{L^\infty(B_n)}$. For any $\rho > 0$ with $2\rho < n$, let ζ be a nonnegative piecewise-smooth cutoff function such that $\zeta \equiv 1$ on B_ρ, $\operatorname{supp}(\zeta) \subset B_{2\rho}$, and $|\nabla \zeta| \leq 1/\rho$. Then we multiply (5.4) by $\varphi(x,t) = v_n^m \zeta^p$, integrate it over $B_\rho \times (\tau, t)$, and obtain

$$\int_\tau^t \int_{B_\rho} |\nabla v_n^m|^p dx d\tau \leq C[\rho^N \|v(\cdot, \tau)\|^{m+1}_{L^\infty(B_{2\rho})} + (t-\tau)\rho^{N-p} \|v_n\|^{mp}_{L^\infty(B_{2\rho} \times (\tau, t))}].$$

Taking the limit as $n \to \infty$, we can prove Lemma 5.1 by the same argument as in Lemma 3.2. □

By an argument similar to that of Lemma 3.4, we obtain the following lemma, which is an extension of Lemma 3–1 in [6].

LEMMA 5–2. *Let $r > 0$ and v be a solution of* (5.1) *and*

$$\phi(t) = \sup_{\tau\in(0,t)} \tau^{N/\kappa} \sup_{\rho\geq r} \frac{\|v(\cdot,\tau)\|_{\infty,B_\rho}}{\rho^{p/d}}. \tag{5.5}$$

Then the solution v satisfies the following inequality:

$$\|v(\cdot,t)\|_{\infty,B_\rho} \leq C[K(t)]^{(N+p)/\kappa^*} \left(\int_{t/4}^{t} \int_{B_{2\rho}} v^{mp} dx d\tau\right)^{p/\kappa^*} \tag{5.6}$$

for all $\rho \geq r$, where

$$K(t) = t^{-Nd/\kappa} \phi^d(t) + t^{-1}. \tag{5.7}$$

Proof. We set $R_1 = \rho$ and $R_2 = 2\rho$ and define B_n, Q_n, and ζ_n as in Lemma 3.4. Furthermore we set

$$\varphi_n(x,t) = \begin{cases} (u-k_n)_+^{p-1}\zeta_n^p & \text{if } p \geq 2, \\ (u-k_n)_+\zeta_n^p & \text{if } 1 < p < 2 \end{cases}$$

and multiply φ_n by (5.1) to obtain that if $p \geq 2$,

$$\begin{aligned} &\int_{B_n} \Lambda_{n,1}(x,\tau)\zeta_n^p dx \Big|_{\tau=t_n} + \iint_{Q_n} (u-k_n)_+^{p-2} |\nabla u|^p \zeta^p dx d\tau \\ &\leq C \iint_{Q_n} (u-k_n)_+^{2(p-1)} |\nabla\zeta|^p dx d\tau + C \iint_{Q_n} \Lambda_{n,1}(x,\tau) \frac{\partial}{\partial t}\zeta_n^p dx d\tau \end{aligned} \tag{5.8}$$

and if $1 < p < 2$,

$$\begin{aligned} &\int_{B_n} \Lambda_{n,2}(x,\tau)\zeta_n^p dx \Big|_{\tau=t_n} + \iint_{Q_n} |\nabla(u-k_n)_+|^p \zeta^p dx d\tau \\ &\leq C \iint_{Q_n} (u-k_n)_+^{p} |\nabla\zeta|^p dx d\tau + C \iint_{Q_n} \Lambda_{n,2}(x,\tau) \frac{\partial}{\partial t}\zeta_n^p dx d\tau, \end{aligned} \tag{5.9}$$

where

$$\Lambda_{n,1}(x,t) = \int_{k_n^\beta}^{u^\beta(x,t)} (s^{1/\beta}-k_n)_+^{p-1} ds \quad \text{and} \quad \Lambda_{n,2}(x,t) = \int_{k_n^\beta}^{u^\beta(x,t)} (s^{1/\beta}-k_n)_+ ds.$$

Here we divide the proof of (5.6) into three cases,

(A) $m \geq 1$, $p \geq 2$, (B) $0 < m < 1$, $p \geq 2$, (C) $m \geq 1$, $1 < p < 2$,

and treat each case respectively.

Case (A): $m \ge 1$, $p \ge 2$. By (5.8) and $m \ge 1$,

$$\|u\|_{\infty,Q_n}^{-(m-1)/m} \sup_{t_n \le \tau \le t} \int_{B_n} (u-k_n)_+^p \zeta_n^p dx + \iint_{Q_n} |D[(u-k_n)_+^{2(p-1)/p} \zeta_n]|^p dxd\tau$$
$$\le C2^{np}\left[\|u\|_{\infty,Q_n}^{p-2}(\sigma\rho)^{-p} + k^{-(m-1)/m}(\sigma t)^{-1}\right] \iint_{Q_n} (u-k_n)_+^p dxd\tau.$$

Set $\omega_n = (u-k_n)_+^{2(p-1)/p}$ and $s = p^2/2(p-1)$, and we have

$$\|u\|_{\infty,Q_n}^{-(m-1)/m} \sup_{t_n \le \tau \le t} \int_{B_n} \omega_n^s \zeta_n^p dx + \iint_{Q_n} |D[\omega_n \zeta_n]|^p dxd\tau$$
$$\le C2^{np}\left[\|u\|_{\infty,Q_n}^{p-2}(\sigma\rho)^{-p} + k^{-(m-1)/m}(\sigma t)^{-1}\right] \iint_{Q_n} \omega_n^s dxd\tau.$$

Using the Gagliardo–Nirenberg inequality and the Hölder inequality as in Lemma 3.4, we have

$$\iint_{Q_{n+1}} (u-k_{n+1})_+^p dxd\tau \le Cb^n \left[\|u\|_{\infty,Q_n}^{p-2}(\sigma\rho)^{-p} + k^{-(m-1)/m}(\sigma t)^{-1}\right]^{(N+p)s/Nq}$$
$$\times \|u\|_{\infty,Q_0}^{(m-1)ps/mq} k^{-p(1-s/q)} \left(\iint_{Q_n} (u-k_n)_+^p dxd\tau\right)^{1+ps/Nq},$$

where $q = p(1+s/N)$ and b is a constant independent of n. By Lemma 5–6 in [8], we have the following inequality:

$$\|u\|_{\infty,Q_\infty}^{[N(p-2)+p^2]/p} \le C\|u\|_{\infty,Q_0}^{(m-1)/m}$$
$$\times \left[\|u\|_{\infty,Q_0}^{p-2}(\sigma\rho)^{-p} + \|u\|_{\infty,Q_\infty}^{-(m-1)/m}(\sigma t)^{-1}\right]^{1+N/p} \iint_{Q_0} u^p dxd\tau$$
$$\le C\|u\|_{\infty,Q_0}^{(m-1)/m} \|u\|_{\infty,Q_\infty}^{-(m-1)(1+N/p)/m}$$
$$\times \sigma^{-(p+N)} \left[\|u\|_{\infty,Q_0}^{p-2+(m-1)/m} \rho^{-p} + t^{-1}\right]^{1+N/p} \iint_{Q_0} u^p dxd\tau.$$

By $u = v^m$,

$$(5.10)\quad \|v\|_{\infty,Q_\infty}^{p(m-1)+\kappa_{mp}} \le C\|v\|_{\infty,Q_0}^{p(m-1)} \sigma^{-N-p} \left[\frac{\|v\|_{\infty,Q_0}^d}{\rho^p} + t^{-1}\right]^{N+p} \left(\iint_{Q_0} v^{mp} dxd\tau\right)^p,$$

where $Q_0 = B_{(1+\sigma)\rho} \times ((1-\sigma)t/2, t)$ and $Q_\infty = B_\rho \times (t/2, t)$.

Setting Q_s as in Lemma 3.4, and applying (5.10) to the pair of cylinders $Q_s \subset Q_{s+1}$, we have

$$X_s^{p(m-1)+\kappa_{mp}} \le CX_{s+1}^{p(m-1)} 2^{s(N+p)} \left[\frac{\|v\|_{\infty,Q_0}^d}{\rho^p} + t^{-1}\right]^{N+p} \left(\iint_{Q_0} v^{mp} dxd\tau\right)^p.$$

The Young inequality yields

$$X_s \le \nu X_{s+1} + C(\nu)(2^{N+p})^s \left[\frac{\|v\|_{\infty,Q_0}^d}{\rho^p} + t^{-1}\right]^{(N+p)/\kappa^*} \left(\iint_{Q_0} v^{mp} dxd\tau\right)^{p/\kappa^*},$$

where $\nu = 2^{-(N+p)-1}$. Iterating these inequalities and taking the limit as $s \to \infty$, we obtain

$$\|v\|_{\infty,Q_\rho} \le C\left[\frac{\|v\|^d_{\infty,Q_{2\rho}}}{\rho^p} + t^{-1}\right]^{(N+p)/\kappa_{mp}} \left(\int_{t/4}^{t}\int_{B_{2\rho}} v^{mp}\,dx\,d\tau\right)^{p/\kappa_{mp}}. \tag{5.11}$$

Therefore, by (5.5) and (5.11), we obtain inequality (5.6).

Case (B): $0 < m < 1$, $p \ge 2$. By (5.8) and $0 < m < 1$,

$$\begin{aligned} k^{-(m-1)/m} \sup_{t_n \le \tau \le t} \int_{B_n} (u-k_n)_+^p \zeta_n^p\,dx + \iint_{Q_n} |D[(u-k_n)_+^{2(p-1)/p}\zeta_n]|^p\,dx\,d\tau \\ \le C2^{np}\left[\|u\|^{p-2}_{\infty,Q_n}(\sigma\rho)^{-p} + \|v\|^{-(m-1)/m}_{\infty,Q_n}(\sigma t)^{-1}\right] \iint_{Q_n} (u-k_n)_+^p\,dx\,d\tau. \end{aligned}$$

Applying the Gagliardo–Nirenberg inequality, the Young inequality, and Lemma 5–6 in [10] as in Case (A), instead of (5.10), we have

$$\begin{aligned} \|v\|^{(1-m)(N+p)+\kappa_{mp}}_{\infty,Q_\infty} &\le C\|v\|^{(1-m)(N+p)}_{\infty,Q_0}\sigma^{-(N+p)p} \\ &\quad\times\left[\frac{\|v\|^d_{\infty,Q_0}}{\rho^p} + t^{-1}\right]^{N+p}\left(\iint_{Q_0} v^{mp}\,dx\,d\tau\right)^p. \end{aligned} \tag{5.12}$$

By the Young inequality and the argument of iteration, we can obtain (5.6) as in Case (A).

Case (C): $m \ge 1$, $1 < p < 2$. By (5.9) and $m \ge 1$,

$$\begin{aligned} \|u\|^{-(m-1)/m}_{\infty,Q_n} \sup_{t_n \le \tau \le t} \int_{B_n} (u-k_n)_+^2\,dx + \iint_{Q_n} |D(u-k_n)_+|^p\,dx\,d\tau \\ \le C2^{np}\left[\frac{1}{(\sigma\rho)^p} + \frac{k^{-(m-1)/m}}{\sigma t}\|u\|^{2-p}_{\infty,Q_0}\right]\iint_{Q_n}(u-k_n)_+^p\,dx\,d\tau, \end{aligned}$$

and by $1 < p < 2$,

$$\begin{aligned} \|u\|^{-(m-1)/m}_{\infty,Q_n}\left(\frac{k}{2^n}\right)^{2-p} \sup_{t_n \le \tau \le t} \int_{B_n} (u-k_{n+1})_+^p\,dx + \iint_{Q_n} |D(u-k_{n+1})_+|^p\,dx\,d\tau \\ \le C2^{np}\left[\frac{1}{(\sigma\rho)^p} + \frac{k^{-(m-1)/m}}{\sigma t}\|u\|^{2-p}_{\infty,Q_0}\right]\iint_{Q_n}(u-k_n)_+^p\,dx\,d\tau. \end{aligned}$$

By the Gagliardo–Nirenberg inequality and the Hölder inequality, we have

$$\begin{aligned} \iint_{Q_{n+1}} (u-k_{n+1})_+^p\,dx\,d\tau &\le Cb^n \|u\|^{(m-1)(1-p/q)/m}_{\infty,Q_0} k^{-p(1-p/q)+(p-2)(1-p/q)} \\ &\quad\times\left[(\sigma\rho)^{-p} + \frac{k^{-(m-1)/m}}{\sigma t}\|u\|^{2-p}_{\infty,Q_0}\right]\left(\iint_{Q_n}(u-k_n)_+^p\,dx\,d\tau\right)^{1+(q-p)/p}, \end{aligned}$$

where $q = p(1+p/N)$ and b is a constant independent of n.

By $m \geq 1$,

$$
\begin{aligned}
(\sigma\rho)^{-p} + \frac{k^{-(m-1)/m}}{\sigma t}\|u\|_{\infty,Q_0}^{2-p}
&\leq \|u\|_{\infty,Q_\infty}^{-(m-1)/m}\|u\|_{\infty,Q_0}^{2-p}\sigma^{-p}\left[\frac{\|u\|_{\infty,Q_0}^{(m-1)/m}\|u\|_{\infty,Q_0}^{2-p}\sigma^{-p}}{\rho^p} + t^{-1}\right] \\
&\leq \|v\|_{\infty,Q_\infty}^{-(m-1)}\|v\|_{\infty,Q_0}^{m(2-p)}\sigma^{-p}\left[\frac{\|v\|_{\infty,Q_0}^{d}}{\rho^p} + t^{-1}\right].
\end{aligned}
$$

Therefore, by Lemma 5–6 in [8], we obtain

$$
\begin{aligned}
\|v\|_{\infty,Q_\infty}^{2m+(m-1)(1+N/p)} \leq C\|v\|_{\infty,Q_0}^{m-1+m(2-p)(1+N/p)} \\
\times\left[\frac{\|v\|_{\infty,Q_0}^{d}}{\rho^p} + t^{-1}\right]^{1+N/p}\sigma^{-(p+N)}\iint_{Q_0} v^p dx d\tau.
\end{aligned}
$$

By the same argument as in Case (A), we obtain inequality (5.5) and complete the proof of Lemma 5.2. □

Let $\psi(t)$ be a function defined by

$$
\psi(t) = \sup_{\tau\in(0,t)} |||v(\cdot,\tau)|||_r. \tag{5.13}
$$

Then we can prove the following lemma.

LEMMA 5.3. *Let $\rho \geq r > 0$ and $\zeta(x)$ be a nonnegative piecewise-smooth cutoff function in $B_{2\rho}$ such that $\zeta \equiv 1$ on B_ρ, $\operatorname{supp}\zeta \subset B_{2\rho}$, and $|\nabla\zeta| < \rho^{-1}$. Then for any $t > 0$,*

$$
\begin{aligned}
&\int_0^t \int_{B_{2\rho}} |Dv^m|^{p-1}\zeta^{p-1} dx d\tau \\
&\quad\leq C\rho^{1+\kappa/d}\left(\int_0^t \tau^{(p+1)/\kappa-1}\phi^{d(p+1)/p}(\tau)\psi(\tau)d\tau\right. \\
&\qquad\left. + \int_0^t \tau^{1/\kappa-1}\phi^{d/p}(\tau)\psi(\tau)d\tau\right)^{(p-1)/p}\left(\int_0^t \tau^{1/\kappa-1}\phi^{d/p}(\tau)\psi(\tau)d\tau\right)^{1/p}.
\end{aligned}
$$

Proof. The Hölder inequality yields

$$
\begin{aligned}
&\int_0^t \int_{B_{2\rho}} |\nabla v^m|^{p-1}\zeta^{p-1} dx d\tau \\
&\leq\left(\int_0^t \int_{B_{2\rho}} \tau^{1/p}|\nabla v^m|^p (v^m+\epsilon)^{-(m+1)/mp}\zeta^p dx d\tau\right)^{1-1/p} \\
&\qquad\times\left(\int_0^t \int_{B_{2\rho}} \tau^{-(p-1)/p}(v^m+\epsilon)^{(m+1)(p-1)/mp} dx d\tau\right)^{1/p}
\end{aligned}
\tag{5.14}
$$

for any $\epsilon > 0$. Multiplying the equation by $g_\epsilon(v) = t^{1/p}(v^m+\epsilon)^{d/pm}\zeta^p$ and integrating it over $B_{2\rho}$, we have

$$
\begin{aligned}
&\int_0^t \int_{B_{2\rho}} \tau^{1/p}|\nabla v^m|^p (v^m+\epsilon)^{d/mp-1}\zeta^p dx d\tau \\
&\leq C\rho^{-p}\int_0^t \int_{B_{2\rho}} \tau^{1/p}(v^m+\epsilon)^{(mp^2-m-1)/mp} dx d\tau + C\int_0^t \int_{B_{2\rho}} \tau^{-(p-1)/p}\Lambda(\epsilon) dx d\tau,
\end{aligned}
\tag{5.15}
$$

where

$$\Lambda(\epsilon) = \int_0^v (s^m + \epsilon)^{d/mp} ds.$$

Taking the limit as $\epsilon \to 0$, by (5.14) and (5.15), we have

$$\int_0^t \int_{B_{2\rho}} |Dv^m|^{p-1} \zeta^{p-1} dx d\tau \le C(I_1 + I_2)^{(p-1)/p} I_2^{1/p},$$

where

$$I_1 \equiv \rho^{-p} \int_0^t \int_{B_{2\rho}} \tau^{1/p} v^{(mp^2-m-1)/p} dx d\tau, \quad I_2 \equiv \int_0^t \int_{B_{2\rho}} \tau^{-(p-1)/p} v^{(m+1)(p-1)/p} dx d\tau.$$

By (5.13),

$$I_1 \le C\rho^{1+\kappa/d} \int_0^t \tau^{(p+1)/\kappa - 1} \phi^{(p+1)d/p}(\tau) \psi(\tau) d\tau$$

and

$$I_2 \le C\rho^{1+\kappa/d} \int_0^t \tau^{1/\kappa - 1} \phi^{d/p}(\tau) \psi(\tau) d\tau.$$

Therefore, the proof of Lemma 5.3 is complete. □

By Lemmas 5.2 and 5.3, we can prove the following proposition by an argument similar to that of §3 in [6].

PROPOSITION 5.4. *Let u be a weak solution of* (5.1). *Then there exist constants C_i, $i = 0, 1, 2$, such that*

$$\phi(t) \le C_1 |||\mu|||_r^{p/\kappa}, \qquad \psi(t) \le C_2 |||\mu|||_r \tag{5.16}$$

for all $0 < t < C_0 |||\mu|||_r^{-d}$. Furthermore, there exist constants C_i, $i = 3, 4$, such that for any $\rho \ge r > 0$,

$$\|v(\cdot, t)\|_{\infty, B_\rho} \le C_4 t^{-N/\kappa} \rho^{p/d} |||\mu|||_r^{p/k}, \tag{5.17}$$

$$\int_0^t \int_{B_\rho} |\nabla v^m|^{p-1} dx d\tau \le C_5 t^{1/\kappa} \rho^{1+\kappa/d} |||\mu|||_r^{1+d/\kappa} \tag{5.18}$$

for all $0 < t < C_0 |||\mu|||_r^{-d}$.

By Proposition 5.4, we can prove Theorem 1.2. In fact, for any σ-finite nonnegative Borel measure μ satisfying $|||\mu|||_r$ for some $r > 0$, there exists a sequence $\{\mu_n\}_{i=1}^\infty \subset C_0^\infty(\mathbf{R}^N)$ such that

$$\lim_{n \to \infty} \int_{\mathbf{R}^N} \mu_n \varphi dx = \int_{\mathbf{R}^N} \varphi d\mu$$

for any $\varphi \in C_0(\mathbf{R}^N)$ and $|||\mu_n|||_r \to |||\mu|||_r$ as $n \to \infty$ for any $r > 0$. Let u_n be a solution of (5.1) for initial condition μ_n. Then by (5.3), (5.17), and (5.18), taking the limit as $n \to \infty$, we can see that there exists a solution of (1.1) for initial condition μ

by the same argument as in the proof of Theorem 1.3. Therefore, by Proposition 5.4, we can complete the proof of Theorem 1.2.

6. $L^\infty_{\rm loc}(\mathbf{R}^N)$ estimate for Case (III).

In order to prove Theorem 1.4, we need some $L^\infty_{\rm loc}(\mathbf{R}^N)$ estimates of the nonnegative solutions. By Theorems 2–1 and 2–2 in [19], we have the following proposition.

PROPOSITION 6.1. *Let v be a locally bounded nonnegative local weak solution of*

$$\frac{\partial}{\partial t}v = \operatorname{div}(|\nabla v^m|^{p-2}\nabla v^m) \quad in \quad \mathbf{S}_\infty, \quad d = m(p-1)-1<0. \tag{6.1}$$

If $\kappa_r > 0$, then there exist constants $C_i = C_i(N,m,p,r)$, $i = 1,2,3,4$, such that for any $t > 0$ and $\rho > 0$,

$$\sup_{x\in B_\rho} v(x,t) \le C_1 t^{-N/\kappa_r}\left(\sup_{0<\tau<t}\int_{B_{2\rho}} v^r(x,\tau)dx\right)^{p/\kappa_r} + C_2\left(\frac{t}{\rho^p}\right)^{p/\kappa_r} \tag{6.2}$$

and

$$\sup_{0<\tau<t}\int_{B_\rho} v^r(x,\tau)dx \le C_3\int_{B_{2\rho}} v^r(x,0)dx + C_4\left(\frac{t^r}{\rho^{\kappa_r}}\right)^{1/-d}. \tag{6.3}$$

We can also prove Proposition 6.1 in view of the arguments of the previous sections and [7]. We remark that Lemma 5.1 holds for equation (6.1).

The following lemma is presented via the same arguments as in Lemma I.4.1 and Corollary III.3.1 of [7] with minor changes.

LEMMA 6.2. *There exists a constant $C = C(N,m,p)$ such that for any $0 < s < t$, $\rho > 0$,*

$$\begin{aligned}\frac{1}{\rho}\int_s^t\int_{B_\rho} |\nabla v^m|^{p-1}dxd\tau \le C\left(\frac{t-s}{\rho^\kappa}\right)^{1/p} \\ \times\left\{\int_{B_{2\rho}} v(x,0)dx + \left(\frac{t}{\rho^\kappa}\right)^{-d}\right\}^{(m+1)(p-1)/p}\end{aligned} \tag{6.4}$$

Therefore, by (6.2)–(6.4), we can complete the proof of Theorem 1.4 in the same way as in that of Theorem 1.2.

REFERENCES

[1] D. G. ARONSON, *Non-negative solutions of linear parabolic equations*, Ann. Scuola. Norm. Sup. Pisa Cl. Sci., 22 (1968), pp. 607–694.

[2] ———, *Widder's inversion theorem and the initial distribution problem*, SIAM J. Math. Anal., 12 (1981), pp. 639–651.

[3] V. BARBU, *Nonlinear Semi-group and Differential Equations in Banach Spaces*, Noordhoff, Leiden, The Netherlands, 1976.

[4] P. BENILAN, M. G. CRANDALL, AND M. PIERRE, *Solutions of the porous medium equation under optimal conditions on initial values*, Indiana Univ. Math. J., 33 (1984), pp. 51–87.

[5] E. DIBENEDETTO, *Degenerate Parabolic Equations*, Springer-Verlag, Berlin, New York, Heidelberg, 1993.

[6] E. DIBENEDETTO AND M. A. HERRERO, *On the Cauchy problems and initial traces for a degenerate parabolic equation*, Trans. Amer. Math. Soc., 314 (1989), pp. 187–224.

[7] E. DiBenedetto and M. A. Herrero, *Nonnegative solutions of the evolution p-Laplacian equations: Initial traces and Cauchy problem when* $1 < p < 2$, Arch. Rational Mech. Anal., 111 (1990), pp. 225–290.

[8] E. DiBenedetto and T. C. Kwong, *Harnack estimates and extinction profile for weak solutions of certain singular parabolic equations*, Trans. Amer. Math. Soc., 330 (1992), pp. 783–811.

[9] Y. Jingxue, *Solutions with compact support for nonlinear diffusion equations*, Nonlinear Anal., 19 (1992), pp. 309–321.

[10] O. A. Ladyzeskayaya, N. S. Solonikov, and N. N. Ural'tseva, *Linear and quasilinear equations of parabolic type*, Transl. Math. Monographs, 23 (1968), p. 95.

[11] J. L. Lions, *Quelques Méthodes de Résolution des Problémes aux Limites Non Linéaires*, Dunod and Gauthier–Villars, Paris, 1969.

[12] M. M. Porzio and V. Vespri, *Hölder estimates for local solutions of some doubly nonlinear degenerate parabolic equations*, J. Differential Equations, 103 (1993), pp. 146–178.

[13] P. A. Raviart, *Sur la résolution de certains équations paraboliques nonlinéaires*, J. Funct. Anal., 5 (1970), pp. 299–328.

[14] S. Tacklind, *Sur les classes quasianalitiques des solutions des équations aux derivée partielles du type parabolique*, Acta Reg. Soc. Sci. Uppsaliensis, 10 (1936), pp. 3-55.

[15] N. S. Trudinger, *Pointwise estimates and quasilinear parabolic equations*, Comm. Pure Appl. Math., 21 (1968), pp. 206–226.

[16] M. Tsutsumi, *On solutions of some doubly nonlinear degenerate parabolic equations*, J. Math. Anal. Appl., 132 (1988), pp. 187–212.

[17] A. N. Tychonov, *Thórèmes d'unicité pour l'équation de la chaleur*, Mat. Sb., 42 (1935), pp. 199–216.

[18] V. Vespri, *On the local behaviour of solutions of a certain class of doubly nonlinear parabolic equations*, Manuscripta Math., 75 (1992), pp. 65–80.

[19] ———, *Harnack type inequalities for solutions of certain doubly nonlinear parabolic equations*, J. Math. Anal. Appl., 181 (1994), pp. 104–131.

[20] D. V. Widder, *Positive temperatures in an infinite rod*, Trans. Amer. Math. Soc., 55 (1944), pp. 85–95.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1261–1269, September 1996

004

GLOBAL STABILITY OF TRAVELING FRONTS AND CONVERGENCE TOWARDS STACKED FAMILIES OF WAVES IN MONOTONE PARABOLIC SYSTEMS*

JEAN-MICHEL ROQUEJOFFRE†, DAVID TERMAN‡, AND VITALY A. VOLPERT§

Abstract. A class of parabolic systems for which the maximum principle is valid is investigated. When two stable rest points can be connected by a traveling front, any solution of the Cauchy problem which initially has these two rest points as spatial limits will become monotone in finite time on every compact interval and converge to a traveling front. As an application, convergence to stacked waves is discussed.

Key words. global stability, traveling waves, wave stacks

AMS subject classifications. 35B40, 35K45

1. Introduction.

The problem under consideration is the long-time behaviour of the solutions of the Cauchy problem for a class of semilinear parabolic systems of the form

$$\begin{aligned} u_t - Du_{xx} &= F(u), \\ u(0,x) &= u_0(x). \end{aligned} \tag{1.1}$$

The unknown is the real-vector function $u(t,x) =. (u_1(t,x),\dots,u_n(t,x))$, the nonlinearity $F(u) = (F_1(u),\dots,F_n(u))$ is smooth, say of class C^1 in u, and D is a diagonal matrix with positive diagonal coefficients $(D_1,\dots,D_n)$. Such a system is called monotone if the following additional condition holds:

$$\forall i \in [1,n],\ \forall j \neq i,\quad \frac{\partial F_i}{\partial u_j} \geq 0. \tag{1.2}$$

It is well known [G], [VV1] that (1.2) implies a comparison principle: if $u_{10} \leq u_{20}$, then $u_1 \leq u_2$ as long as both solutions exist. The inequalities are meant to hold componentwise.

Traveling-front solutions of (1.1) are solutions of the form $u(t,x) = w(x+ct)$; if $F(w_-) = F(w_+) = 0$, then the front is said to connect the two rest points w_- and w_+ if it goes to w_- (resp. w_+) as $x \to -\infty$ (resp. as $x \to +\infty$). Denoting $w' = \frac{dw}{dx}$, we write the differential system satisfied by $w(x)$:

$$\begin{aligned} -w'' + cw' &= F(w), \\ \lim_{x\to\pm\infty} w(x) &= w_\pm. \end{aligned} \tag{1.3}$$

The existence of a solution (c,u) for (1.3) has been investigated in a number of frameworks. The scalar version is now understood. Systems of equations have also been

*Received by the editors May 6, 1994; accepted for publication (in revised form) April 14, 1995.

†UFR MIG, UMR CNRS 5640, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse cedex, France.

‡Department of Mathematics, Ohio State University, Columbus, OH 43210. The research of this author was supported in part by NSF grant DMS-9203299LE.

§Laboratoire d'Analyse Numérique, Université de Lyon I, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne cedex, France.

extensively studied. For example, a model for competing species was considered in [G], while a model for chemical activity on isothermal catalyst surfaces was studied in [FT]. These are both systems with two equations; the more general case was considered in [VV1]. Nonmonotone systems also arise in many applications—in particular, combustion theory. Systems with two equations are considered in [BNS] and [T1], while models involving complex chemistry are studied in [B] and [T2]. This list is, of course, by no means exhaustive.

Henceforth, system (1.1) will always be assumed to be monotone, and this will not be mentioned again. Assume that F has two stable rest points $w_- < w_+$. Also assume that in the interval $[w^-, w^+]$, the only rest points w are such that the matrix $F'(w)$ has at least one eigenvalue in the right half-plane. This assumptions is meant to disallow waves that connect intermediate rest points, which could prevent the existence of a wave connecting w_- and w_+. Other stable rest points, not in the interval $[w_-, w_+]$, are allowed.

From [VV1], there indeed exists a unique $c \in I\!R$ and a unique—up to the translations—profile w such that (1.3) holds; furthermore, $w' > 0$. Henceforth, we will always denote by w the unique solution of (1.3) such that

$$w(0) = \frac{w^- + w^+}{2}. \tag{1.4}$$

It is proved in [VV2] that if the initial datum is monotone in x and satisfies decay conditions at $\pm\infty$, then $u(t)$ converges exponentially to a front. It is the purpose of this paper to remove the monotonicity assumption and to examine some consequences of this result when F has multiple stable rest points.

Let us recall that global stability for the bistable scalar case was treated by Fife and McLeod [FML]. The local stability of traveling fronts for monotone systems with two equations was considered in [G] and [FT]. In several space dimensions with nonhomogeneous convection, the problem has recently been solved by the first author in [R1] (monotone initial data) and [R2] (nonmonotone initial data).

The paper is organized as follows. In §2, we state our main results, and §§3 and 4 are devoted to their proofs.

2. Notations and main results. Let us denote by $\mathcal{S}(t)u_0$ the solution of (1.1)—which may, by the way, not exist globally. Let $UC(I\!R, I\!R^n)$ be the set of all bounded, uniformly continuous functions from $I\!R$ to $I\!R^n$; for $u \in UC(I\!R, I\!R^n)$ and $h \in I\!R$, define $\tau_h u(x) := u(x+h)$.

We will call $w_0 \in I\!R^n$ a *stable rest point* of F if $F(w_0) = 0$ and the eigenvalues of the matrix $F'(w_0)$ have negative real parts.

THEOREM 2.1. *Assume that $w^- \leq w^+$ are two stable rest points connected by a monotone front w with speed c. Let $u_0 \in UC(I\!R, I\!R^n)$ satisfy $w^- \leq u_0(x) \leq w^+$, and set*

$$\varepsilon(u_0) = \sup\left(\limsup_{x\to-\infty} |u_0(x) - w^-|, \ \limsup_{x\to+\infty} |u_0(x) - w^+| \right). \tag{2.1}$$

Then, if $\varepsilon(u_0)$ is small enough, $\mathcal{S}(t)u_0$ is defined for all time and there exist $x_0 \in I\!R$ and $\omega > 0$ such that

$$\|\mathcal{S}(t)u_0 - \tau_{x_0+ct}w\|_\infty = O(e^{-\omega t}).$$

When multiple ordered stable rest points exist, the long-time behaviour of $\mathcal{S}(t)u_0$ can be a difficult problem. However, the following natural case can be treated.

THEOREM 2.2 *Assume that* $w^- = w^1_\infty < \cdots < w^l_\infty < \cdots < w^k_\infty = w^+$, $1 \le l \le k$, *are* k *rest points such that the eigenvalues of the matrices* $F'(w^l)$ *have negative real parts. Assume that for all* $l \in [1,k]$, w^l_∞ *and* w^{l+1}_∞ *are connected by a monotone front* w^l *with speed* c^l *and that* $c^{l+1} < c^l$.

Let $u_0 \in UC(I\!R, I\!R^n)$ *satisfy* $w^- \le u_0(x) \le w^+$. *Then if* $\varepsilon(u_0)$ *is small enough—recall that* $\varepsilon(u_0)$ *is defined by* (2.1)—(1.1) *has a global solution* $\mathcal{S}(t)u_0$ *and there exist* $(x^1, \ldots, x^k) \in I\!R^k$ *and* $\omega > 0$ *such that*

$$\left\| \mathcal{S}(t)u_0 - w^1_\infty - \sum_{l=1}^{k} \left(\tau_{x^l + c^l t} w^l - w^l_\infty\right) \right\|_\infty = O(e^{-\omega t}).$$

Remarks. 1. Theorem 2.2 implies nonexistence results for certain solutions of (1.3); in particular, its assumptions preclude the existence of a single wave connecting w^- to w^+.

2. In contrast with the preceding remark, we may consider the union of the monotone fronts w^l as a stable wave train which connects w^- with w^+. The existence of stacked waves for systems with two equations was considered in [FT]. It was demonstrated there that if w^- and w^+ are any distinct, stable rest points and all of the rest points are nondegenerate, then there must exist a wave train which connects w^- with w^+. The stacked family may, of course, consist of a single traveling-wave solution. Theorem 2.2 demonstrates that this wave train is globally, asymptotically stable. A similar result concerning the existence of wave trains is proved in [T2], where a nonmonotone system is considered. The extension of this result to more general systems will be addressed elsewhere.

3. Convergence to traveling fronts.

Assume that the assumptions of Theorem 2.1 are satisfied. In particular, w is a monotone front with speed c. The first task in studying the stability of w is to rewrite (1.1) in the reference frame of the traveling front. This yields

$$\begin{aligned} u_t - Du_{xx} + cu_x &= F(u), \\ u(0,x) &= u_0(x). \end{aligned} \tag{3.1}$$

The solution will still be called $\mathcal{S}(t)u_0$. As is known, three ingredients are required in the proof of convergence: local stability, precompactness of the orbits, and quasiconvergence. One cannot hope to prove quasiconvergence without precompactness, and global stability is implied by local stability and quasiconvergence. Local stability is a consequence of Theorem 4.1 in [VV2], and we state it in a more general form which will be useful in §4.

THEOREM 3.1. *For* $v_0 \in UC(I\!R, I\!R^n)$ *and* $h \in UC(I\!R_+ \times I\!R, I\!R^n)$, *let* $u(t,x)$ *be the solution of*

$$\begin{aligned} u_t - Du_{xx} + cu_x &= F(u) + h(t), \\ u(0) &= w + v_0. \end{aligned} \tag{3.2}$$

Assume that $\|h(t)\|_\infty \le \delta e^{-2\omega t}$*—the same* ω *as in Theorem* 2.1. *There exist* $\gamma(v_0,h)$ *and* $C(v_0,h) > 0$ *defined and bounded for small* δ *and* $\|v_0\|_\infty$ *such that*

$$\|\mathcal{S}(t)u_0 - \tau_{\gamma(v_0,h)} w\|_\infty \le C(v_0,h)e^{-\omega t}.$$

We now turn to precompactness. The most common way to prove this is to trap $\mathcal{S}(t)u_0$ between two translates of the wave, as is stated below. Set $p_0 := (1, \ldots, 1)$.

LEMMA 3.2. *Let u_0 be as in Theorem 2.1. There exist $q > 0$, $\omega > 0$, and $\xi_1 < \xi_2$ such that for t_0 large enough, it holds that*

$$\forall t \geq t_0, \quad \tau_{\xi_1} w - qp_0 e^{-\omega t} \leq \mathcal{S}(t)u_0 \leq \tau_{\xi_2} w + qp_0 e^{-\omega t}. \tag{3.3}$$

Proof. Choose $\varepsilon_0 > 0$ so small that for all v_0 lying in the cube $[-\varepsilon_0, \varepsilon_0]^n$, the matrices $F'(w_- + v)$ and $F'(w_+ + v)$ have eigenvalues of negative real parts. As a consequence, the solutions $v_\pm(t)$ of the two differential systems

$$\frac{dv_-}{dt} = F(v_-), \quad v_-(0) = w^- + v_0,$$
$$\frac{dv_+}{dt} = F(v_+), \quad v_+(0) = w^+ + v_0$$

approach $w_\pm$ exponentially as $t \to +\infty$. Therefore, if $\varepsilon(u_0) \leq \varepsilon_0$, the function $\varepsilon(\mathcal{S}(t)u_0)$ goes exponentially to 0. In Theorem 3.1, we set $h \equiv 0$ and select $t_0 > 0$ large enough so that Theorem 3.1 works with $v_0 \equiv \varepsilon(\mathcal{S}(t_0)u_0)p_0$. Furthermore, there exists $M > 0$ such that

$$\tau_{-M}(w - v_0) \leq \mathcal{S}(t_0)u_0 \leq \tau_M(w + v_0).$$

From the comparison principle, we have, for all $t \geq t_0$,

$$\mathcal{S}(t - t_0)\big(\tau_{-M}(w - v_0)\big) \leq \mathcal{S}(t)u_0 \leq \mathcal{S}(t - t_0)\big(\tau_M(w + v_0)\big).$$

Since the semigroup commutes with the translations, we get (3.3). □

Lemma 3.2 implies global existence and precompactness. Therefore, it remains to prove quasiconvergence, and we will use the method of [R2]. We first need a lemma to replace the Harnack inequalities that were used in [R2], and in this scope we consider the linear parabolic system

$$u_t - Du_{xx} + cu_x = A(t,x)u, \quad x \in I\!R, \tag{3.4}$$

where $A(t,x) := \big(a_{ij}(t,x)\big)_{1 \leq i,j \leq n}$ is bounded and continuous and satisfies $a_{ij}(t,x) \geq 0$ if $i \neq j$. Henceforth, if $u(t,x)$ is a bounded solution of (3.4), we will set $\sup_{x \in I\!R} u(t,x) := \sup_{1 \leq i \leq n,\ x \in I\!R} u_i(t,x)$, and the same convention will be used for the infimum.

LEMMA 3.3. *Let $u(t,x) \geq 0$ be a globally bounded solution of* (3.4) *for which there exist $\mu, \alpha > 0$ such that*

$$\sup_{x \in I\!R} \left(\inf_{1 \leq i \leq n} u_i(t,x) \right) = \max_{x \in [-\alpha,\alpha]} \left(\min_{1 \leq i \leq n} u_i(t,x) \right) = \mu.$$

Choose $M > \alpha$ and $t \geq 1$. There exists $\eta(M) > 0$ such that $\inf_{x \in [-M,M]} u(t,x) \geq \eta(M)$.

Proof. For $t \geq 1$, let $x_t \in [-\alpha, \alpha]$ such that $u(t, x_t) \leq \mu p_0$. From the parabolic estimates, u_x is globally bounded; therefore, there exists $K > 0$ such that

$$u(t,y) \geq (\mu - K|x_t - y|)p_0$$

for all $y \in I\!R$; as a consequence, we get, for $|x_t - y|$, $u(t,y) \geq \frac{\mu}{2}p_0$.

Let us assume without loss of generality that $N := \frac{2K}{\mu}$ is a nonzero integer. For $j \in [-N, N]$, set $x_j := \frac{j}{N}$ and $I_j := [x_j, x_{j+1}]$. We choose $j \in [-N, N]$ such that $x_t \in I_j$; for $y \in I_j$, we may write—thanks once again to the parabolic estimates—

$$\forall \tau \in \left[-\frac{1}{2}, \frac{1}{2}\right], \quad u(t+\tau, y) \geq u(t, y) - K'\tau p_0,$$

where K' depends on neither t nor τ—recall that t was taken > 1. Thus for $0 \leq |\tau| \leq \delta := \frac{1}{4KK'}$, we have

$$u(t+\tau, y) \geq \frac{\mu}{4} p_0. \tag{3.5}$$

We now define by induction an increasing sequence $(t_p)_p$ such that $(t_{p+1} - t_p)_p$ is bounded away from 0. Set $t_0 := 0$ and for $p \geq 1$ assume that, for all $k \leq p$, we have managed to define t_k and an integer $j_k \in [-N, N]$ such that $x_{t_k} \in I_{j_k}$; also, assume that $t_{k+1} - t_k \geq \delta$ for $k \leq n-1$. The time t_{p+1} will be chosen to be the first instant when the inequality $u(t, y) \geq \frac{\mu}{4} p_0$ ceases to hold in I_{j_p}; obviously, from (3.5), we have $t_{p+1} - t_p \geq \delta$; however, t_{p+1} may be infinite, but this will not bother us too much.

Notice that for all n, the abscissa x_{t_p} can be chosen to remain in $[-\alpha, \alpha]$; this is allowed by the assumption.

Now choose $M > \alpha$; without loss of generality, we take $M \geq 1 + \alpha$. Set $\underline{a}_{ij} := \inf_{t \geq 0, x \in I\!R} a_{ij}(t, x)$ and $\underline{A} := (\underline{a}_{ij})_{1 \leq i,j \leq n}$. Setting $s_p := t_p - \frac{\delta}{2}$, we define the sequence $(u_p(t, x))_p$ by

$$\begin{aligned} \partial_t u_p - \partial_{xx} u_p + c\partial_x u_p &= \underline{A} u_p, \quad (t, x) \in]s_p, t_{p+1}] \times] - \infty, x_{j_p}], \\ u_p(t, x_{j_p}) &= \frac{\mu}{4}, \\ u_p(s_p, x) &= 0, \end{aligned} \tag{3.6}$$

and we also define the sequence $(v_p(t, x))_p$ by the same boundary problem as (3.6), but replacing $] - \infty, x_{j_p}]$ by $[x_{j_{p+1}}, +\infty[$. As is well known, the functions u_p and v_p are time-increasing for every p; moreover, we have $u_p(t, .) = \tau_{x_{j_p} - x_0} u_0(t - s_p)$ (resp. $v_p(t, .) = \tau_{x_{j_{p+1}} - x_0} v_0(t - s_p)$). Furthermore, the comparison principle implies that for every p and every $t \in [s_p, t_{p+1}]$ and $x \notin I_{j_p}$, $\sup(u_p(t), v_p(t)) \leq u(t)$. This together with (3.5) implies

$$\inf_{x \in [-M, M]} u(t, x) \geq \inf\left(\left(\mu, \inf_{x \in [-M, M]} (u_0(t_0, x), v_0(t_0, x))\right)\right) := \eta(M).$$

From (3.6) and the properties of u_p and v_p, we have $\eta(M) > 0$. □

We may now turn to the proof of Theorem 2.1. We will use the idea of [R2], with minor modifications. Notice that by precompactness, $\omega(u_0)$ is nonempty and compact in $UC(I\!R, I\!R^n)$.

LEMMA 3.4. *Let u_0 satisfy the assumptions of Theorem 1.1. Then there exists an x-increasing element in $\omega(u_0)$.*

Proof. From Lemma 3.2, there exist two real numbers $h_1 \leq h_2$ such that

$$\forall \psi \in \omega(u_0), \quad \tau_{h_1} w \leq \psi \leq \tau_{h_2} w. \tag{3.7}$$

Denote by h the following application, defined on $\omega(u_0)$:

$$\forall\psi \in \omega(u_0), \quad h(\psi) = \mathrm{Inf}\{h > 0 : \ \forall k \geq h, \tau_k\psi \geq \psi\};$$

from (3.7), we see that $h(\psi)$ is finite for $\psi \in \omega(u_0)$. Also, by compactness, h attains its minimum h_0 at some point $\psi_0 \in \omega(u_0)$. Assume that $h_0 > 0$. Notice that from the definition of h_0 and the maximum principle, we have $h\big(\mathcal{S}(t)\psi_0\big)= h_0$.

Let us first notice that, because of (3.7), there exists $\mu > 0$ such that

$$\forall t \geq 0, \quad \min_{1\leq i \leq n}\left\{\sup_{x\in\mathbb{R}}\Big(\mathcal{S}(t)\psi_{0,i}(x+h_0) - \mathcal{S}(t)\psi_{0,i}(x)\Big)\right\} \geq \mu. \tag{3.8}$$

For every $\delta \in \mathbb{R}$, let $v^\delta(t,x,y)$ be defined as $v^\delta = \tau_{h_0-\delta}\mathcal{S}(t)\psi_0 - \mathcal{S}(t)\psi_0$. The function v^0 solves

$$v^0_t + Dv^0_{xx} + cv^0_x = A(t,x)v^0 \quad \text{in } \mathbb{R}$$

with $A(t,x) = \int_0^1 F'(\sigma\tau_{h_0-\delta}\mathcal{S}(t)\psi_0 + (1-\sigma)\mathcal{S}(t)\psi_0 \, d\sigma$. The matrix $A(t,x)$ satisfies the assumptions of Lemma 3.3; therefore, for every large enough $M > 0$, there exist $\delta_0(M) \in]0,h_0[$ and $\mu(M) > 0$ such that

$$\forall\delta \in [0,\delta_0(M)[, \ \forall t \geq 1, \forall x \in [-M,M], \quad v^\delta(t,x) \geq \mu(M).$$

Let ε_0 be defined as in Lemma 3.2 and $M > 0$ be large enough so that for all $k \in [0,h]$ all $x \notin [-M,M]$ and all $t \geq 1$, we have $\tau_k\mathcal{S}(t)\psi_0(x) \in [-\varepsilon_0,\varepsilon_0]^n$. We are now going to study the evolution of $\mathcal{S}(t)\psi_0$. By compactness, there exist a sequence $(t_n)_n$ and $\psi_\infty \in \omega(u_0)$ such that

$$\lim_{n\to+\infty} \|\mathcal{S}(t_n)\psi_0 - \psi_\infty\|_\infty = 0.$$

From the definition of h_0, we know that for all $x \in [-M,M]$ and $k \geq h_0 - \delta_0$, $\psi_\infty(x+k) \geq \psi_\infty(x)$. Let us see what happens for $x \geq M$. From the definition of ε_0 and the maximum principle, we have

$$v^\delta(t,x) \geq -\varepsilon_0 e^{\int_0^t A(s,x)\,ds} p_0. \tag{3.9}$$

From the definition of ε_0, the right-hand side of (3.9) goes exponentially to 0. Furthermore, a similar phenomenon occurs for $x \leq -M$.

To sum up, we have just constructed $\psi_\infty \in \omega(u_0)$ such that for every $k \geq h_0 - \delta_0$, the inequality $\tau_k\psi_\infty \geq \psi_\infty$ holds, which implies that $h(\psi_\infty) \leq h_0 - \delta_0$. This contradicts the definition of h_0. □

Proof of Theorem 2.1. Let $\psi_0 \in \omega(u_0)$ be increasing in x. From Theorem 5.1 of [VV2]—convergence to traveling fronts for x-increasing initial data—$\mathcal{S}(t)\psi_0$ converges to some $\tau_{x_0}w$. From the closedness of $\omega(u_0)$, we infer that $\tau_{x_0}\psi_0 \in \omega(u_0)$, which is exactly quasiconvergence. □

4. Convergence to stacked waves. The first idea proceeds as in [FML]: first, construct sub- and supersolutions to make sure that everything happens in the right reference frames; we then use Theorem 3.1 to complete the proof. However, this method cannot be carried along the whole way: to make it work, one would need an additional assumption, namely, the existence of $p > 0$ such that every $F'(w^l_\infty)p > 0$ for all l.

This assumption is not irrelevant, but we prefer another method which works without it. Let us first remark that combining Theorem 2.1 with Theorem 3.1 yields the following result.

THEOREM 4.1. *Let u_0 satisfy the assumptions of Theorem* 2.1. *For $h \in UC(I\!R_+ \times I\!R, I\!R^n)$, let $u(t,x)$ be the solution of*

$$u_t - Du_{xx} + cu_x = F(u) + h(t), \qquad u(0) = w + u_0. \tag{4.1}$$

Assume that $h(t) = O(e^{-2\omega t})$. There exists $\varepsilon_0 > 0$ such that if $\varepsilon(u_0^l) \leq \varepsilon_0$ is small enough—see (2.1) *for the definition—$u(t)$ is defined for all time and there exists $x_0 \in I\!R$ such that*

$$\|u(t) - \tau_{x_0+ct}w\|_\infty = O(e^{-\omega t}).$$

The proof is standard and will not be given. The following consequence of Theorem 4.1 will be needed.

LEMMA 4.2. *Let α, k, and γ be positive. With the notations of Theorem* 2.1, *let $u(t,x)$ solve*

$$\begin{aligned} u_t - Du_{xx} + cu_x &= F(u), \quad x \geq -\alpha t, \\ u(t, -\alpha t) &= w_- - kp_0 e^{-\gamma t}, \\ u(t, +\infty) &= w_+. \end{aligned} \tag{4.2}$$

There exists $x_0 \in I\!R$ such that $\|u(t) - \tau_{x_0+ct}w\|_{L^\infty([-\alpha t, +\infty[)} = O(e^{-\omega t})$.

Proof. The proof requires two steps. Throughout the proof, the notation of §3 will be used.

Step 1: *Estimates at $x = -\alpha t$.* Let $\overline{u}_0$ be x-increasing and such that

$$\lim_{x \to -\infty} \overline{u}_0(x) = w_-, \quad \overline{u}_0 \geq u(0).$$

Clearly, $u \leq \overline{u}$; therefore, by Theorem 3.1, we have $u(t,x) \leq \tau_{x_1}w + O(e^{-\omega t})$.

Now recall that $F'(w_-)^{-1}$ sends the nonpositive cone of $I\!R^n$ onto its interior. To see this, we only have to write

$$F'(w_-)^{-1} = \lim_{t \to +\infty} \int_0^t e^{(t-s)F'(w_-)}\, ds,$$

which is obviously nonpositive by the comparison principle, and even negative. By continuity and Frobenius–Perron theorem, there exists $\varepsilon_0 > 0, \mu > 0$ and a vector $p > 0$ such that

$$\forall v \in [-\varepsilon_0, \varepsilon_0]^n, \quad F'(w_- + v)q \leq -\mu q. \tag{4.3}$$

Choosing $\gamma_1 < \inf(\gamma, \mu)$, we see that the function $\underline{u}(t) := w_- - \varepsilon q e^{-\gamma_1 t}$ is a subsolution to (4.2) as soon as ε is small enough. As a consequence, we get $\underline{u}(t) \leq u(t,x) \leq \overline{u}(t,x)$ once all the parameters are conveniently chosen. This yields an estimate of the form $\|u(t) - w_-\|_{L^\infty([-\alpha t, -\alpha t+2[)} = O(e^{-\omega t})$; from boundary estimates, we finally infer

$$|u_x(t, -\alpha t)| + |u_x(t, -\alpha t)| + |u_t(t, -\alpha t)| = O(e^{-\gamma_1 t}). \tag{4.4}$$

Step 2: *Convergence.* Let Γ be a C^∞ real-valued, nonnegative, nondecreasing function that is equal to 0 on $I\!R_-$ and to 1 on $[1,+\infty[$. Define componentwise the function $v(t,x)$ on the whole real line as

$$v_i(t,x) = (w_-)_i + \big(u_i(t,x) - (w_-)_i\big)\Gamma(x + \alpha t - 1). \tag{4.5}$$

From Step 1 above, the function v solves an equation of the type (4.1), and we may apply Theorem 4.1 to obtain Lemma 4.2. □

Let us now prove that, as we stated above, everything happens in the right reference frames.

LEMMA 4.3: *Let u_0 fulfill the assumptions of Theorem* 2.2. *There exist* $2k$ *real numbers* $\xi_\alpha^1 < \cdots < \xi_\alpha^l < \xi_\alpha^k$, $\alpha \in \{1,2\}$, *and* $q, \omega > 0$ *such that*

$$w_\infty^1 + \sum_{l=1}^{k}\big(\tau_{\xi_1^l + c^l t} w^l - w_\infty^l\big) - qp_0 e^{-\omega t} \le \mathcal{S}(t)u_0$$
$$\le w_\infty^1 + \sum_{l=1}^{k}\big(\tau_{\xi_2^l + c^l t} w^l - w_\infty^l\big) + qp_0 e^{-\omega t}.$$

Proof. Only the lower bound will be dealt with since the upper bound is similar. By an easy induction on $l \in [1,k]$, it is possible to find k functions $u^1(t,x), \dots, u^l(t,x), \dots, u^k(t,x)$ defined on some time interval of the form $[t_0, +\infty[$, $t_0 > 0$, large enough such that

- for all l, u^l is a solution of (4.2) with $\alpha = \frac{c^l + c^{l+1}}{2}$ and $c = c^l$,
- $\lim_{x \to +\infty} u^l(t,x) = w_\infty^{l+1}$,
- $u^l\big(t, \frac{c^l - c^{l+1}}{2}t\big) = u^{l+1}\big(t, -\frac{c^l - c^{l+1}}{2}t\big)$ for $l \le k$,
- $u^1(t,x) \le \mathcal{S}(t)u_0$,
- $u^l(t) \le \tau_{-c^l t}\mathcal{S}(t)u_0$, for $t \ge t_0$, $x \ge \frac{c^l - c^{l-1}}{2}$, and $l \ge 1$.

Now define $\underline{u}(t,x)$ as

$$\underline{u}(t,x) = \begin{cases} \tau_{c^1 t} u^1(t,x) & \text{if } \ x \le -\dfrac{c^1 + c^2}{2}t, \\ \tau_{c^l t} u^l(t,x) & \text{if } \ -\dfrac{c^l + c^{l-1}}{2}t \le x \le -\dfrac{c^l + c^{l+1}}{2}t, \\ \tau_{c^k t} u^k(t,x) & \text{if } \ x \ge -\dfrac{c^k + c^{k-1}}{2}t. \end{cases} \tag{4.5}$$

Obviously, $\underline{u}(t,x) \le \mathcal{S}(t)u_0$; application of Lemma 4.2 yields Lemma 4.3. □

Proof of Theorem 2.2. From Lemma 4.3, the vector function v^l, defined as in (4.5) with $\alpha^l = c^l$, satisfies an equation of the type (4.1). Therefore, Theorem 4.1 yields Theorem 2.2. □

Acknowledgments. The first author acknowledges an invitation by the University of Lyon I, where he was able to work in a stimulating atmosphere.

The second author is grateful to the CMAP for inviting him in June 1992.

REFERENCES

[BNS] H. BERESTYCKI, B. NICOLAENKO, AND B. SCHEURER, *Travelling wave solutions to combustion models and their singular limits,* SIAM J. Math. Anal., 16 (1985), pp. 1207–1242.

[B] A. BONNET, *Travelling waves for planar flames with complex chemistry reaction network*, Comm. Pure Appl. Math., 45 (1992), pp. 1271–1302.

[FT] M. FEINBERG AND D. TERMAN, *Travelling composition waves on isothermal catalyst surfaces*, Arch. Rational Mech. Anal., 116 (1991), pp. 35–69.

[FML] P. C. FIFE AND J. B. MCLEOD, *The approach of solutions of nonlinear diffusion equations by travelling front solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 335–361.

[G] R. GARDNER, *Existence and stability of travelling wave solutions of competition models: A degree theoretic approach*, J. Differential Equations, 44 (1982), pp. 335–379.

[R1] J.-M. ROQUEJOFFRE, *Convergence to travelling waves for solutions of a class of semilinear parabolic equations*, J. Differential Equations, 108 (1996), pp. 262–295.

[R2] ———, *Eventual monotonicity and convergence to travelling fronts for the solutions of parabolic equations in cylinders,* Ann. Inst. H. Poincaré Anal. Non Linéare, to appear.

[T1] D. TERMAN, *Stability of planar wave solutions to a combustion model*, SIAM J. Math. Anal., 21 (1990), pp. 1139–1171.

[T2] ———, *Traveling wave solutions arising from a two step combustion model*, SIAM J. Math. Anal., 19 (1988), pp. 1057–1080.

[VV1] A. I. VOLPERT AND V. A. VOLPERT, *Applications of the rotation theory of vector fields to the study of wave solutions of parabolic equations,* Trans. Moscow Math. Soc., 52 (1990), pp. 59–108.

[VV2] ———, *Travelling wave solutions of monotone parabolic systems*, to appear.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1270–1285, September 1996

005

TRAVELING-WAVE SOLUTIONS TO COMBUSTION MODELS FOR A REVERSIBLE REACTION*

ALEXIS BONNET†

Abstract. We prove the existence of traveling flames for a combustion model involving a reversible chemical reaction $\sum_{i=1}^{n} \nu_i A_i \rightleftharpoons \sum_{i=1}^{n} \mu_i A_i$. The proof of existence involves a degree theory argument and refined a priori estimates.

Key words. combustion, traveling waves, system of ordinary differential equations

AMS subject classifications. 34A34, 80A25, 80A30

1. Introduction. The study of planar flame is one of the fundamental problems in combustion (see, for example, [ZFK], [CFN], [JN], [W]). The rigorous mathematical treatment of this question was initiated by Berestycki, Nicolaenko, and Scheurer, who proved the existence of traveling waves for a simple combustion model [BNS]. They gave a complete analysis of the case of a one-step nonreversible reaction $A \longrightarrow B$ (see also [M] and [B3] for other qualitative properties and uniqueness results on this model). Later, Terman [T] and Heinze [H] addressed the question of complex chemistry. In their work, they considered only exothermic reactions. In [B1], we studied some complex chemical networks with the help of graph theory. We obtained the existence of traveling waves for complex chemical networks involving endothermic reactions. However, purely reversible reactions are not covered by these works.

In this paper, we consider the deflagration-wave problem for a compressible reacting gas with two or more species involved in a reversible chemical reaction.

$$A \rightleftharpoons B \tag{1.1}$$

or

$$\sum_{i=1}^{n} \nu_i A_i \rightleftharpoons \sum_{i=1}^{n} \mu_i A_i. \tag*{(1.1)$'$}$$

In the limit of small Mach number, the one-dimensional traveling-wave problem for the reversible chemical reaction (1.1) can be reduced to a system of two reaction-diffusion equations (see §2 for a review of the model)

$$\begin{cases} -T'' + cT' = -Y_1 f_1(T) + (1 - Y_1) f_2(T), \\ -d_1 Y_1'' + cY_1' = -Y_1 f_1(T) + (1 - Y_1) f_2(T) \end{cases} \tag{1.2}$$

with boundary conditions

$$\begin{cases} T(-\infty) = 0, \quad T(+\infty) = T^+, \\ Y_1(-\infty) = Y_1^-, \quad Y_1(+\infty) = Y_1^+. \end{cases} \tag{1.3}$$

*Received by the editors February 4, 1994; accepted for publication (in revised form) April 17, 1995. This work was partly done while the author was staying at the Courant Institute of Mathematical Sciences, New York University, and was supported by a Bourse Lavoisier of Ministère des Affaires Etrangères (France).

†Université de Cergy-Pointoise, 2 Avenue Adolphe Chauvin, 95302 Cergy-Pointoise cedex, France and Departement de Mathématiques et d'Informatique, École Normale Supérieure, 75230 Paris cedex 05, France.

The real c is the unknown mass flux of the wave.

In (1.2)–(1.3), T denotes the renormalized temperature of the mixture, Y_1 is the concentration of species A, and $1 - Y_1$ is the concentration of species B. Here we assume that A and B have the same diffusion coefficient d_1. For simplicity, d_1 is taken constant.

In system (1.2), $Y_1 f_1(T)$ is the rate of the semireaction $A \to B$ and $(1-Y_1)f_2(T)$ is the rate of the semireaction $B \to A$.

More precisely, we assume that the rates are of mass-action Arrhenius form

$$f_i = B_i(T)\exp\left(-\frac{E_i}{R(\alpha T + t_0)}\right), \quad i = 1, 2, \tag{1.4}$$

where E_i is the activation energy of the semireaction, $\alpha T + t_0$ is the absolute temperature of the mixture, and R a physical constant.

Notice that under conditions (1.3), the right-hand side of the equations in (1.2) vanishes for ($T = 0$, $Y_1 = Y_1^-$) and for ($T = T^+$, $Y_1 = Y_1^+$). Since we expect a flame to propagate, the mixture is not in equilibrium in the fresh gas ($T = 0$, $Y_1 = Y_1^-$). In other words, $-Y_1^- f_1(0) + (1 - Y_1^-)$ is not zero. However, this term is very small and the reaction is frozen because of the exponential dependance in temperature of the Arrhenius term. Thus the right-hand side of (1.2) is very small but not rigorously zero. This is what is refered as the cold-boundary difficulty (see [JN], [W], and [ZBLM]). Commonly, we introduce an ignition temperature $\theta > 0$ such that $B_i(T) = b_i\phi(T)\chi_{[\theta,+\infty)}$ for some $b_i > 0$. Under this hypothesis, one may expect system (1.2)–(1.3) to admit nontrivial solutions. A limit case is when $\theta = 0$ and ϕ is a smooth function such that $\phi(0) = 0$ and $\phi(T) > 0$ for $T > 0$. This hypothesis is used in some biological models and is known to yield to different qualitative results: the set of possible speed is no longer discrete (see [KPP]).

In (1.3), (T^+, Y_1^+) is the only pair (T, Y_1) with $T > \theta$ satisfying

$$\begin{cases} T = Y_1 - Y_1^-, \\ -Y_1 f_1(T) + (1 - Y_1)\, f_2(T) = 0. \end{cases} \tag{1.5}$$

We will prove the existence of traveling waves for the problem (1.2)–(1.3) in the cases $d_1 = 1$ (§3) and $d_1 \neq 1$ (§4). If $d_1 = 1$, the problem is reduced to a form to which the result of [BNS] in the scalar case applies.

The proof of existence given here for $d_1 \neq 1$ involves degree theory. The a priori estimates needed for the degree argument are obtained by mean of a backward-shooting method.

In §5, we discuss the limit $\theta \to 0$.

Finally, in §6, we extend our results to the more general chemical reaction (1.1)′. There, we have to solve a system of $n + 1$ reaction-diffusion equations

$$\begin{cases} -T'' + cT' = -\omega_1(T, Y) + \omega_2(T, Y), \\ -d_i Y_i'' + cY_i' = (\mu_i - \nu_i)\omega_1(T, Y) + (\nu_i - \mu_i)\omega_2(T, Y) \quad \text{for } i = 1, \dots, n, \end{cases} \tag{1.6}$$

where d_i is the diffusion coefficient of species A_i and ω_1 and ω_2 are the rates of the two semireactions. Our main theorem (Theorem 6.7 in §6) gives the existence of a flame for general kinetic rate ω_1 and ω_2 satisfying the monotonicity condition (6.7) in §6. This condition is satisfied in particular by the kinetic rates given by the law of mass action and the Arrhenius law. However, the result is not restricted to these laws.

2. The model for a reversible reaction $A \rightleftharpoons B$.

2.1. The model. When the hydrodynamic effects are neglected, the propagation of a deflagration flame is described by a thermodiffusive model. We are interested in the existence of traveling waves. These are flames propagating at a constant speed c. In a frame of reference moving at the same speed as the flame, the problem is written as a system of three reaction-diffusion equations of unknown (T, Y_1, Y_2, c),

$$\left\{\begin{aligned} -T'' + cT' &= Q_1 Y_1 f_1(T) + Q_2 Y_2 f_2(T), \\ -d_1 Y_1'' + cY_1' &= -Y_1 f_1(T) + Y_2 f_2(T), \\ -d_2 Y_2'' + cY_2' &= Y_1 f_1(T) - Y_2 f_2(T), \end{aligned}\right. \tag{2.1}$$

with boundary conditions in the fresh gas at $-\infty$,

$$T(-\infty) = 0, \quad Y_1(-\infty) = Y_1^-, \quad \text{and} \quad Y_2(-\infty) = Y_2^-, \tag{2.2}$$

where Y_1 (resp. Y_2) is the concentration of species A (resp. B). In (2.1), Q_1 and $Y_1 f_1(T)$ (resp. Q_2 and $Y_2 f_2(T)$) are the heat release and the rate of the semireaction $A \to B$ (resp. $B \to A$). As specified in the introduction, we assume that the rates are ruled by the mass-action law and the Arrhenius law:

$$f_i = B_i(T) \exp\left(-\frac{E_i}{R(\alpha T + t_0)}\right). \tag{2.3}$$

To resolve the cold-boundary problem, we assume that

$$B_i(T) = b_i\, \phi(T)\, \chi_{[\theta,+\infty)}, \tag{2.4}$$

where $b_i > 0$, $\theta > 0$ is the ignition temperature, and ϕ is some C^1 function with $\phi(0) = 0$ and $\phi(T) > 0$ on $(0, +\infty)$.

2.2. The diffusion coefficients. If we assume that A and B have the same diffusion coefficient, $d_1 = d_2$. Since Y_1 and Y_2 are bounded, an integration of the sum of the equations in Y_1 and Y_2 of system (2.1) gives the relation

$$Y_1 + Y_2 = Y_1^- + Y_2^-. \tag{2.5}$$

Assume $Y_1^- + Y_2^- = 1$; system (2.1) is then reduced to two reaction-diffusion equations as in (1.2):

$$\left\{\begin{aligned} -T'' + cT' &= -Y_1 f_1(T) + (1 - Y_1)\, f_2(T), \\ -d_1 Y_1'' + cY_1' &= -Y_1 f_1(T) + (1 - Y_1)\, f_2(T). \end{aligned}\right. \tag{2.6}$$

Remark 2.1. The coefficient d_1 is, in fact, the inverse of the Lewis number

$$Le = \frac{\lambda}{\rho C_p D}, \tag{2.7}$$

where λ is the thermal conductivity of the mixture, ρ is the density, C_p is the specific heat at constant pressure, and D is the diffusion coefficient of species A and B.

2.3. The heat release. We remark that by the first principle of thermodynamics, the heat release of the two semireactions are related:

$$Q_1 + Q_2 = 0.$$

The situation of an athermic reaction ($Q_1 = Q_2 = 0$) is not relevant in this model. If $Q_1 = Q_2 = 0$, the temperature will be a constant and system (1.2) reduces to $-d_1 Y_1'' + cY_1' = -Y_1 a_1 + (1 - Y_1)a_2$, where a_1 and a_2 are positive constants. This equation has no bounded solution except the constant $Y_1 \equiv a_2/(a_1 + a_2)$.

Without loss of generality, we may assume that Q_2 is positive and for a renormalized temperature, we may choose $Q_2 = 1$ and $Q_1 = -1$.

2.4. The boundary condition and the curve of chemical equilibrium. The boundary condition (2.2) prescribes only the concentrations and the temperature of the fresh gas. The composition and the temperature of the burnt gas are the unknowns of the problem. An integration of the difference of the two equations in (2.6) between $-\infty$ and $+\infty$ gives (if the limits exists at $+\infty$)

$$c(T(+\infty) - T(-\infty)) - c(Y_1(+\infty) - Y_1(-\infty)) = 0. \tag{2.8}$$

Since $T(-\infty) = 0$ and $Y_1(-\infty) = Y_1^-$, we get the necessary condition $T(+\infty) = Y_1(+\infty) - Y_1^-$. Moreover, if the limits T^+ and Y_1^+ of T and Y_1 exist at $+\infty$ for a solution of (1.2), then T^+ and Y^+ satisfy $-Y_1^+ f_1(T^+) + (1 - Y_1^+) f_2(T^+) = 0$. Therefore, Y_1^+ and T^+ are characterized by the set of equations

$$\begin{cases} T^+ = Y_1^+ - Y_1^-, \\ -Y_1^+ f_1(T^+) + (1 - Y_1^+) f_2(T^+) = 0. \end{cases} \tag{2.9}$$

We will now study the curve of chemical equilibrium defined by the second equation in (2.9). The approximation of activated states gives the relation between the activation energies and the heat release: $E_1 = E_2 + Q_2 = E_2 + 1$. Then, on the curve of chemical equilibrium, we define Y_1 as a function $Y_c(T)$ of T on $[\theta, +\infty)$:

$$Y_c(T) = \frac{1}{1 + \frac{b_1}{b_2} \exp\left(-\frac{1}{R(\alpha T + t_0)}\right)}. \tag{2.10}$$

By a straightforward calculation, $Y_c'(T)$ is negative and the equation $Y_c(T) = T + Y_1^-$ defines a unique T_0 and gives the existence and uniqueness of Y_1^+, $T^+ > 0$ satisfying (2.9) provided that $\theta < T_0$. In the following, the inequality $\theta < T_0$ will be a necessary and sufficient condition of existence of a traveling flame.

3. Lewis number equal to 1. When the Lewis number is equal to 1 (or equivalently $d_1 = 1$), system (1.2) reduces to a scalar equation. Indeed, the boundary conditions in (1.3) and the difference of the two equations of (1.2) give the relation $T = Y_1 - Y_1^-$. We are reduced to a scalar equation:

$$\begin{cases} -T'' + cT' = -(T + Y_1^-)f_1(T) + (1 - T - Y_1^-) f_2(T), \\ T(-\infty) = 0, \; T(0) = \theta. \end{cases} \tag{3.1}$$

The condition $T(0) = \theta$ removes the translational invariance. We may denote $k(T) = -(T+Y_1^-)f_1(T)+(1-T-Y_1^-)$. Then (3.1) is the scalar equation studied by Berestycki, Nicolaenko, and Scheurer, but here $k(T)$ is not positive everywhere:

$$\begin{cases} k(T) = 0 & \text{on} \quad [0, \theta], \\ k(T) > 0 & \text{on} \quad (\theta, T_0), \\ k(T) < 0 & \text{on} \quad (T_0, 1). \end{cases} \tag{3.2}$$

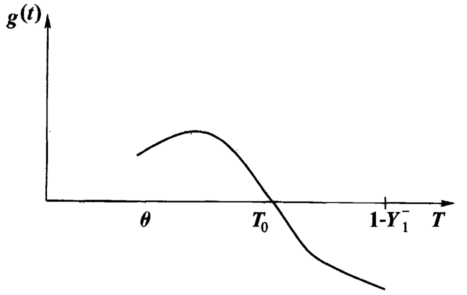


FIG. 1

The real T_0 is the unique solution of $Y_c(T) = T + Y_1^-$ (see identity (2.10)). (See Figure 1.)

Since $Y_2 = 1 - Y_1 = 1 - T - Y_1^-$ is the concentration of species B, it makes no physical sense to define g for $T > 1 - Y_1^-$. We have the following result.

THEOREM 3.1. *Under assumptions* (2.3) *and* (2.4), *if* $T_0 > \theta$, *then there exists a unique solution* $T : \mathbb{R} \to [0,1]$ *and* $c > 0$ *of problem* (3.1). *Moreover,* T *is of class* C^1, *it is of class* C^2 *on* $\mathbb{R} - \{0\}$, *and* $T(+\infty) = T_0$. *The condition* $T_0 > \theta$ *is necessary for the existence of a solution of* (3.1).

Proof. We will reduce the problem to the scalar case of [BNS]. For this purpose, let us prove the following lemma.

LEMMA 3.2. *For any bounded solution of* (3.1), *we have* $T(x) \leq T_0$ *on* $\mathbb{R}$.

COROLLARY 3.3. *If* $T_0 \leq \theta$, *problem* (3.1) *has no bounded solutions.*

Proof. For $T \geq T_0$, the inequality $-T'' + cT' < 0$ and the maximum principle imply that T cannot reach a local maximum at a point where $T > T_0$. Then if a solution (T, c) of (3.1) is such that T is larger than T_0 at some point x_0 in $\mathbb{R}$, T is increasing on $(x_0, +\infty)$. We have assumed that T was bounded, and we just saw that it was increasing on $(x_0, +\infty)$; therefore, T has a limit at $+\infty$: $T^+ > T_0$. Then, however, T^+ satisfies $k(T^+) = 0$, which implies $T^+ = T_0$—a contradiction. This completes the proof of Lemma 3.2.

A similar contradiction argument on a solution of (3.1) less than T_0 on $\mathbb{R}$ gives that for any bounded solution of (3.1), T has a limit at $+\infty$ and $T(+\infty) = T_0$.

To complete the proof of Theorem 3.1, we remark that since g is positive on (θ, T_0) and vanishes for $T = T_0$ and $T \in [0, \theta]$, the argument of Berestycki, Nicolaenko, and Scheurer gives the existence and uniqueness of a solution (T, c) of (3.1) with T increasing and $T(+\infty) = T_0$.

4. The reaction $A \rightleftharpoons B$ for Lewis number $\neq 1$.

4.1. Setting of the problem.

When the Lewis number $\neq 1$ (or equivalently $d_1 \neq 1$), the equations with unknown (T, Y, c) are

$$\left\{ \begin{array}{l} -T'' + cT' = -Y_1 f_1(T) + (1 - Y_1) f_2(T), \\ -d_1 Y_1'' + cY_1' = -Y_1 f_1(T) + (1 - Y_1) f_2(T), \\ T(-\infty) = 0, \ Y_1(-\infty) = Y_1^-. \end{array} \right. \tag{4.1}$$

We assume that the rate functions f_1 and f_2 satisfy assumptions (2.3) and (2.4).

To prove the existence of some solution (T, Y, c) of problem (4.1), we follow the same proof sketch as in [BNS]. On a truncated domain $[-a, +a]$, we prove the existence of solutions of the differential equations in (4.1) with adapted boundary condition (zero flux at $-a$). The existence is obtained by a degree argument for a fixed-point problem. The difficulty is to find some a priori bounds for T, Y_1, T', Y_1', and c. Then by passing to the limit $a = +\infty$, we prove the existence of a solution of (4.1).

4.2. Truncated problem on $\overline{I_a} = [-a, +a]$. Let T^+ and Y^+ be the temperature and concentration of the burnt gas as defined in (2.9). (We implicitly assume that $T_0 > \theta$.) We study the problem on $[-a, +a]$:

$$(4.2)\qquad \begin{cases} -T'' + cT' = -Y_1 f_1(T) + (1 - Y_1)\, f_2(T), \\ -d_1 Y_1'' + cY_1' = -Y_1 f_1(T) + (1 - Y_1)\, f_2(T), \\ -T'(-a) + cT(-a) = 0, \quad T(+a) = T^+, \quad T(0) = \theta, \\ -d_1 Y_1'(-a) + cY_1(-a) = cY_1^-, \quad Y_1(+a) = Y^+. \end{cases}$$

The condition $T(0) = \theta$ removes the translational invariance of the problem.

The proof of existence of a solution (T, Y_1, c) for problem (4.2) will involve degree theory. Let $X = C^1(\overline{I_a}) \times C^1(\overline{I_a}) \times I\!R$; X is a Banach space equipped with the norm

$$\|T, Y_1, c\|_X = \max\left(\|T\|_{C^1(\overline{I_a})}, \|Y_1\|_{C^1(\overline{I_a})}, |c|\right).$$

For $0 \le \tau \le 1$, we consider the mapping that sends (t, y, c) of X onto the unique solution (T, Y) of the linear system

$$(4.3)\qquad \begin{cases} -T'' + cT' = -\tau y f_1(t) + \tau(1 - y)\, f_2(t), \\ -d_1 Y'' + cY' = -\tau y f_1(t) + \tau(1 - y)\, f_2(t), \\ -T'(-a) + cT(-a) = 0, \quad T(+a) = T^+, \\ -d_1 Y'(-a) + cY(-a) = cY_1^-, \quad Y(+a) = Y^+ \end{cases}$$

and define $K_\tau : X \to X$, $(t, y, c) \mapsto (T, Y, c - T(0) + \theta)$. Then (t, y, c) is a solution of (4.2) if and only if (t, y, c) is a fixed point of K_1.

Remark 4.1. For the definition of f_1 and f_2, we take a smooth approximation χ_ϵ of $\chi_{[\theta,+\infty)}$ to ensure the existence and uniqueness of the solution of (4.3):

$$(4.4)\qquad \begin{cases} \chi_\epsilon = 0 \quad \text{on } [0, \theta], \quad \chi_\epsilon = 1 \quad \text{on } [\theta + \epsilon, +\infty), \\ f_i(T) = b_i \chi_\epsilon(T) \exp\left(-\dfrac{E_i}{R(\alpha T + t_0)}\right), \quad i = 1, 2. \end{cases}$$

To prove the existence of a solution of (4.2), we will compute the degree of $F_\tau = I - K_\tau$ at 0. For this purpose, let us consider an open subset Ω of X:

$$(4.5)\qquad \Omega = \left\{(t, y, c) \in X, \ \|t\|_{C^1(\overline{I_0})} < M, \ \|y\|_{C^1(\overline{I_0})} < M, \ \underline{c} < c < \overline{c}\right\}.$$

If for every τ, $F_\tau(\partial\Omega) \not\ni 0$, then $\deg(F_1, \Omega, 0) = \deg(F_0, \Omega, 0)$ by the homotopy property of the degree.

Step 1: *Construction of* Ω *such that* $F_\tau(\partial\Omega) \not\ni 0$. The main result here is as follows.

PROPOSITION 4.2. *There exist positive constants* M, $\underline{c}$, *and* $\overline{c}$ *such that* $F_\tau(\partial\Omega) \not\ni 0$ *for every* τ *in* $[0, 1]$.

The existence of M is a consequence of the following proposition.

PROPOSITION 4.3. *If* (T, Y, c) *is such that* $F_\tau(T, Y, c) = 0$, *then* Y *and* T *are strictly increasing:* $Y' > 0$ *and* $T' > 0$ *on* $[-a, a]$.

The proof will involve the following lemmas.

LEMMA 4.4. *Let* (T, Y, c) *satisfy* $F_\tau(T, Y, c) = 0$. *If* $T(x) = \theta$, *then* $x = 0$.

Proof. If the lemma is not true, there is at least one point x_0 where $T(x_0) = \theta$ and $T'(x_0) \leq 0$. Then, however, by the uniqueness of the solution of differential equations, we should have

$$\begin{cases} T(x) = \theta + \dfrac{T'(x_0)}{c}\left(e^{c(x-x_0)} - 1\right), \\ \text{and } Y(x) = Y(x_0) + \dfrac{Y'(x_0)}{c}\left(e^{c(x-x_0)} - 1\right) \quad \text{for } x \geq x_0, \end{cases} \tag{4.6}$$

which leads to a contradiction with the boundary condition at $+a$.

In conclusion, for $x > 0$, we must have $T > \theta$.

COROLLARY 4.5. *On* $[-a, 0]$, T *and* Y *are explicitly given by*

$$\begin{cases} T(x) = \theta\, e^{cx}, \\ Y(x) = Y_1^- + b\, e^{cx}; \end{cases} \tag{4.7}$$

moreover, $Y(0) > 0$.

Indeed, if $Y(0) \leq 0$, then $Y'(0) \leq 0$. We remark that $-Y f_1(T) + (1-Y) f_2(T) \leq 0$ for $Y \leq 0$. Therefore, the maximum principle implies that Y remains nonpositive on $[-a, a]$, which contradicts the boundary conditions at $+a$.

For simplicity of notation, we introduce the function

$$g(T, Y) = -Y f_1(T) + (1 - Y) f_2(T)$$

and the sets

$$\begin{cases} U^+ = \{(T,Y) \in I\!R^2,\ T > \theta,\ g(T,Y) \geq 0\}, \\ U^- = \{(T,Y) \in I\!R^2,\ T > \theta,\ g(T,Y) \leq 0\}, \\ \Gamma_e = U^+ \cap U^-. \end{cases} \tag{4.8}$$

The set Γ_e is the equilibrium curve for $T > \theta$ of the chemical reaction $A \rightleftharpoons B$. (See Figure 2.)

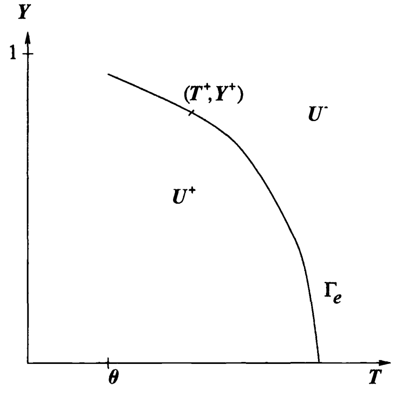


FIG. 2

LEMMA 4.6. *For* T, Y, *and* c *as in Proposition* 4.3, *we have*

$$T'(+a) > 0 \quad \text{and} \quad Y'(+a) > 0.$$

Proof. By integration of (4.3), we get $T'(+a) = Y'(+a)$. The discussion is then as follows:

(i) If $T'(+a) = Y'(+a) = 0$, then $T(x) = T^+$ and $Y(x) = Y^+$ is a solution of the two differential equations in (4.3). This is actually the only solution that satisfies $T(+a) = T^+$, $Y(+a) = Y^+$, and $T'(+a) = Y'(+a) = 0$. This leads to a contradiction with the boundary conditions at $-a$.

(ii) If $T'(+a) < 0$ and $Y'(+a) < 0$, the boundary conditions at $-a$ imply that T' and Y' cannot remain negative on $[-a, +a]$. Let x_0 be the largest x in $[-a, +a]$ where either T' or Y' is nonnegative (we have $x_0 < a$). Take, for instance, $T'(x_0) \geq 0$. Since T is C^1, we get $T'(x_0) = 0$. Now notice that if $(p, q) \in U^-$, then

$$\{(p', q') \in I\!R^2, p < p', q < q'\} \subset \overset{o}{U}{}^-,$$

where $\overset{o}{U}{}^-$ denotes the interior of U^-. By the definition of x_0, we have $T(x_0) > T(+a)$ and $Y(x_0) > Y(+a)$; therefore, $(T(x_0), Y(x_0)) \in \overset{o}{U}{}^-$, which leads to $g(T(x_0), Y(x_0)) < 0$. Then we remark that $-T''(x_0) + cT'(x_0) = g(T(x_0), Y(x_0)) < 0$. Since $T'(x_0) = 0$, we get $T''(x_0) > 0$ and consequently $T'(x) > 0$ on an interval $(x_0, x_0 + \epsilon)$ for some $\epsilon > 0$, which contradicts the maximality of x_0.

A similar contradiction is achieved if we assume that $Y'(x_0) \geq 0$. This completes the proof of Lemma 4.6.

We are now ready to prove Proposition 4.3 by contradiction. Indeed, if Proposition 4.3 is not true, there is a maximum x_0 where either $T'(x_0) \leq 0$ or $Y'(x_0) \leq 0$. Corollary 4.5 implies that $x_0 > 0$ and $T(x_0) > \theta$. Then the same argument as in Lemma 4.6 with the set U^+ in place of U^- gives a contradiction.

The proof of Proposition 4.3 is now complete.

Let us now go back to Proposition 4.2. Since T and Y are increasing and $Y_c(T)$ is decreasing, (T, Y) remains in U^+ on $[-a, a]$ and, by integration of (4.3), we get

$$-T' + cT \geq 0 \quad \text{and} \quad -d_1 Y' + cY \geq 0. \tag{4.9}$$

We have thus completed the proof of the following proposition.

PROPOSITION 4.7. *If* $F_\tau(T, Y, c) = 0$, *then* $0 < T \leq T^+$, $0 < Y \leq Y^+$, *and* $0 < T' \leq cT^+$, $0 < Y' \leq \frac{cY^+}{d_1}$.

The bound M in the definition (4.5) of Ω will be found as soon as c is bounded from above.

We get an upper bound for c independent of τ and $a > a_0$. As in [BNS], we have the following result.

PROPOSITION 4.8. *Let* $M = \sup_{\substack{\theta \leq t \leq T^+ \\ 0 \leq y \leq Y^+}} -yf_1(t) + (1-y)f_2(t)$ *and let* a_0 *be fixed. For every* $a \geq a_0$, *if* $F_\tau(T, Y, c) = 0$, *then* c *satisfies*

$$c \leq \max\left(-\frac{\log\theta}{a_0},\ \max\left(2M, \frac{\sqrt{2M}}{\theta}\right)\right). \tag{4.10}$$

Let us now find a lower bound for c that is independent of τ. The inequalities in (4.9) and the fact that $T(0) = \theta$ and $T'(0) = c\theta$ lead easily to $T(x) \leq \theta\, e^{cx}$ on $[0, +a]$ by Gronwall's lemma. As $T(+a) = T^+$, this gives

$$T^+ \leq \theta\, e^{ca} \tag{4.11}$$

and the following proposition.

PROPOSITION 4.9. *If* $F_\tau(T, Y, c) = 0$, *then* $c \geq \frac{1}{a}\log\frac{T^+}{\theta}$.

This lower bound of c is independent of τ but dependent on a. For the passage to the limit $a = +\infty$, we will need a lower bound on c independent of a, which will be developed in §4.3.

Propositions 4.7, 4.8, and 4.9 give the bounds M, $\underline{c}$, and $\overline{c}$ of Proposition 4.1. This completes the construction of the open-bounded set Ω with $F_\tau(\partial\Omega) \not\ni 0$ for $\tau \in [0,1]$. Moreover, $M = \max\left(Y^+, \overline{c}Y^+, \overline{c}\,\frac{Y^+}{d_1}\right)$ since $T^+ = Y^+ - Y_1^-$.

Step 2: *Justification of the degree.* As in [BNS], it is easy to verify that

$$\deg(F_0, \Omega, 0) = -1. \tag{4.12}$$

Indeed, for $\tau = 0$, we explicitly compute $K_0(t,y,c) = (T, Y, c - T(0) + \theta)$. We obtain $T(x) = T^+\ e^{c(x-a)}$, $Y(x) = T^+ e^{\frac{c}{d_1}(x-a)} + Y_1^-$, and $c - T(0) + \theta = c - T^+e^{-ca} + \theta$. Then $F_0(t,y,c) = 0$ if and only if $c = \frac{1}{a}\log\frac{T+}{\theta}$, $t(x) = T^+\ e^{c(x-a)}$, and $y(x) = T^+\ e^{\frac{c}{d_1}(x-a)} + Y_1^-$. The mapping K_0 is homotopic to $\Phi : X \to X$, $(t,y,c) \mapsto (T^+\ e^{c(x-a)}, T^+\ e^{\frac{c}{d_1}(x-a)} + Y_1^-, c - T^+e^{-ca} + \theta)$. The multiplicative property of the degree gives $\deg(Id - \Phi) = -1$ (notice that $c \mapsto T^+e^{-ca}$ is decreasing).

In conclusion, we have proved the following theorem.

THEOREM 4.10. *Under assumption* (4.4), *problem* (4.2) *on* $[-a, a]$ *has at least one solution* (T, Y, c) *in* X. *Moreover, there exist* $\underline{c}$, $\overline{c}$ $(0 < \underline{c} < \overline{c})$, *and* $M(\overline{c})$ *such that* $\|T\|_{\mathcal{C}^1(\overline{I_a})} < M(\overline{c})$ *and* $\|Y\|_{\mathcal{C}^1(\overline{I_a})} < M(\overline{c})$.

Let $\epsilon \to 0$ in (4.4). The existence theorem and the bounds above remain true for $f_i(T) = b_i\chi_{[\theta,+\infty)} \exp\left(-\frac{E_i}{R(\alpha T + t_0)}\right)$ and $f_i(T) = b_i\Phi(x)\,\chi_{[\theta,+\infty)} \exp\left(-\frac{E_i}{R(\alpha T + t_0)}\right)$.

4.3. The passage to the limit $a = +\infty$**.** For the passage to the limit, we need estimates independent of a. Proposition 4.8 gives an upper bound $\overline{c}$ of c that is independent of $a > a_0$. Let us now find a lower bound of c that is independent of a.

Step 1: *Lower bound of c independent of a.* An integration of the equation in T of system (4.2) gives

$$-T'(+a) + cT(+a) = \int_{-a}^{a} g(T,Y) = \int_{0}^{a} g(T,Y). \tag{4.13}$$

As $0 < T' \le \frac{T^+}{c}$, identity (4.13) gives

$$cT^+ > \int_0^a g(T,Y)\,dx.$$

We estimate the integral

$$\int_0^a g(T,Y)\,dx = \int_\theta^{T^+} \frac{g(t, Y(T^{-1}(t)))}{T'(T^{-1}(t))}\,dt$$
$$\ge \frac{1}{cT^+}\int_\theta^{T^+} g(t, Y(T^{-1}(t)))\,dt.$$

The two inequalities above give

$$(cT^+)^2 > \int_0^a g(t, Y(T^{-1}(t)))\,dt. \tag{4.14}$$

Since $g(t, Y(T^{-1}(t)\,)) > 0$, we only need to obtain a lower bound of this function on a finite interval of t. Let us remark that for $\epsilon < \frac{T^+ - \theta}{2}$, we have $g(Y,T) > 0$ on

$[\theta+\epsilon, T^+-\epsilon]\times[0\le Y\le Y^+]\subset \overset{o}{U}^-$. Then, by compactness, there exists $\delta_\epsilon>0$ such that

$$g(T,Y)>\delta_\epsilon \quad \text{on} \quad [\theta+\epsilon, T^+-\epsilon]\times[0\le Y\le Y^+].$$

For a solution (T,Y,c) of (4.2), we know that $0\le Y\le Y^+$. Then for $\epsilon_0=(T^+-\theta)/4$, we have

$$(cT^+)^2>\epsilon_0\delta_0 \quad \text{and} \quad c>c_0=\frac{\sqrt{\epsilon_0\delta_{\epsilon_0}}}{T^+}. \tag{4.15}$$

The real $c_0=\sqrt{\epsilon_0\delta_{\epsilon_0}}/T^+>0$ is a lower bound of c independent of a.

Step 2: *Passage to the limit* $a=+\infty$. Using the a priori estimate of §4.1 and the lower bound c_0 of c obtained above, we will prove the existence of a solution of problem (4.1).

THEOREM 4.11. *Assume that $\theta<T_0$. Assume that one of the conditions* (4.4) *or* (2.4) *on f_1 and f_2 is satisfied. Then there exists an increasing sequence $\{a_n\}_{n\in\mathbb{N}}$ with* $\lim a_n=+\infty$ *such that $(T_{a_n},Y_{a_n},c_{a_n})$ is a solution of* (4.2) *on $(-a_n,+a_n)$ and converges in $C^1_{\text{loc}}(\mathbb{R})\times C^1_{\text{loc}}(\mathbb{R})\times\mathbb{R}$ to a solution (T,Y,c) of* (4.1). *Moreover, (T,Y,c) satisfies*

$$\begin{gathered}T(+\infty)=T^+, \qquad Y(+\infty)=Y^+,\\ 0\le T\le T^+, \qquad 0\le Y\le Y^+,\\ 0\le T'\le cT^+, \qquad 0\le Y'\le\frac{cY^+}{d_1},\\ Y,T\in W^{2,\infty}(\mathbb{R}),\\ 0<c_0\le c\le\bar{c}<+\infty.\end{gathered}\tag{4.16}$$

The condition $T_0>\theta$ is necessary for the existence of a nontrivial bounded solution of (4.1).

Proof. The arguments are the same as in [BNS]. We use the bounds of T, Y, c, T', Y', and g to get a bound for T'' and Y'' and to prove the local convergence to a solution of the differential system in (4.1). The boundary conditions at $-\infty$ are obviously satisfied. For the limit at $+\infty$, we notice that T and Y are increasing and bounded. Their limits exist and we have $Y'(+\infty)=Y''(+\infty)=T'(+\infty)=T''(+\infty)=0$ and consequently $g(T(+\infty),Y(+\infty))=0$ with $T(+\infty)>\theta$. By integration of (4.1), $T(+\infty)=Y(+\infty)-Y_1^-$ and thus $T(+\infty)=T^+$ and $Y(+\infty)=Y^+$. The inequalities in (4.16) are straightforward.

Finally, we prove the following lemma.

LEMMA 4.12. *If $T_0\le\theta$ then* (4.1) *has no bounded nontrivial solutions.*

Proof. By contradiction, assume that $T_0\le\theta$ and let (T,Y_1,c) be a nontrivial bounded solution of (4.1). In other words, we assume that T and Y_1 are not constant. Then $c>0$ and there is a point x_0 where $T(x_0)=\theta$. We may assume that $x_0=0$ and that $T<\theta$ on $(-\infty,0)$. Therefore, on $(-\infty,0)$ we explicitly have $T=\theta\ e^{cx}$.

(i) If T is increasing on $\mathbb{R}$, then it has a limit $T^+>\theta$ at $+\infty$ since we assumed that T is bounded. Consequently, $-T'+cT$, which is also increasing, converges to cT^+ at $+\infty$. However, we know that $-T'+cT=-d_1Y_1'+c(Y_1-Y_1^-)$. Thus $-d_1Y_1'+c(Y_1-Y_1^-)$ converges to cT^+ at $+\infty$. This implies that $Y_1\to Y^+=T^++Y_1^-$ as $x\to+\infty$. Then $g(T^+,Y^+)=0$, and this gives $T^+=T_0$—a contradiction since $T_0\le\theta<T^+$

(ii) If T is not increasing on $\mathbb{R}$, it has a local maximum x_0, $T(x_0)>\theta$. At x_0, we have $-T''(x_0)+cT'(x_0)\ge0$, that is, $(T(x_0),Y_1(x_0))\in U^+$ (or, equivalently, $Y_1(x_0)\le Y_c(T(x_0))$). Since $T_0\le\theta$ and $Y_c(T)$ is decreasing, we have $T(x_0)>Y_c(T(x_0))-Y_1^-$ and we deduce that $d_1Y_1'(x_0)=c(Y_1(x_0)-Y_1^--T(x_0))<0$. Therefore, in a right neighborhood $(x_0, x_0+\epsilon)$ of x_0, we have $(T(x), Y_1(x)) \in U^+$. However, we know that in U^+, $-T''+cT' \geq 0$ and $-d_1 Y_1'' + cY_1' \geq 0$. Moreover, we know that if $(t,y) \in U^+$, then for all T and Y such that $T \leq t$ and $Y \leq y$, $g(T,Y) \geq 0$. The same contradiction argument as in Lemma 4.6 gives that if at a point x_1 we have $T'(x_1) < 0$, $Y_1'(x_1) < 0$ and $(T(x_1), Y_1(x_1)) \in U^+$, then $T'(x) < 0$, $Y_1'(x) < 0$, and $g(T(x), Y_1(x)) \geq 0$ for all $x \geq x_1$. Moreover, we have $T'' \leq cT' < 0$ and then $T \to -\infty$ as $x \to +\infty$, which contradicts the assumption: T and Y_1 are bounded. This completes the proof of the lemma and of Theorem 4.11.

5. The limit $\theta \to 0$. In §§3 and 4, we proved the existence of traveling waves for the problem with ignition temperature. Here we will study the limit $\theta \to 0$. This section uses the ideas of Marion [M], who studied the case of a direct reaction $A \to B$.

We assume here that f_1 and f_2 are given by $f_i(T) = b_i\,\phi(T)\exp\big(-\frac{E_i}{R(\alpha T+t_0)}\big)$, where $\phi(T)$ is chosen such that $\phi(0) = 0$, $\phi(T) > 0$ on $(0,+\infty)$ and ϕ is locally Lipschitz.

We consider the rate functions f_1^θ and f_2^θ defined as $f_i^\theta(T) = \chi_{[\theta,+\infty)}(T) f_1(T)$.

We will first prove that for $\theta \to 0$, translates of solutions of (4.1) for the rate functions f_i^θ converge to a solution (T, Y_1, c) of

$$\begin{cases} -T''+cT' = -Y_1 f_1(T) + (1-Y_1)\,f_2(T), \\ -d_1 Y_1'' + cY_1^1 = -Y_1 f_1(T) + (1-Y_1)\,f_2(T), \\ T(-\infty) = 0, \quad T(+\infty) = T^+, \\ Y_1(-\infty) = Y_1^-, \quad Y_1(+\infty) = Y^+. \end{cases} \tag{5.1}$$

In §4, we found a lower bound c_0 of c that is independent of θ for $\theta < \theta_0$ for some fixed $\theta_0 < T^+$. However, the upper bound of c and thus the bounds for Y' and T' found in §4 are not independent of θ (see (4.10)).

Let $(T_\theta, Y_\theta, c_\theta)$ be a solution of (4.1); since T_θ is increasing, we may define $h_\theta(s) = T'_\theta(T_\theta^{-1}(s))$ in the same way as in [M]. We have

$$\begin{aligned} h_\theta(T^+) &= h_\theta(0) = 0, \\ h'_\theta(s) &= c_\theta - \frac{g^\theta(s, Y_\theta(T_\theta^{-1}(s))}{h_\theta(s)}. \end{aligned} \tag{5.2}$$

For $x = 0$, notice that $T = \theta$ and $T' = c\theta$—that is to say, $h_\theta(\theta) = c_\theta\theta$. Then, since $h_\theta(T^+) = 0$, there exists $\theta_1 \in (\theta, T^+)$ such that

$$\forall s \in [\theta, \theta_1], \quad h_\theta(s) \geq \frac{c_\theta}{2}\, s \quad \text{and} \quad h_\theta(\theta_1) = \frac{c_\theta}{2}\,\theta_1. \tag{5.3}$$

Then $h'_\theta(\theta_1) \leq \frac{c_\theta}{2}$ and

$$\frac{c_\theta}{2} \geq h'_\theta(\theta_1) \geq c_\theta - \frac{g(\theta_1, Y(T_\theta^{-1}(\theta_1))}{\frac{c_\theta}{2}\,\theta_1}, \tag{5.4}$$

which gives

$$\left(\frac{c_\theta}{2}\right)^2 \leq \frac{g(\theta_1, Y(T_\theta^{-1}(\theta_1))}{\theta_1} \leq \frac{f_2(\theta_1)}{\theta_1}. \tag{5.5}$$

However, $\frac{f_2^\theta(\theta_1)}{\theta_1} = b_2 \frac{\phi(\theta_1)}{\theta_1} \exp\left(-\frac{E_2}{R(\alpha\theta_1+t_0)}\right)$ is bounded on $(0, T^+]$ since $\phi(0) = 0$ and since ϕ is locally Lipschitz. We have an upper bound of c independent of θ:

$$c_\theta \leq 2\sqrt{\sup_{s\in(0,T^+]} \frac{f_2(s)}{s}}. \tag{5.6}$$

By the same argument as in [M], it is easy to prove the following theorem.

THEOREM 5.1. *For each $\theta < \frac{T^+}{2}$, let z_θ be such that $T(z_\theta) = \frac{T^+}{2}$. There is a sequence $\{\theta_n\}_{n\in N}$, $0 < \theta_n < \frac{T^+}{2}$, $\lim \theta_n = 0$ such that $(T_{\theta_n}(x - z_{\theta_n}), Y_{\theta_n}(x - z_{\theta_n}), c_{\theta_n})$ converges to a solution (T, Y, c) of* (5.1).

Moreover, we can prove the following result easily as in [M].

THEOREM 5.2. *There exist $0 < \underline{c} < \bar{c}$ such that* (5.1) *has no solution (T, Y_1) for c in $[0, \underline{c}]$ and* (5.1) *has a solution for c in $[\bar{c}, +\infty)$. If $d_1 \leq 1$, then $\underline{c} = \bar{c}$.*

6. General case of reversible reaction.

6.1. Setting of the problem.

In §4, we made some special assumptions to reduce the problem to the study of a system of two reaction-diffusion equations. In this section, we address the general case of a reversible reaction involving n species A_i:

$$\sum_{i=1}^{n} \nu_i A_i \rightleftharpoons \sum_{i=1}^{n} \mu_i A_i. \tag{6.1}$$

The nonnegative constants ν_i and μ_i are the stoichiometric coefficients of the reaction. We may have $\nu_i > 0$ and $\mu_i > 0$ at the same time since that is the case in the branching step of chemical-reaction networks. For reasons explained in Remark 6.1 below, we will assume that $\nu_i \neq \mu_i$.

We will call the forward reaction $\sum_{i=1}^n \nu_i A_i \to \sum_{i=1}^n \mu_i A_i$ semireaction 1 and the backward reaction $\sum_{i=1}^n \mu_i A_i \to \sum_{i=1}^n \nu_i A_i$ semireaction 2. Without loss of generality, we assume that semireaction 2 is exothermic with $Q_2 = 1$ the heat released. The equations with unknown (T, Y, c) for the propagation of a planar traveling wave are

$$\begin{cases} -T'' + cT' = -\omega_1(T, Y) + \omega_2(T, Y), \\ -d_i Y_i'' + cY_i' = (\mu_i - \nu_i)\omega_1(T, Y) + (\nu_i - \mu_i)\omega_2(T, Y) \ \text{ for } i = 1, \ldots, n, \end{cases} \tag{6.2}$$

where Y denotes the vector $(Y_1, \ldots, Y_n)$. This system can be written as

$$\begin{cases} -T'' + cT' = -\omega_1(T, Y) + \omega_2(T, Y), \\ -d_i Y_i'' + cY_i' = (\nu_i - \mu_i)\big(-\omega_1(T, Y) + \omega_2(T, Y)\big) \ \text{ for } i = 1, \ldots, n. \end{cases} \tag{6.3}$$

We prescribe the concentrations of the species in the fresh gas at $-\infty$:

$$T(-\infty) = 0, \ Y_1(-\infty) = Y_1^-, \ldots, \ Y_n(-\infty) = Y_n^-. \tag{6.4}$$

In (6.3), the constants d_i are the diffusion coefficients of the chemical species A_i. Since they are not supposed to be equal to 1, we cannot reduce (6.3) to one equation as in §3. Moreover, we do not assume that the d_i's are equal to each other. This situation occurs, for example, in diluted reacting gases.

The usual expression of the rates of the semireactions, ω_1 and ω_2, is given by the mass-action and Arrhenius laws:

$$\omega_1(T,Y) = \prod_{\nu_i>0} Y_i^{\nu_i} B_1(T) \exp\left(-\frac{E_1}{R(\alpha T + t_0)}\right), \tag{6.5}$$
$$\omega_2(T,Y) = \prod_{\mu_i>0} Y_i^{\mu_i} B_2(T) \exp\left(-\frac{E_2}{R(\alpha T + t_0)}\right),$$

where E_j is the activation energy of each semireaction and $\alpha T + t_0$ is the absolute temperature of the mixture. As we explained in §2, the activation energies and the heats released, Q_1 and Q_2, are related to each other: $Q_1 + Q_2 = 0$ and $E_1 = E_2 + Q_2$. Above, we assume that semireaction 2 is exothermic ($Q_2 = 1$), and, consequently, we have $E_1 > E_2$. The cutoff function B_j is introduced to resolve the cold-boundary difficulty:

$$B_j(T) = b_j \chi_{[\theta,+\infty)}, \tag{6.6}$$

where b_j is a positive constant and $\theta > 0$ is the ignition temperature. In (6.6), the indicator function $\chi_{[\theta,+\infty)}$ can be replaced by a smooth approximation $\chi^{\epsilon}_{[\theta,+\infty)}$ as in (4.4).

In fact, the proof of existence applies to some more general functions ω_1 and ω_2. The only assumption we need is the Lipschitz continuity of ω_1 and ω_2 for $T > \theta$ and the following property:

$$\begin{cases} \text{for } T > \theta, \ \dfrac{\omega_1}{\omega_2} \text{ is increasing with respect to the } n+1 \text{ uplet} \\ (T, (\mu_1 - \nu_1)Y_1, (\nu_2 - \mu_2)Y_2, \ldots, (\nu_n - \mu_n)Y_n). \end{cases} \tag{6.7}$$

This means that when $T, (\nu_1 - \mu_1)Y_1, (\nu_2 - \mu_2)Y_2, \ldots, (\nu_n - \mu_n)Y_n$ are simultaneously increased, the value of $\frac{\omega_1}{\omega_2}(T,Y)$ is increased. Physically, this means that the rate functions $\omega_i \geq 0$ satisfy the following condition:

$$\text{for } T > \theta, \quad \begin{cases} \omega_1(T,Y) = 0 \Longleftrightarrow \exists i, & \nu_i > 0 \quad \text{and} \quad Y_i = 0, \\ \omega_2(T,Y) = 0 \Longleftrightarrow \exists i, & \mu_i > 0 \quad \text{and} \quad Y_i = 0. \end{cases} \tag{6.8}$$

This condition reads as follows: the rate of the reaction vanishes if the concentration of one of the reactants vanishes; the rate of the reaction is positive if the concentrations of all the reactants are positive.

Remark 6.1. Notice here that if for some i_0, $\mu_{i_0} = \nu_{i_0}$, then the chemical species A_{i_0} is globally neither consumed nor produced by the reversible reaction. Therefore, Y_{i_0} is a constant. Then if $Y_{i_0}^- = 0$, no reaction occurs ($\omega_1 = \omega_2 \equiv 0$). If $Y_{i_0}^- > 0$, the concentration Y_{i_0} appears as a constant in ω_1 and ω_2 and the equation in Y_{i_0} can be removed from system (6.3). Therefore, without loss of generality, we assume that $\mu_i \neq \nu_i$ for all i.

LEMMA 6.2. *The rate functions ω_1 and ω_2 defined by identities* (6.5) *and* (6.6) *satisfy properties* (6.7) *and* (6.8).

Proof. This is straightforward since $E_1 = E_2 + Q_2 = E_2 + 1$.

We are now interested in the composition of the burnt gas. Formally, the burnt gas should be in a state of chemical equilibrium. If the limits of Y_i and T at $+\infty$ exist, then an integration between $-\infty$ and $+\infty$ of linear combinations of equations in (6.3) gives the n relations

$$Y_i(+\infty) - Y_i(-\infty) = (\nu_i - \mu_i)(T(+\infty) - T(-\infty)) \quad \text{for } i = 1, \ldots, n.$$

Thus, with the boundary conditions (6.4), we get

$$Y_i(+\infty) = Y_i^- + (\nu_i - \mu_i)T(+\infty) \quad \text{for } i = 1, \dots, n. \tag{6.9}$$

Denote by Y_i^+ the right side of (6.9) and define $T^+ = T(+\infty)$. The reals Y_i^+ depend linearly on T^+. If the limit exists at $+\infty$, then the right-hand side of equation (6.3) is null, which gives a necessary condition on T^+, $Y^+ = (Y_1^+, \dots, Y_n^+)$:

$$\omega_1(T^+, Y^+) - \omega_2(T^+, Y^+) = 0. \tag{6.10}$$

Since Y^+ is a function of T^+, (6.10) gives a condition on T^+ that we can write for $T^+ > \theta$ as

$$\frac{\omega_1}{\omega_2}(T^+, Y^+(T^+)) = 1. \tag{6.11}$$

Also, we remark here that $(\nu_i - \mu_i)Y_i^+(T^+)$ is an increasing function of T^+. Therefore, Lemma 6.2 (or condition (6.7)) gives that $\frac{\omega_1}{\omega_2}(T^+, Y^+(T^+))$ is an increasing function of T^+. This gives the following proposition.

PROPOSITION 6.3. *There is a real $T^+ > \theta$ satisfying* (6.11) *if and only if*

$$\lim_{s \searrow \theta} \frac{\omega_1}{\omega_2}(s, Y^+(s)) < 1.$$

Moreover, T^+ is unique.

To prove the existence of a traveling-wave solution of (6.3) and (6.4), we will proceed as in §4.

6.2. Truncated problem on $\overline{I_a} = [-a, +a]$. We now study the problem on $[-a, +a]$.

$$\begin{cases} -T'' + cT' = -\omega_1(T, Y) + \omega_2(T, Y), \\ -d_iY_i'' + cY_i' = (\nu_i - \mu_i)\big(-\omega_1(T, Y) + \omega_2(T, Y)\big) \quad \text{for } i = 1, \dots, n \\ -T'(-a) + cT(-a) = 0, \quad T(+a) = T^+, \quad T(0) = \theta, \\ -d_iY_i'(-a) + cY_i(-a) = cY_i^-, \quad Y_i(+a) = Y_i^+ \quad \text{for } i = 1, \dots, n. \end{cases} \tag{6.12}$$

We now prove the following theorem.

THEOREM 6.4. *System* (6.2) *admits a solution (T^a, Y^a, c^a) on $[-a, +a]$, and there are positive constants $\underline{c}$, $\bar{c}$, and M independent of $a > a_0$ such that $||T^a||_{C^1(\overline{I_a})} < M$, $||Y_i^a||_{C^1(\overline{I_a})} < M$, and $\underline{c} < c < \bar{c}$.*

Proof. The proof of existence of a solution (T, Y, c) for problem (6.12) will involve degree theory as in §3. Since the proof follows the same sketch, we will detail only the specific results needed here.

6.2.1. A priori estimates and the monotonicity of T and Y_i. Some a priori estimates are needed for degree theory (see §4.1). These a priori estimates are easily derived as soon as T and Y_i are proved to be monotonic. The monotonicity of T and Y_i is obtained by the same kind of argument as in the proof of Proposition 4.3. Indeed, we first present the following lemma.

LEMMA 6.5. *For (T, Y, c), a solution of* (6.12), *we have*

$$T'(+a) > 0 \quad \textit{and} \quad (\nu_i - \mu_i)Y_i'(+a) > 0. \tag{6.13}$$

Proof. Indeed, by integration of (6.12), we get

$$Y_i'(+a) = (\nu_i - \mu_i)T'(+a). \tag{6.14}$$

Then the argument is the same as in Lemma 4.6. We introduce the sets U^+ and U^- associated with the function $g(T, Y) = -\omega_1(T, Y) + \omega_2(T, Y)$ as in (4.8). Identity (6.14) associated with Lemma 6.2 allows us to argue by contradiction on $T, Y_1, \ldots, Y_n$ exactly as we did in Lemma 4.6 with only T and Y.

With the same assumptions as in Proposition 4.3, we prove the following.

PROPOSITION 6.6. *T is increasing; Y_i is increasing if $\nu_i - \mu_i$ is positive and decreasing if $\nu_i - \mu_i$ is negative.*

With Proposition 6.6, we are able to derive the a priori estimates on T, Y, and c and get the existence of a traveling wave on a bounded domain.

6.3. Conclusion. The passage to the limit $a = +\infty$ is the same as in §4, and we can state the following theorem.

THEOREM 6.7. *Assume that $Q_2 > 0$ and that conditions* (6.7) *and* (6.8) *on ω_1 and ω_2 hold. If*

$$\lim_{s \searrow \theta} \frac{\omega_1}{\omega_2}(s, Y^+(s)) < 1, \tag{6.15}$$

then problem (6.3)–(6.4) *has at least a nontrivial solution (T, Y, c) in $\left(W^{2,\infty}(I\!R)\right)^{n+1} \times I\!R$. Moreover, T and Y_i are monotonic and satisfy*

$$\begin{cases} T(+\infty) = T^+, \\ Y_i(+\infty) = Y_i^+(T^+) = Y_i^- + (\nu_i - \mu_i)T^+. \end{cases} \tag{6.16}$$

REFERENCES

[BL] H. BERESTYCKI AND B. LARROUTUROU, *Quelques aspects mathématiques de la propagation des flammes prémélangées*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar 10, H. Brezis and J.-L. Lions, eds., Pitman–Longman, Harlow, UK, 1990, pp. 65–129.

[BLN] H. BERESTYCKI, B. LARROUTUROU, AND L. NIRENBERG, *A nonlinear elliptic problem describing the propagation of a curved premixed flame*, Mathematical Modeling in Combustion and Related Topics, C.-M. Brauner and C. Schmidt-Lainé, eds., NATO ASI Series E, vol. 140, Martinus Nijhoff, Dordrecht, the Netherlands, 1988.

[BNS] H. BERESTYCKI, B. NICOLAENKO, AND B. SCHEURER, *Traveling wave solutions to combustion models and their singular limits*, SIAM J. Math. Anal., 16 (1985), pp. 1207–1242.

[Be] C. BERGE, *Graphes et Hypergraphes*, Dunod, Paris, 1970.

[B1] A. BONNET, *Travelling waves for flames with complex chemistry reactions network*, Comm. Pure Appl. Math., XV (1992), pp. 1269–1302.

[B2] A. BONNET, *Propagation of flames in the limit of zero ignition temperature*, Arch. Rational Mech. Anal., to appear.

[B3] ———, *Non unicité dans le modèle de propagation de flamme plane quand le nombre de Lewis est inférieur à* 1*/Non-uniqueness for planar flame propagation model when Lewis number is less than* 1, C. R. Acad. Sci. Paris Sér. II Méc. Phys. Chim. Sci. Univers. Sci. Terve, 315 (1992), pp. 421–426.

[CFN] P. CLAVIN, P. C. FIFE, AND B. NICOLAENKO, *Multiplicity and related phenomena in competing reaction flames*, SIAM J. Appl. Math., 47 (1987), pp. 296–331.

[H] S. HEINZE, *Traveling waves in combustion processes with complex chemical networks*, Trans. Amer. Math. Soc., 304 (1987), pp. 405–416.

[JN] W. E. JOHNSON AND W. NACHBAR, *Laminar flame theory and the steady linear burning of a monopropellant*, Arch. Rational Mech. Anal., 12 (1963), pp. 58–91.

[J] W. E. JOHNSON, *On a first order boundary value problem for laminar flame theory*, Arch. Rational Mech. Anal., 13 (1963), pp. 46–54.

[K1] J. I. KANEL, *Stabilization of solutions of the Cauchy problem for equations encountered in combustion theory*, Mat. Sb., 59 (1962), pp. 245–288.

[K2] ———, *On steady state solutions to systems of equations arising in combustion theory*, Dokl. Akad. Nauk UzSSR, 149 (1963), pp. 367–369.

[KPP] A. KOLMOGOROFF, I. PETROVSKY, AND N. PISCOUNOFF, *Study of the diffusion equation with growth of the quantity of matter and its application to a biology problem*, Byul. Moskov. Gos. Univ. Matem., 17 (1937), pp. 1–26.

[M] M. MARION, *Qualitative properties of a nonlinear system for laminar flames without ignition temperatures*, Nonlinear Anal., 9 (1985), pp. 1269–1292.

[S] G. I. SIVASHINSKY, *Instabilities, pattern formation and turbulence in flames*, Ann. Rev. Fluid Mech., 15 (1983), pp. 179–199.

[T] D. TERMAN, *Traveling wave solutions arising from a two-step combustion model*, SIAM J. Math. Anal., 19 (1988), pp. 1057–1080.

[W] F. WILLIAMS, *Combustion Theory*, Addison–Wesley, Reading, MA, 1983.

[ZBLM] Y. B. ZELDOVICH, G. I. BARENBLATT, V. B. LIBOVICH, AND G. M. MAHVILADZE, *Mathematical Theory of Combustion and Detonation*, Nauka, Moscow, 1985.

[ZFK] Y. B. ZELDOVICH AND D. A. FRANK-KAMENETSKII, *A theory of thermal propagation of flame*, Acta Phys. Chim., 2 (1938), p. 341.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1286–1310, September 1996

006

TRAVELING-WAVE SOLUTIONS OF CONVECTION-DIFFUSION SYSTEMS IN NONCONSERVATION FORM*

LIONEL SAINSAULIEU†

Abstract. Hyperbolic systems in nonconservation form are found in several domains of mathematical physics, but the definitions of their shockwave solutions rely on the definition of the product of a Heavyside-type function with Dirac-type distribution. For systems in nonconservation form extracted from a convection-diffusion system, we prove a conjecture of Le Floch. This relies on the construction of traveling-wave solutions of a second-order system.

Key words. hyperbolic systems in nonconservation form, convection-diffusion systems, shockwaves, traveling waves, Lax entropy conditions

AMS subject classifications. 35A07, 35B45, 35J15

1. Introduction.

Many models used in fluid mechanics are written in the following form (in one-dimensional slab geometry):

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x} - \frac{\partial}{\partial x}\left(\mathbf{D}(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x}\right) = 0, \tag{1.1}$$

where $\mathbf{A}$ and $\mathbf{D}$ are two $p \times p$ C^2 matrix-valued functions defined on a subset Ω of $\mathbb{R}^p$. (See, for instance, [1].) When the matrix-valued function $\mathbf{A}$ is written in the form $\mathbf{A}(\mathbf{u}) = \mathbf{df}(\mathbf{u})$ for some flux function $\mathbf{f}$, we say that system (1.1) is in conservation form. However, it is now well established that some systems modeling two-phase fluid flows (see, for instance, [2]–[4]) or nonlinear elasticity (see [5]) are written in the form of a hyperbolic system in nonconservation form.

In order to avoid considering the small-scale effects connected with the diffusion phenomena or as the first step of a numerical method, consider the following first-order system extracted from (1.1):

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x} = 0. \tag{1.2}$$

Usually, the matrix $\mathbf{A}(\mathbf{u})$ has real eigenvalues so that system (1.2) is hyperbolic. However, systems in nonconservation form present a number of unconventional mathematical features, and this leads to interesting new mathematical problems. Indeed, in the same manner as for hyperbolic systems in conservation form, we expect the formation of shockwaves in the solutions of (1.2), even with smooth initial data. But when $\mathbf{u} : \mathbb{R}_+ \times \mathbb{R} \to \Omega$ is a discontinuous function, the following product in nonconservation form,

$$\mathbf{A}(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x}, \tag{1.3}$$

lacks meaning as a distribution and the usual theory of hyperbolic systems of conservation laws does not apply. In fact, many authors consider hyperbolic systems in

*Received by the editors March 4, 1994; accepted for publication (in revised form) May 5, 1995.

†CERMICS, École Nationale des Ponts et Chaussees, Central 2, La Courtine, 93167 Noisy-le-Grand cedex, France.

nonconservation form to be valuable models, which explains the great amount of work recently performed to give meaning to the product (1.3) when $\mathbf{u}$ is a discontinuous function.

The first author who gave meaning to (1.3) when $\mathbf{u}$ is the step function

$$\mathbf{u}(x,t) = \begin{cases} \mathbf{u}^L, & x < \sigma t, \\ \mathbf{u}^R, & x > \sigma t, \end{cases} \tag{1.4}$$

was probably Volpert (see [6]). His definition is, to some extent, the simplest one:

$$\mathbf{A}(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x} = \frac{1}{2}\left(\mathbf{A}(\mathbf{u}^L) + \mathbf{A}(\mathbf{u}^R)\right)(\mathbf{u}^R - \mathbf{u}^L)\delta_{x=\sigma t}, \tag{1.5}$$

where δ denotes Dirac's delta function. The authors of [7] noticed that this definition is not satisfactory since it depends on the system of dependent variables $\mathbf{u}$. They generalized Volpert's definition by replacing the right-hand side of (1.5) with an average along some given path, namely,

$$\mathbf{A}(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x} = \left(\int_0^1 \mathbf{A}(\phi(s;\mathbf{u}^L,\mathbf{u}^R))\frac{\partial \phi}{\partial s}(s;\mathbf{u}^L,\mathbf{u}^R)s\right)\delta_{x=\sigma t}, \tag{1.6}$$

where the function $\phi : [0,1] \times \Omega \times \Omega \to \Omega$ is Lipschitz continuous and satisfies

$$\phi(0;\mathbf{u}^L,\mathbf{u}^R) = \mathbf{u}^L, \quad \phi(1;\mathbf{u}^L,\mathbf{u}^R) = \mathbf{u}^R. \tag{1.7}$$

This definition is independent of the system of dependent variables, but it relies on the choice of the function ϕ. When the matrix $\mathbf{A}$ is the derivative of a flux function $\mathbf{f}$, the definition (1.6) is consistent with the theory of distributions. The right-hand side of (1.6) is independent of the choice of ϕ. On the contrary, when there is no flux function $\mathbf{f}$ such that $\mathbf{A} = \mathbf{df}$, the definition (1.6) indeed depends on ϕ. Given a family of paths ϕ, the authors give a solution of the Riemann problem

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x} = 0, \quad \mathbf{u}(x,0) = \begin{cases} \mathbf{u}^L, & x < 0, \\ \mathbf{u}^R, & x > 0 \end{cases}$$

provided that $|\mathbf{u}^R - \mathbf{u}^L|$ is small enough.

In [8]–[9], the authors describe their own mathematical definition of the product in (1.3). This definition relies only on the matrix $\mathbf{A}$: some equations in (1.2) are considered in a strong sense and others in a weak sense (see [8] or [9] for more details).

In our opinion, the definition of a shockwave solution of a system in nonconservation form should depend on the dissipation mechanisms. Following [10], we define a shockwave solution of system (1.2) extracted from the second-order system (1.1) as the limit, when diffusion is neglected, of a traveling-wave solution of (1.1). It is important to notice that the points of view in [7] and [8] are very close to this definition. Indeed, compare the "microscopic structure" of a shockwave in the work of Colombeau with the path ϕ that connects the left and right states of a shockwave in the work of Dal Maso, Le Floch, and Murat, or a shockwave solution of (1.2) viewed as the limit, when diffusion is neglected, of a solution of (1.1). In each case, some extra information is added to the shockwave in the form of a continuous profile that connects the left and right states of the shock. When a second-order system in the form of (1.1) is a relevent model, our approach offers a simple definition of the shockwave solution of (1.2). However, a convenient choice of the function ϕ in the framework of Dal Maso, Le Floch, and Murat or a convenient treatment of system (1.2) using Colombeau's calculus gives the same shockwaves.

A continuous traveling wave $\mathbf{u}$ with speed σ that connects states $\mathbf{u}^L$ and $\mathbf{u}^R$ is a continuous bounded function in the form

$$\mathbf{u}(x,t) = \hat{\mathbf{u}}(x - \sigma t), \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+, \tag{1.8}$$

with the following limits:

$$\lim_{x \to -\infty} \mathbf{u}(x,t) = \mathbf{u}^L, \qquad \lim_{x \to +\infty} \mathbf{u}(x,t) = \mathbf{u}^R. \tag{1.9}$$

The function $\mathbf{u}$ is a solution of (1.1) if $\hat{\mathbf{u}}$ is a solution of the following system of differential equations:

$$-\sigma \hat{\mathbf{u}}' + \mathbf{A}(\hat{\mathbf{u}})\hat{\mathbf{u}}' - (\mathbf{D}(\hat{\mathbf{u}})\hat{\mathbf{u}}')' = 0, \tag{1.10}$$

where $'$ denotes derivation with respect to x. Next, set

$$\hat{\mathbf{u}}_v(x) = \hat{\mathbf{u}}(x/v), \quad x \in \mathbb{R}.$$

Then the function $\hat{\mathbf{u}}_v$ is a solution of the differential system

$$-\sigma \hat{\mathbf{u}}_v' + \mathbf{A}(\hat{\mathbf{u}}_v)\hat{\mathbf{u}}_v' - v\,(\mathbf{D}(\hat{\mathbf{u}}_v)\hat{\mathbf{u}}_v') = 0 \tag{1.11$_v$}$$

and has the same limits as $\hat{\mathbf{u}}$ as ξ tends to $\pm\infty$. Furthermore, the family $(\hat{\mathbf{u}}_v')_{v>0}$ is bounded in $L^1(\mathbb{R})$ and, as ν tends to zero, tends almost everywhere to the following discontinuous function:

$$\hat{\mathbf{u}}_0(x) = \begin{cases} \mathbf{u}^R & \text{if } x > 0, \\ \mathbf{u}^L & \text{if } x < 0. \end{cases} \tag{1.12}$$

Following [10] and [11], we say that the function (1.12) is a shockwave solution of (1.2) compatible with the diffusion matrix $\mathbf{D}$. Unlike the case of hyperbolic systems of conservation laws, the function $\hat{\mathbf{u}}_0$ indeed depends on the shape of the diffusion matrix.

The mathematical justification of this definition relies on the existence of traveling-wave solutions of (1.1). A particular case is considered in [11]. A general existence result was conjectured in [10]. This paper is devoted to its rigorous proof for a general class of systems of the form of (1.1).

There are several papers dealing with traveling-wave solutions of second-order systems in conservation form. First, a general result may be found in [12]. The authors assume that the diffusion matrix is regular, but this assumption is usually not satisfied by fluid-flow models. To our knowledge, no general result concerning second-order systems with a singular diffusion matrix has yet been published. Several systems arising from fluid mechanics have been considered. In [13], the author proves the existence of traveling-wave solutions of the ZND detonation model. Setting the chemical reaction terms to zero then gives the existence of traveling-wave solutions of the system of Navier–Stokes equations. Two-dimensional viscous isentropic flow equations are considered in [14].

To study the traveling-wave solutions of a second-order system in conservation form, one integrates (1.10) (recall that, by assumption, $\mathbf{A}(\mathbf{u}) = \mathbf{df}(\mathbf{u})$). In [12] and [15], the authors write (1.10) in the form of a dynamical system. The traveling-wave solutions of (1.1) are then obtained using the powerful tools of dynamical-systems theory. Here system (1.1) is in nonconservation form and this method does not apply. In fact, we prove that a solution of (1.1) is in the form $\mathbf{u}(x) = \mathbf{u}^L + u(x)\mathbf{f}_0 + O(\epsilon^2)$, where $\mathbf{f}_0$ is a constant vector and the function u is solution of the following model equation:

$$-\frac{\epsilon}{2}u' + \left(\frac{u^2}{2}\right)' - u'' = 0, \quad u(-\infty) = 0, \quad u(+\infty) = \epsilon < 0.$$

The function u is written

$$u(x) = \frac{\epsilon}{1 + \exp\left(\epsilon\frac{x}{2}\right)}.$$

Section 2 gives our hypotheses and summarizes the results obtained in this paper. In §3, we prove the existence of traveling-wave solutions of (1.1) when the diffusion matrix $\mathbf{D}$ is the identity matrix. Using the results of §3, we deduce in §4 the existence of traveling-wave solutions of (1.1) for a wide class of diffusion matrices. Section 5 is devoted to the proof of a priori estimates of the solutions of (1.1), from which we deduce the uniqueness of the solutions constructed in §4. Finally some technical results concerning a nonlinear differential equation are given in the appendix.

2. Hypotheses and main results. We suppose that system (1.2) is strictly hyperbolic. For any $\mathbf{u} \in \Omega$, the matrix $\mathbf{A}(\mathbf{u})$ has p distinct real eigenvalues

$$\lambda_1(\mathbf{u}) < \cdots < \lambda_p(\mathbf{u}).$$

We denote by $\{\mathbf{r}_k(\mathbf{u})\}_{1\leq k\leq p}$ (resp., $\{\mathbf{l}_k(\mathbf{u})\}_{1\leq k\leq p}$) the right (resp., left) eigenvectors of the matrix $\mathbf{A}(\mathbf{u})$.

Recall that a characteristic field $\mathbf{r}_i$ is genuinely nonlinear (GNL) if

$$\forall\mathbf{u} \in \Omega, \quad \nabla\lambda_i(\mathbf{u}) \cdot \mathbf{r}_i(\mathbf{u}) \neq 0$$

and linearly degenerate (LD) if

$$\forall\mathbf{u} \in \Omega, \quad \nabla\lambda_i(\mathbf{u}) \cdot \mathbf{r}_i(\mathbf{u}) = 0.$$

(The theory of hyperbolic systems of conservation laws is examined in [15] and [16].) We suppose that each field $\mathbf{r}_k$, $1 \leq k \leq p$, is either LD or GNL. We normalize the GNL fields by

$$\forall\mathbf{u} \in \Omega, \quad \nabla\lambda_i(\mathbf{u}) \cdot \mathbf{r}_i(\mathbf{u}) = 1. \tag{2.1}$$

The vectors $\mathbf{l}_k$, $1 \leq k \leq p$, are normalized by

$$\mathbf{l}_k(\mathbf{u}) \cdot \mathbf{r}_l(\mathbf{u}) = \delta_{kl}.$$

We assume that the first r equations in (1.1) contain no diffusion terms and that only the last $q = p - r$ equations contain diffusion terms. This means that the matrix $\mathbf{D}(\mathbf{u})$ is of the form

$$\mathbf{D}(\mathbf{u}) = \begin{pmatrix} 0 & 0 \\ \mathbf{D}_1(\mathbf{u}) & \mathbf{D}_2(\mathbf{u}) \end{pmatrix}, \tag{2.2}$$

where $\mathbf{D}_1(\mathbf{u})$ is a $q \times r$ matrix and $\mathbf{D}_2(\mathbf{u})$ is a $q \times q$ matrix. Note that the diffusion matrix in a one-dimensional model of fluid flow is usually of the form of (2.2). For instance, the total-mass-conservation equation does not contain any diffusion term. Then we write the matrix $\mathbf{A}(\mathbf{u})$ in the following form:

$$\mathbf{A}(\mathbf{u}) = \begin{pmatrix} \mathbf{A}_1(\mathbf{u}) & \mathbf{A}_2(\mathbf{u}) \\ \mathbf{A}_3(\mathbf{u}) & \mathbf{A}_4(\mathbf{u}) \end{pmatrix}. \tag{2.3}$$

Here $\mathbf{A}_1(\mathbf{u})$ is a $r \times r$ matrix, $\mathbf{A}_2(\mathbf{u})$ is a $r \times q$ matrix, $\mathbf{A}_3(\mathbf{u})$ is a $q \times r$ matrix, and $\mathbf{A}_4(\mathbf{u})$ is a $q \times q$ matrix.

For $\mathbf{u}^L$ given in Ω and an index i such that $\mathbf{r}_i$ is GNL, we obtain below a half-curve of states that can be connected to $\mathbf{u}^L$ by a traveling-wave solution of (1.1). This half-curve is tangent in $\mathbf{u}^L$ to $\mathbf{r}_i(\mathbf{u}^L)$. However, we need the following two conditions to be satisfied:

$$\text{The matrix } \mathbf{A}_1(\mathbf{u}) - \sigma 1_r \text{ is regular.} \tag{2.4}$$

$$\ker \mathbf{D}(\mathbf{u}) \cap \left(\mathbf{A}(\mathbf{u}) - \sigma \mathbf{1}_p\right)^{-1} \text{range} \mathbf{D}(\mathbf{u}) = \{\mathbf{0}\}. \tag{2.5}$$

In fact, when one of these conditions is not satisfied, the only possible traveling-wave solutions of (1.1) have some discontinuous components (cf. [17]).

When $(\lambda_i(\mathbf{u}^L), \mathbf{u}^L)$ satisfies (2.4) and (2.5), we can find a positive number $\hat{\epsilon}$ such that for any $\mathbf{u} \in \Omega$ and $\sigma \in \mathbb{R}$ with

$$|\mathbf{u} - \mathbf{u}^L| \leq \hat{\epsilon}, \quad |\sigma - \lambda_i(\mathbf{u}^L)| \leq \hat{\epsilon}, \tag{2.6}$$

the matrix $\mathbf{A}_1(\mathbf{u}) - \sigma 1_r$ is regular and $\ker \mathbf{D}(\mathbf{u}) \cap \left(\mathbf{A}(\mathbf{u}) - \sigma \mathbf{1}_p\right)^{-1} \text{range} \mathbf{D}(\mathbf{u}) = \{\mathbf{0}\}$.

Next, we assume that for any GNL field $\mathbf{r}_i$,

$$\mathbf{l}_i(\mathbf{u}) \cdot \mathbf{D}(\mathbf{u}) \mathbf{r}_i(\mathbf{u}) \neq 0 \quad \forall \mathbf{u} \in \Omega. \tag{2.7}$$

This assumption may be found in [12] and is essential with regard to the stability of the wave.

A natural function space for studying (1.1) is the following:

$$W = \{\mathbf{u} \in L^\infty(\mathbb{R}), \ \mathbf{u}' \in L^1(\mathbb{R}), \ \mathbf{u}(-\infty) = \mathbf{0}\}.$$

With the norm

$$\|\mathbf{u}\|_W = \|\mathbf{u}'\|_{L^1(\mathbb{R})},$$

W is a Banach space.

The existence result is stated as follows.

THEOREM 2.1. *Consider a second-order system* (1.1) *whose diffusion matrix is of the form of* (2.2). *Let* $\mathbf{u}^L$ *belong to* Ω *and let* $i \in \{1, \ldots, p\}$ *such that* $\mathbf{r}_i$ *is GNL. Assume that the pair* $(\lambda_i(\mathbf{u}^L), \mathbf{u}^L)$ *satisfies* (2.4) *and* (2.5). *Then we can find a positive number* ϵ_0 *such that for any* $\sigma \in \left]\lambda_i(\mathbf{u}^L) - \epsilon_0/2, \lambda_i(\mathbf{u}^L)\right[$, *system* (1.1) *has a traveling-wave solution* $\mathbf{u}_\sigma : (x,t) \in \mathbb{R}_+ \times \mathbb{R} \to \mathbf{u}_\sigma(x - \sigma t)$ *with speed* σ *and* $\mathbf{u}_\sigma(-\infty) = \mathbf{u}^L$. *Furthermore,*

$$\mathbf{u}_\sigma(+\infty) = \mathbf{u}^L + 2(\sigma - \lambda_i(\mathbf{u}^L))\mathbf{r}_i(\mathbf{u}^L) + O\left(\left(\sigma - \lambda_i(\mathbf{u}^L)\right)^2\right). \tag{2.8}$$

These solutions are unique up to a translation. Indeed, we prove the following.

THEOREM 2.2. *Assume that all the characteristic fields $\mathbf{r}_i$, $1 \le i \le p$, are GNL. Let $\mathbf{u}^L \in \Omega$ and $\sigma \in \mathbb{R}$ such that the pair $(\sigma, \mathbf{u}^L)$ satisfies* (2.4) *and* (2.5). *We can find two positive numbers $\hat{c}$ and ϵ_1 such that if $\mathbf{v} : (x,t) \to \mathbf{v}(x - \sigma t)$ is a traveling-wave solution of* (1.1) *with speed σ and $\mathbf{v}(-\infty) = \mathbf{u}^L$ that satisfies*

$$\begin{gathered} \epsilon = |\mathbf{v}(+\infty) - \mathbf{u}^L| \le \epsilon_1, \\ \|\mathbf{v} - \mathbf{u}^L\|_W \le \hat{c}\epsilon, \end{gathered} \tag{2.9}$$

then $\lambda_i(\mathbf{u}^L) - \epsilon_0/2 < \sigma < \lambda_i(\mathbf{u}^L)$ and

$$\mathbf{v}(x) = \mathbf{u}_\sigma(x + a), \quad x \in \mathbb{R},$$

for some real number a.

3. Existence in the case of the identity-diffusion matrix. To begin, we prove Theorem 2.1 when the matrix $\mathbf{D}$ is the identity matrix. The proof is much simpler in this case, and the general case can be reduced to this simpler case, which is done in §4 below.

We are thus concerned with the construction of traveling-wave solutions of the following system:

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x} - \frac{\partial^2 \mathbf{u}}{\partial x^2} = 0. \tag{3.1}$$

We assume that the first-order system extracted from (3.1),

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x} = 0, \tag{3.2}$$

is strictly hyperbolic. Conditions (2.4) and (2.5) are obviously satisfied by any pair $(\sigma, \mathbf{u}) \in \mathbb{R} \times \Omega$ since $r = 0$ and $\ker \mathbf{D}(\mathbf{u}) = \{\mathbf{0}\}$. Then let $\mathbf{u}^L \in \Omega$ be fixed and choose an index i such that the field $\mathbf{r}_i$ is GNL.

As proved in §1, finding traveling-wave solutions of (3.1) amounts to finding bounded solutions of the following system of differential equations:

$$-\sigma \mathbf{u}' + \mathbf{A}(\mathbf{u})\mathbf{u}' - \mathbf{u}'' = 0, \tag{3.3}$$

where σ denotes the speed of the traveling wave. A function $\mathbf{u}$ with $\mathbf{u}(-\infty) = \mathbf{u}^L$ is a solution of (3.3) if

$$(\mathbf{A}(\mathbf{u}^L) - \sigma\mathbf{Id})(\mathbf{u}(x) - \mathbf{u}^L) - \mathbf{u}'(x) = \Phi(\mathbf{u})(x) = \int_{-\infty}^{x} \big(\mathbf{A}(\mathbf{u}^L) - \mathbf{A}(\mathbf{u}(s))\big)\mathbf{u}'(s)s. \tag{3.4}$$

We expect that the states that can be connected to a given left state $\mathbf{u}^L$ by a traveling-wave solution of (3.3) belong to some half-curves tangent in $\mathbf{u}^L$ to one of the right eigenvectors of $\mathbf{A}(\mathbf{u}^L)$. Next, we expect that the velocity of the traveling wave is close to the corresponding eigenvalue. For a given $\epsilon < 0$, we are thus led to look for solutions of (3.3) in the set $\Sigma_i^\epsilon \times \mathcal{K}_i^\epsilon$ defined by

$$\Sigma_i^\epsilon = \left\{\sigma \in \mathbb{R}, \ \left|\lambda_i(\mathbf{u}^L) - \sigma + \frac{\epsilon}{2}\right| \le c_0\epsilon^2\right\}, \tag{3.5}$$

(3.6)
$$\mathcal{K}_i^\epsilon = \left\{ \mathbf{u} \in \mathbf{u}^L + W;\ \|\mathbf{u}-\mathbf{u}^L\|_W \le 2|\epsilon||\mathbf{r}_i(\mathbf{u}^L)|,\ \left\|\mathbf{l}_k(\mathbf{u}^L)\cdot\left(\mathbf{u}-\mathbf{u}^L\right)\right\|_W \le c_0\epsilon^2,\ k \neq i \right\},$$

where c_0 is chosen a priori.

The function $\Phi(\mathbf{u})$ has the following behavior in $\mathcal{K}_i^\epsilon$.

LEMMA 3.1. *Let* $\mathbf{u}^1$ *and* $\mathbf{u}^2$ *belong to* $\mathcal{K}_i^\epsilon$. *Then* $\Phi(\mathbf{u}^j)$, $j = 1, 2$, *belongs to* W *and*

$$\|\Phi(\mathbf{u}^j)\|_W \le c_1\epsilon^2, \quad j = 1,\ 2, \qquad \|\Phi(\mathbf{u}^1)-\Phi(\mathbf{u}^2)\|_W \le c_1|\epsilon|\|\mathbf{u}^1-\mathbf{u}^2\|_W, \tag{3.7}$$

where c_1 *is independent of* c_0.

Here and below, C denotes a generic a priori constant which varies from relation to relation. We use C whenever the constants have no special significance in the proof.

LEMMA 3.2. *Let* $\mathbf{u}^1$ *and* $\mathbf{u}^2$ *belong to* $\mathcal{K}_i^\epsilon$. *Set* $u_i^j(x) = \mathbf{l}_i(\mathbf{u}^L)\cdot\left(\mathbf{u}^j(x)-\mathbf{u}^L\right)$, $j = 1, 2$. *Then*

$$\mathbf{l}_i(\mathbf{u}^L)\cdot\Phi(\mathbf{u}^j)(x) = \frac{(u_i^j(x))^2}{2} + \theta_i(\mathbf{u}^j)(x) \tag{3.8}$$

for some function θ_i *that satisfies*

$$\|\theta_i(\mathbf{u}^j)\|_W \le C|\epsilon|^3, \quad j = 1,\ 2, \tag{3.9a}$$

$$\|\theta_i(\mathbf{u}^1)-\theta_i(\mathbf{u}^2)\|_W \le C\left(\epsilon^2\|u_i^1-u_i^2\|_W + |\epsilon|\sum_{k\neq i}\left\|\mathbf{l}_k(\mathbf{u}^L)\cdot\left(\mathbf{u}^1-\mathbf{u}^2\right)\right\|_W\right). \tag{3.9b}$$

Proof. The proof of Lemma 3.1 is straightforward. Let us prove Lemma 3.2. Let be given $\mathbf{u}$ in $\mathcal{K}_i^\epsilon$. Set

$$u_i(x) = \mathbf{l}_i(\mathbf{u}^L)\cdot\left(\mathbf{u}(x)-\mathbf{u}^L\right).$$

Since $\mathbf{A}$ is a C^2 matrix-valued function, Taylor's formula gives

$$\mathbf{l}_i(\mathbf{u}^L)\cdot\left((\mathbf{A}(\mathbf{u}^L)-\mathbf{A}(\mathbf{u}))\mathbf{u}'(x)\right) = \mathbf{l}_i(\mathbf{u}^L)\cdot\left[\left(\mathbf{A}(\mathbf{u}^L)\cdot(\mathbf{u}^L-\mathbf{u}(x))\right)\mathbf{u}'(x)\right]+\beta_i(x), \tag{3.10a}$$

where the function β_i satisfies

$$|\beta_i(x)| \le C|\mathbf{u}(x)-\mathbf{u}^L|^2|\mathbf{u}'(x)|. \tag{3.10b}$$

However, the function $\mathbf{u}$ belongs to $\mathcal{K}_i^\epsilon$, and we write

$$\begin{aligned}\mathbf{l}_i(\mathbf{u}^L)&\cdot\left((\mathbf{A}(\mathbf{u}^L)-\mathbf{A}(\mathbf{u}))\mathbf{u}'(x)\right)\\ &= -\mathbf{l}_i(\mathbf{u}^L)\cdot\left(\mathbf{A}(\mathbf{u}^L)\cdot\left(\mathbf{r}_i(\mathbf{u}^L),\mathbf{r}_i(\mathbf{u}^L)\right)\right)u_i(x)u_i'(x)+\zeta_i(x),\end{aligned} \tag{3.11a}$$

where the function ζ_i is given by

$$\begin{aligned}\zeta_i(\mathbf{u}^j)(x) = {}&\beta_i(\mathbf{u}^j)(x)+\mathbf{l}_i(\mathbf{u}^L)\cdot\left\{\mathbf{A}(\mathbf{u}^L)\cdot\left[\mathbf{u}^L-\mathbf{u}+u_i(x)\mathbf{r}_i(\mathbf{u}^L)\right]\right)\mathbf{u}'(x)\right\}\\ &+u_i(x)\mathbf{l}_i(\mathbf{u}^L)\cdot\left\{\left[\mathbf{A}(\mathbf{u}^L)\cdot\mathbf{r}_i(\mathbf{u}^L)\right]\left[\mathbf{u}'(x)-u_i'(x)\mathbf{r}_i(\mathbf{u}^L)\right]\right\}.\end{aligned} \tag{3.11b}$$

Since the function $\mathbf{u}$ belongs to $\mathcal{K}_i^\epsilon$, we have the following estimate:

$$\|\zeta_i\|_{L^1(\mathbb{R})} \le C|\epsilon|^3. \tag{3.11c}$$

On the other hand, for $\mathbf{v} \in \Omega$, $(\lambda_i(\mathbf{v}), \mathbf{r}_i(\mathbf{v}))$ is an eigenpair of $\mathbf{A}(\mathbf{v})$:

$$\mathbf{A}(\mathbf{v})\mathbf{r}_i(\mathbf{v}) = \lambda_i(\mathbf{v})\mathbf{r}_i(\mathbf{v}).$$

Derive this identity with respect to $\mathbf{v}$ in the direction $\mathbf{r}_i(\mathbf{v})$ to obtain

$$\mathbf{A}(\mathbf{v})\cdot\big(\mathbf{r}_i(\mathbf{v}),\mathbf{r}_i(\mathbf{v})\big)+\mathbf{A}(\mathbf{v})\big(\mathbf{r}_i(\mathbf{v})\cdot\mathbf{r}_i(\mathbf{v})\big) = \big(\nabla\lambda_i(\mathbf{v})\cdot\mathbf{r}_i(\mathbf{v})\big)\mathbf{r}_i(\mathbf{v})+\lambda_i(\mathbf{v})\big(\mathbf{r}_i(\mathbf{v})\cdot\mathbf{r}_i(\mathbf{v})\big).$$

Next, apply the left eigenvector $\mathbf{l}_i(\mathbf{v})$ of $\mathbf{A}(\mathbf{v})$ on the left to this identity. Recalling that $\mathbf{l}_i(\mathbf{v}) \cdot \mathbf{r}_i(\mathbf{v}) = 1$, we get

$$\begin{aligned}\mathbf{l}_i(\mathbf{v}) \cdot \big(\mathbf{A}(\mathbf{v}) \cdot \big(\mathbf{r}_i(\mathbf{v}),\mathbf{r}_i(\mathbf{v})\big)\big) &+ \lambda_i(\mathbf{v})\big(\mathbf{r}_i(\mathbf{v}) \cdot \mathbf{r}_i(\mathbf{v})\big)\\ &= \big(\nabla\lambda_i(\mathbf{v}) \cdot \mathbf{r}_i(\mathbf{v})\big)\mathbf{r}_i(\mathbf{v}) + \lambda_i(\mathbf{v})\big(\mathbf{r}_i(\mathbf{v}) \cdot \mathbf{r}_i(\mathbf{v})\big).\end{aligned}$$

But the field $\mathbf{r}_i$ is GNL and normalized by (2.1). Hence

$$\mathbf{l}_i(\mathbf{v}) \cdot \big(\mathbf{A}(\mathbf{v}) \cdot \big(\mathbf{r}_i(\mathbf{v}),\mathbf{r}_i(\mathbf{v})\big)\big) = 1.$$

Finally, we have

$$\mathbf{l}_i(\mathbf{u}^L) \cdot \big((\mathbf{A}(\mathbf{u}^L) - \mathbf{A}(\mathbf{u}))\mathbf{u}'(x)\big) = -u_i(x)u_i'(x) + \zeta_i(x), \tag{3.12}$$

where the function ζ_i satisfies the estimate (3.11c). By integration of the formula (3.11c), we obtain that the function $\mathbf{l}_i(\mathbf{u}^L) \cdot \Phi(\mathbf{u})$ is written in the form of (3.8), where the function $\theta_i(\mathbf{u})$ is given by

$$\theta_i(\mathbf{u}^j)(x) = \int_{-\infty}^{x} \zeta_i(\mathbf{u}^j)(s)\, s.$$

The estimate (3.9a) follows from (3.11c).

Next, let be given $\mathbf{u}^1$ *and* $\mathbf{u}^2$ in $\mathcal{K}_i^\epsilon$. Set

$$\theta_i(\mathbf{u}^j)(x) = \mathbf{l}_i(\mathbf{u}^L) \cdot \Phi(\mathbf{u}^j)(x) + \frac{\big(u_i^j(x)\big)^2}{2}, \quad j = 1, 2.$$

The function ζ_i is of the third order in $\mathbf{u}^j$, and since $\mathbf{u}^j$ belongs to $\mathcal{K}_i^\epsilon$, the estimate in (3.9) follows from (3.10b) and (3.12). □

The system of differential equations (3.3) has been replaced by the fixed-point problem (3.4). More generally, we consider the following fixed-point problem:

$$\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)\big(\mathbf{u} - \mathbf{u}^L\big) - \mathbf{u}' = \hat{\Phi}(\sigma, \mathbf{u}), \tag{3.13}$$

where $\hat{\mathbf{A}}$ is a C^1 $p \times p$ matrix-valued function and $\hat{\Phi}$ is a continuous mapping from $\mathbb{R} \times \big(\mathbf{u}^L + W\big)$ to W. The main result of this section is the existence of fixed-point solutions of (3.13). Of course, as a consequence, this gives the existence of traveling-wave solutions of (3.1). The main reason we introduce this generality is that it allows us to handle the case of degenerate $\mathbf{D}$ without repeating the proof of Theorem 3.3 below.

We suppose that $\hat{\mathbf{A}}$ and $\hat{\Phi}$ satisfy the following:

[H1] Let $\mathbf{u}^L$ be fixed. There exist p distinct numbers $\hat{\lambda}_k(\mathbf{u}^L)$, $1 \le k \le p$, a subset $\mathcal{I}$ of $\{1, \dots, p\}$, and a positive number ϵ_2 such that if

$$|\sigma - \hat{\lambda}_k(\mathbf{u}^L)| \le \epsilon_2 \tag{3.14}$$

for some index $k \in \mathcal{I}$, the matrix $\hat{A}(\sigma, \mathbf{u}^L)$ has a simple eigenvalue $\tau_k(\sigma)$. Furthermore,

$$\left|\tau_k(\sigma) - \hat{\lambda}_k(\mathbf{u}^L) + \sigma\right| \le C|\hat{\lambda}_k(\mathbf{u}^L) - \sigma|^2. \tag{3.15}$$

Let $\mathbf{f}_k(\sigma)$ (resp., $\mathbf{g}_k(\sigma)$) denote the right (resp., left) eigenvector of $\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)$ associated with $\tau_k(\sigma)$. The vector $\mathbf{g}_k(\sigma)$ is normalized by

$$\mathbf{g}_k(\sigma) \cdot \mathbf{f}_k(\sigma) = 1$$

and the functions $\sigma \to \tau_k(\sigma)$, $\mathbf{f}_k(\sigma)$, and $\mathbf{g}_k(\sigma)$ are C^1. Setting $\mathbf{F}_k(\sigma) = \big\{\mathbf{f} \in \mathbb{R}^p;\ \mathbf{g}_k(\sigma) \cdot \mathbf{f} = 0\big\}$, we have that

$$|\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)\mathbf{f}| \ge c_2|\mathbf{f}| \quad \forall \mathbf{f} \in \mathbf{F}_k(\sigma) \tag{3.16}$$

for some positive number c_2.

Assumption [H1] is obviously satisfied when $\hat{\mathbf{A}}(\sigma, \mathbf{u}^L) = \mathbf{A}(\mathbf{u}^L) - \sigma\mathbf{Id}$. The numbers $\hat{\lambda}_k$, $1 \le k \le p$, are indeed the eigenvalues of $\mathbf{A}(\mathbf{u}^L)$; $\mathcal{I} = \{1, \dots, p\}$ and the eigenpair $(\hat{\tau}_k(\sigma), \hat{\mathbf{f}}_k(\sigma))$ is $\big(\lambda_k(\mathbf{u}^L) - \sigma, \mathbf{r}_k(\mathbf{u}^L)\big)$. Next, the vector space $\hat{\mathbf{F}}_k(\sigma)$ is given by

$$\hat{\mathbf{F}}_k(\sigma) = \bigoplus_{l \ne k} \mathbb{R}\mathbf{r}_l(\mathbf{u}^L).$$

Then, taking into account that the matrix $\mathbf{A}(\mathbf{u}^L)$ has p distinct eigenvalues and setting $c_2 = \inf_{l \ne k} |\lambda_l - \lambda_k| > 0$, condition (3.16) is met.

Note that when $\hat{\mathbf{A}}$ satisfies [H1], by virtue of (3.15), the pair $\big(\hat{\lambda}_k(\mathbf{u}^L), \mathbf{f}_k\big(\hat{\lambda}_k(\mathbf{u}^L)\big)\big)$ is an eigenpair of the matrix $\hat{\mathbf{A}}\big(\hat{\lambda}_k(\mathbf{u}^L), \mathbf{u}^L\big)$. For simplicity, we write

$$\hat{\lambda}_k(\mathbf{u}^L) = \hat{\lambda}_k^0, \quad \mathbf{f}_k(\hat{\lambda}_k(\mathbf{u}^L)) = \mathbf{f}_k^0, \quad \mathbf{g}_k(\hat{\lambda}_k(\mathbf{u}^L)) = \mathbf{g}_k^0.$$

Next, denote by $\mathbf{Q}_k(\sigma)$ the projector defined by the expression

$$\mathbf{Q}_k(\sigma)\mathbf{f} = \mathbf{f} - \big(\mathbf{g}_k(\sigma) \cdot \mathbf{f}\big)\mathbf{f}_k(\sigma), \quad \mathbf{f} \in \mathbb{R}^p, \tag{3.17}$$

and set $\mathbf{Q}_k^0 = \mathbf{Q}_k\big(\hat{\lambda}_k(\mathbf{u}^L)\big)$. The range of $\mathbf{Q}_k(\sigma)$ is the vector space $\mathbf{F}_k(\sigma)$. The two sets $\mathcal{K}_k^\epsilon$ and Σ_k^ϵ are naturally replaced by the following two sets:

$$\hat{\Sigma}_k^\epsilon = \Big\{\sigma \in \mathbb{R},\ \big|\hat{\lambda}_k(\mathbf{u}^L) - \sigma + \epsilon/2\big| \le c_3\epsilon^2\Big\}, \tag{3.18}$$

and

$$\hat{\mathcal{K}}_k^\epsilon = \big\{\mathbf{u} \in \mathbf{u}^L + W;\ \|\mathbf{u} - \mathbf{u}^L\|_W \le 2|\epsilon|\,|\mathbf{f}_k^0|,\ \big\|\mathbf{Q}_k^0(\mathbf{u} - \mathbf{u}^L)\big\|_W \le c_3\epsilon^2\big\}, \tag{3.19}$$

where c_3 is chosen a priori. Next, the function $\hat{\Phi}$ is assumed to satisfy the following two hypotheses:

[H2] For any pair $(\mathbf{u}^1, \mathbf{u}^2)$ that belongs to $\hat{\mathcal{K}}_i^\epsilon$, we have

$$\|\hat{\Phi}(\sigma, \mathbf{u}^j)\|_W \le c_4\epsilon^2, \ j = 1,\ 2, \quad \|\hat{\Phi}(\sigma, \mathbf{u}^1) - \hat{\Phi}(\sigma, \mathbf{u}^2)\|_W \le c_4|\epsilon|\|\mathbf{u}^1 - \mathbf{u}^2\|_W, \tag{3.20}$$

where c_4 is independent of c_3 in (3.18) and (3.19).

[H3] Let $\mathbf{u}^1$ *and* $\mathbf{u}^2$ belong to $\hat{\mathcal{K}}_i^\epsilon$. Set

$$u_i^j = \mathbf{g}_i(\sigma) \cdot \big(\mathbf{u}^j(x) - \mathbf{u}^L\big), \quad j = 1, 2. \tag{3.21}$$

Then

$$\mathbf{g}_i(\sigma) \cdot \hat{\Phi}(\sigma, \mathbf{u}^j) = \frac{\big(u_i^j(x)\big)^2}{2} + \hat{\theta}_i(\sigma, \mathbf{u}^j)(x), \tag{3.22}$$

where the function $\hat{\theta}_i$ satisfies

$$\begin{gathered} \|\hat{\theta}_i(\sigma, \mathbf{u}^j)\|_W \le C|\epsilon|^3, \quad j = 1, 2, \\ \|\hat{\theta}_i(\sigma, \mathbf{u}^1) - \hat{\theta}_i(\sigma, \mathbf{u}^2)| \le C \left(\epsilon^2 \|u_i^1 - u_i^2\|_W + |\epsilon| \|\mathbf{Q}_i(\sigma)\big(\mathbf{u}^1 - \mathbf{u}^2\big)\|_W\right). \end{gathered} \tag{3.23}$$

We can solve the fixed-point problem (3.13).

THEOREM 3.3. *Let $(\sigma, \mathbf{u}) \to \hat{\mathbf{A}}$ be a C^1 $p \times p$ matrix-valued function and let $\hat{\Phi}$ be a continuous mapping from $\mathbb{R} \times \big(\mathbf{u}^L + W\big)$ to W. Assume that* [H1], [H2], *and* [H3] *hold. Then we can find a positive number ϵ_0 such that for any index $i \in \mathcal{I}$, $\epsilon \in (-\epsilon_0, 0)$, and $\sigma \in \hat{\Sigma}_i^\epsilon$, the fixed-point problem* (3.13) *has a unique solution $\mathbf{u}$ in $\hat{\mathcal{K}}_i^\epsilon$.*

Proof. Let be given $\epsilon \in (-\epsilon_0, 0)$, where ϵ_0 is chosen below, $i \in \mathcal{I}$, and $\sigma \in \hat{\Sigma}_i^\epsilon$. We decompose a function $\mathbf{u}$ that belongs to $\hat{\mathcal{K}}_i^\epsilon$ into

$$\mathbf{u}(x) = \mathbf{u}^L + u_i(x)\mathbf{f}_i(\sigma) + \mathbf{Q}_i(\sigma)\big(\mathbf{u}(x) - \mathbf{u}^L\big). \tag{3.24}$$

The fixed-point problem (3.13) is then equivalent to the following problem:

$$\mathbf{Q}_i(\sigma)\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)\big(\mathbf{u} - \mathbf{u}^L\big) - \mathbf{Q}_i(\sigma)\mathbf{u}' = \mathbf{Q}_i(\sigma)\hat{\Phi}(\sigma, \mathbf{u}), \tag{3.25}$$

$$\tau_i(\sigma)u_i + \frac{u_i^2}{2} - u_i' = \hat{\theta}_i(\sigma, \mathbf{u}). \tag{3.26}$$

We solve (3.25)–(3.26) by using Banach's fixed-point theorem. For $\mathbf{v}$ that belongs to $\mathcal{K}_i^\epsilon$, we consider the following system of differential equations:

$$\mathbf{Q}_i(\sigma)\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)\big(\mathbf{u} - \mathbf{u}^L\big) - \mathbf{Q}_i(\sigma)\mathbf{u}' = \mathbf{Q}_i(\sigma)\hat{\Phi}(\sigma, \mathbf{v}), \tag{3.27}$$

$$\tau_i(\sigma)u_i + \frac{u_i^2}{2} - u_i' = \hat{\theta}_i(\mathbf{v}), \quad u_i(0) = -\tau_i(\sigma). \tag{3.28}$$

Note that the matrices $\mathbf{Q}_i(\sigma)$ and $\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)$ commute so that system (3.27) is a system of $p - 1$ linear differential equations with constant coefficients in $\mathbf{F}_i(\sigma)$. Its solution relies on the following.

LEMMA 3.4. *Let $\mathbf{B}_0$ be a given $p \times p$ matrix and $\mathbf{F}_0$ be a subspace of $\mathbb{R}^p$ with $\mathbf{B}_0\mathbf{F}_0 \subset \mathbf{F}_0$. Assume that*

$$|\mathbf{B}_0\mathbf{f}| \ge \hat{c}|\mathbf{f}| \quad \forall \mathbf{f} \in \mathbf{F}_0. \tag{3.29}$$

Let $\mathbf{Q}_0$ *denote a projector onto* $\mathbf{F}_0$. *Then for* $\mathbf{z} \in W$, *the system of differential equations*

$$\mathbf{B}_0\mathbf{w} - \mathbf{w}' = \mathbf{Q}_0\mathbf{z}$$

has a unique bounded solution $\mathbf{w} : \mathbb{R} \to \mathbf{F}_0$ *with* $\mathbf{w}(-\infty) = \mathbf{0}$. *Furthermore,*

$$\|\mathbf{w}\|_W \leq C\,\|\mathbf{z}\|_W\,, \tag{3.30}$$

where C *only depends on* $\hat{c}$ *and* $\|\mathbf{Q}_0\|$.

Proof. Denote by $\mathbf{F}_0^+$ (resp., $\mathbf{F}_0^-$) the subspace of $\mathbf{F}_0$, invariant by $\mathbf{B}_0$, associated with the positive (resp., negative) eigenvalues of $\mathbf{B}_{0|\mathbf{F}_0}$. Denote by $\mathbf{Q}_0^+$ (resp., $\mathbf{Q}_0^-$) the restriction of $\mathbf{Q}_0$ to the subspace $\mathbf{F}_0^+$ (resp., $\mathbf{F}_0^-$). The unique bounded solution $\mathbf{w} : \mathbb{R} \to \mathbf{F}_0$ with $\mathbf{w}(-\infty) = \mathbf{0}$ of (3.29) is given by

$$\begin{aligned} \mathbf{Q}_0^+\mathbf{w}(x) &= -\int_x^{+\infty} \exp\left(\mathbf{Q}_0^+\mathbf{B}_0(x-s)\right)\mathbf{Q}_0\mathbf{z}(s)s \\ \mathbf{Q}_0^-\mathbf{w}(x) &= \int_{-\infty}^{x} \exp\left(\mathbf{Q}_0^-\mathbf{B}_0(x-s)\right)\mathbf{Q}_0\mathbf{z}(s)s. \end{aligned} \tag{3.31}$$

Finally, by (3.29), the mapping $\mathbf{Q}_0\mathbf{B}_0 : \mathbf{F}_0 \to \mathbf{F}_0$ is regular, and a straightforward computation allows us to conclude the proof of Lemma 3.4. □

Next, we solve the differential equation (3.28) thanks to the following

LEMMA 3.5. *Let* $\zeta \in W$ *and* c_5 *be a positive number. We can find* $\delta_0 > 0$ *such that for any* $\delta \in (-\delta_0, \delta_0)$, *if*

$$\|\zeta\|_W \leq c_5|\delta|^3, \tag{3.32}$$

the unique solution of the following Cauchy problem

$$\frac{\delta}{2}w(x) + \frac{(w(x))^2}{2} - w'(x) = \zeta(x), \quad w(0) = -\frac{\delta}{2} \tag{3.33}$$

is bounded: $w' \in L^1(\mathbb{R})$. *Furthermore,* w *satisfies the following estimate:*

$$w(x) = \frac{-\delta}{1+\exp(-\delta x/2)} + z \quad \text{with} \quad z \in W \quad \text{and} \quad \|z\|_W \leq C\delta^2. \tag{3.34}$$

The proof of Lemma 3.5 is technical and is given in the appendix.

Let us return to the solution of (3.27)–(3.28). Let $\mathbf{v}$ belong to $\mathcal{K}_i^\epsilon$. According to Lemma 3.4, there exists a unique bounded solution $\mathbf{w} : \mathbb{R} \to \mathbf{F}_i(\sigma)$ of the following system of $p-1$ differential equations:

$$\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)(\mathbf{w} - \mathbf{Q}_i(\sigma)\mathbf{u}^L) - \mathbf{w}' = \mathbf{Q}_i(\sigma)\hat{\Phi}(\mathbf{v}).$$

On the other hand, by virtue of (3.15) and (3.18), $\tau_i(\sigma)$ satisfies

$$\left|\tau_i(\sigma) + \frac{\epsilon}{2}\right| \leq C\epsilon^2. \tag{3.35}$$

But the function $\hat{\theta}_i(\mathbf{v})$ satisfies the estimate (3.23) and, provided that ϵ_0 is chosen small enough, we can apply Lemma 3.5. The differential equation (3.28) has a unique solution u_i that belongs to W. (Note that $u_i(-\infty) = 0$ since $\tau_i(\sigma) > 0$.)

Let $\mathbf{u}$ denote the unique function in $\mathbf{u}^L + W$ such that $\mathbf{Q}_i(\sigma)(\mathbf{u} - \mathbf{u}^L) = \mathbf{w}$ and $\mathbf{g}_i(\sigma) \cdot \big(\mathbf{u}(x) - \mathbf{u}^L\big) = u_i(x)$. The function $\mathbf{u}$ is by construction a solution of (3.27)–(3.28). Furthermore,

$$\mathbf{Q}_i^0(\mathbf{u} - \mathbf{u}^L) = \mathbf{Q}_i(\sigma)(\mathbf{u} - \mathbf{u}^L) + (\mathbf{Q}_i^0 - \mathbf{Q}_i(\sigma))(\mathbf{u} - \mathbf{u}^L)$$

and we deduce from (3.30) and (3.18) that

$$\|\mathbf{Q}_i^0(\mathbf{u} - \mathbf{u}^L)\| \le c_6\epsilon^2 + C|\epsilon|^3 m, \tag{3.36}$$

where c_6 is independent of c_3 and C depends continuously on c_3 and other constants. On the other hand, by virtue of (3.18), (3.34), and (3.35), we obviously have·

$$\|\mathbf{u} - \mathbf{u}^L\| \le \frac{3}{2}|\epsilon||\mathbf{f}_i(\sigma)| + C\epsilon^2.$$

Hence we can choose ϵ_0 small enough so that the function $\mathbf{u}$ belongs to $\mathcal{K}_i^\epsilon$. Setting $\mathbf{u} = \mathcal{F}(\sigma, \mathbf{v})$, we have thus constructed a mapping from $\mathcal{K}_i^\epsilon$ to itself. Next, we prove the following lemma.

LEMMA 3.6. *We can choose ϵ_0 such that for fixed $\epsilon \in (-\epsilon_0, 0)$ and $\sigma \in \hat{\Sigma}_i^\epsilon$, the mapping $\mathcal{F}^2$ defined by*

$$\mathcal{F}^2(\sigma, \mathbf{v}) = \mathcal{F}\big(\sigma, \mathcal{F}(\sigma, \mathbf{v})\big)$$

is a contraction.

Proof. We first obtain some estimates of $\mathcal{F}$.

LEMMA 3.7. *Let $\mathbf{v}^1$ and $\mathbf{v}^2$ belong to $\mathcal{K}_i^\epsilon$. Then*

$$\begin{aligned} \|\mathcal{F}(\sigma, \mathbf{v}^1) - \mathcal{F}(\sigma, \mathbf{v}^2)\|_W &\le C\left(|\epsilon|\, \|\mathbf{g}_i(\sigma)\cdot(\mathbf{v}^1 - \mathbf{v}^2)\|_W + \|\mathbf{Q}_i(\sigma)(\mathbf{v}^1 - \mathbf{v}^2)\|_W\right), \\ \big\|\mathbf{Q}_i(\sigma)\big(\mathcal{F}(\sigma, \mathbf{v}^1) - \mathcal{F}(\sigma, \mathbf{v}^2)\big)\big\|_W &\le C|\epsilon|\, \|\mathbf{v}^1 - \mathbf{v}^2\|_W. \end{aligned} \tag{3.37}$$

Assume that (3.37) holds. We obtain that the mapping $\mathbf{v} \in \mathcal{K}_i^\epsilon \to \mathcal{F}^2(\sigma, \mathbf{v})$ is a contraction by iterating (3.37). The proof of Lemma 3.6 is then complete. □

Proof of Lemma 3.7. Let $\mathbf{v}^1$ *and* $\mathbf{v}^2$ belong to $\hat{\mathcal{K}}_i^\epsilon$. Set $\mathbf{u}^j = \mathcal{F}(\sigma, \mathbf{v}^j)$, $j = 1, 2$. The function $\mathbf{w} = \mathbf{Q}_i(\sigma)(\mathbf{u}^1 - \mathbf{u}^2)$ is a solution of the following system of differential equations:

$$\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)\mathbf{w} - \mathbf{w}' = \mathbf{Q}_i(\sigma)\big(\hat{\Phi}(\sigma, \mathbf{v}^1) - \hat{\Phi}(\sigma, \mathbf{v}^2)\big).$$

Then, taking into account the estimate (3.20), Lemma 3.4 gives

$$\big\|\mathbf{Q}_i(\sigma)\big(\mathbf{u}^1 - \mathbf{u}^2\big)\big\|_W \le C|\epsilon|\, \|\mathbf{v}^1 - \mathbf{v}^2\|_W. \tag{3.38}$$

Next, we have the following lemma.

LEMMA 3.8. *Let $\delta \in (-\delta_0, \delta_0)$. and ζ^1, $\zeta^2 \in W$ satisfy* (3.32). *Denote by w^j, $j = 1, 2$, the solution of* (3.33) *when ζ is replaced by ζ^j. Then*

$$\|w^1 - w^2\|_W \le \frac{C}{|\delta|}\, \|\zeta^1 - \zeta^2\|_W.$$

The proof of Lemma 3.8 can be found in the appendix. Denoting by u_i^j, $j = 1, 2$, the solution of the differential equation

$$\tau_i(\sigma)u_i^j + \frac{(u_i^j)^2}{2} - (u_i^j)' = \hat{\theta}_i(\mathbf{v}^j), \qquad u_i^j(0) = -\tau_i(\sigma), \qquad j = 1, 2,$$

and taking into account the estimate (3.23), we deduce from Lemma 3.8 that

$$\|u_i^1 - u_i^2\|_W \leq C\,(|\epsilon|\,\|v_i^1 - v_i^2\| + \|\mathbf{Q}_i(\sigma)(\mathbf{v}^1 - \mathbf{v}^2)\|), \tag{3.39}$$

where $v_i^j(x) = \mathbf{g}_i(\sigma)\cdot(\mathbf{v}(x) - \mathbf{u}^L)$, $j = 1, 2$. By the definition of $\mathcal{F}$, (3.38) and (3.39) yield (3.37).

For $\epsilon \in (-\epsilon_0, 0)$ and $\sigma \in \hat{\Sigma}_i^\epsilon$, the mapping $\mathbf{v} \to \mathcal{F}^2(\sigma, \mathbf{v})$ is a contraction defined in $\hat{\mathcal{K}}_i^\epsilon$. There exists a unique $\mathbf{u}_\sigma \in \mathcal{K}_i^\epsilon$ such that $\mathbf{u}_\sigma = \mathcal{F}(\sigma, \mathbf{u}_\sigma)$. The function $\mathbf{u}_\sigma$ is a solution of (3.13) and the proof of Theorem 3.3 is complete. □

4. Existence for general diffusion matrices.

In §3, we proved the existence of traveling-wave solutions of (3.1). This system is simple to handle because here the diffusion matrix is the identity matrix. Here we consider more general diffusion matrices of the form of (2.3). We prove that a traveling-wave solution of (1.1) is a solution of a fixed-point problem of the form of (3.13). Indeed, let $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2)$ be a traveling wave with speed σ solution of (1.1) such that $\mathbf{u}(-\infty) = \mathbf{u}^L$ and $\|\mathbf{u} - \mathbf{u}^L\|_W \leq \epsilon_2$ for some number ϵ_2 chosen below. Assume that the pair $(\sigma, \mathbf{u}^L)$ satisfies (2.4) and (2.5). Inserting the expressions (2.3) and (2.2) of the matrices $\mathbf{A}(\mathbf{u})$ and $\mathbf{D}(\mathbf{u})$ in (1.1), we obtain that $\mathbf{u}$ is a solution of the following system:

$$\big(\mathbf{A}_1(\mathbf{u}) - \sigma\mathbf{Id}_r\big)\mathbf{u}_1' + \mathbf{A}_2(\mathbf{u})\mathbf{u}_2' = \mathbf{0}, \tag{4.1}$$

$$\mathbf{A}_3(\mathbf{u})\mathbf{u}_1' + \big(\mathbf{A}_4(\mathbf{u}) - \sigma\mathbf{Id}_q\big)\mathbf{u}_2' - \big(\mathbf{D}_1(\mathbf{u})\mathbf{u}_1' + \mathbf{D}_2(\mathbf{u})\mathbf{u}_2'\big) = \mathbf{0}. \tag{4.2}$$

By virtue of (2.4), we obtain that for small enough ϵ_2, for any $x \in \mathbb{R}$, the matrix $\big(\mathbf{A}_1(\mathbf{u}(x)) - \sigma\mathbf{Id}_r\big)$ is regular. Hence

$$\mathbf{u}_1' = -\big(\mathbf{A}_1(\mathbf{u}) - \sigma\mathbf{Id}_r\big)^{-1}\mathbf{A}_2(\mathbf{u})\mathbf{u}_2'. \tag{4.3}$$

This differential equation defines a mapping $\mathbf{u}_2 \to \mathbf{u}_1$.

LEMMA 4.1. *Let $\mathbf{u}^L = (\mathbf{u}_1^L, \mathbf{u}_2^L) \in \Omega$ and $\sigma \in \mathbb{R}$ satisfy* (2.4) *and* (2.5); *we can find a positive ϵ_2 such that for any function $\mathbf{u}_2 \in \mathbf{u}_2^L + W$ with*

$$\|\mathbf{u}_2 - \mathbf{u}_2^L\|_W \leq \epsilon_2, \tag{4.4}$$

the system of differential equations (4.3) *has a unique solution $\mathbf{u}_1 = \mathcal{G}(\mathbf{u}_2) \in \mathbf{u}_1^L + W$. Furthermore, the mapping $\mathcal{G}$ is continuous:*

$$\|\mathcal{G}(\mathbf{u}_2^1) - \mathcal{G}(\mathbf{u}_2^2)\|_W \leq C\,\|\mathbf{u}_2^1 - \mathbf{u}_2^2\|_W\,. \tag{4.5}$$

Proof. A fixed-point method is convenient. Let $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2) \in \mathbf{u}^L + W$ with $\|\mathbf{v} - \mathbf{u}^L\|_W \leq \epsilon_2$ and $\sigma \in \mathbb{R}$ such that $(\sigma, \mathbf{u}^L)$ satisfies (2.4) and (2.5). The unique bounded solution $\mathbf{u}_1$ of the system of differential equations

$$\mathbf{u}_1' = -\big(\mathbf{A}_1(\mathbf{v}) - \sigma\mathbf{Id}_r\big)^{-1}\mathbf{A}_2(\mathbf{v})\mathbf{v}_2', \quad \mathbf{u}_1(-\infty) = \mathbf{u}_1^L$$

is given by

$$\mathbf{u}_1(x) = \mathbf{u}_1^L - \int_{-\infty}^{x} \big(\mathbf{A}_1(\mathbf{v}) - \sigma\mathbf{Id}_r\big)^{-1}\mathbf{A}_2(\mathbf{v})\mathbf{v}_2'\, s. \tag{4.6}$$

For $\mathbf{v} = \mathbf{v}_1^m$, $m = 1, 2$, as above, let $\mathbf{u}_1^m$, $m = 1, 2$, denote the functions defined by (4.6). A straightforward computation allows us to estimate the norm $\|\mathbf{u}_1^1 - \mathbf{u}_1^2\|_W$:

$$\|\mathbf{u}_1^1 - \mathbf{u}_1^2\|_W \le C\epsilon_2 \, \|\mathbf{v}^1 - \mathbf{v}^2\|_W \, .$$

For $(\sigma, \mathbf{u}_2)$ satisfying (4.4), using Banach's fixed-point theorem, it is now a simple matter to obtain the existence and the uniqueness of the function $\mathbf{u}_1 \in \mathbf{u}_1^L + W$ such that the pair $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2)$ is a solution of (4.3). The mapping $\mathcal{G}$ defined by $\mathbf{u}_1 = \mathcal{G}(\mathbf{u}_2)$ then obviously satisfies (4.5). □

Next, let us insert the expression (4.3) in (4.2). We obtain that $\mathbf{u}_2$ is a solution of the following system:

$$\mathbf{B}(\sigma, \mathbf{u})\mathbf{u}_2' - \left(\hat{\mathbf{D}}(\sigma, \mathbf{u})\mathbf{u}_2'\right)' = \mathbf{0}, \tag{4.7}$$

where

$$\mathbf{B}(\sigma, \mathbf{u}) = \mathbf{A}_4(\mathbf{u}) - \sigma\mathbf{Id}_q - \mathbf{A}_3(\mathbf{u})\big(\mathbf{A}_1(\mathbf{u}) - \sigma\mathbf{Id}_r\big)^{-1}\mathbf{A}_2(\mathbf{u}), \tag{4.8}$$

$$\hat{\mathbf{D}}(\sigma, \mathbf{u}) = \mathbf{D}_2(\mathbf{u}) - \mathbf{D}_1(\mathbf{u})\big(\mathbf{A}_1(\mathbf{u}) - \sigma\mathbf{Id}_r\big)^{-1}\mathbf{A}_2(\mathbf{u}). \tag{4.9}$$

LEMMA 4.2. *Assume that* $(\sigma, \mathbf{u}^L)$ *satisfies* (2.4) *and* (2.5). *We can choose* ϵ_2 *such that for any state* $\mathbf{u} \in \Omega$ *with* $|\mathbf{u} - \mathbf{u}^L| \le \epsilon_2$, *the matrix* $\hat{\mathbf{D}}(\sigma, \mathbf{u})$ *is regular.*

Proof. Let $\mathbf{h}_2$ belong to the kernel of $\hat{\mathbf{D}}(\sigma, \mathbf{u})$. Then the vector $\mathbf{h}$ defined by

$$\mathbf{h} = \begin{pmatrix} \big(\mathbf{A}_1(\mathbf{u}) - \sigma\mathbf{Id}_r\big)^{-1}\mathbf{A}_2(\mathbf{u})\mathbf{h}_2 \\ \mathbf{h}_2 \end{pmatrix}$$

belongs to the vector space $\ker\mathbf{D}(\mathbf{u}) \cap \big(\mathbf{A}(\mathbf{u}) - \sigma\mathbf{1}_p\big)^{-1}\mathrm{range}\mathbf{D}(\mathbf{u})$. By virtue of assumption (2.5), this is impossible when ϵ_2 is small enough. □

In fact, to write system (4.7) in the form of (3.13), we formally set

$$\mathbf{v}_2' = \hat{\mathbf{D}}(\sigma; (\mathcal{G}(\mathbf{u}_2), \mathbf{u}_2))\mathbf{u}_2', \quad \mathbf{v}_2(-\infty) = \mathbf{u}_2^L, \tag{4.10}$$

and we wish to write a system whose solution is $\mathbf{v}_2$. These considerations are made rigorous with the following lemma.

LEMMA 4.3. *Let* $(\sigma, \mathbf{u}^L)$ *satisfy* (2.4) *and* (2.5). *Let* $\mathbf{u}_2$ *belong to* $\mathbf{u}_2^L + W$ *that satisfies* (4.4). *The function* $\mathbf{v}_2 \in \mathbf{u}_2^L + W$, *whose derivative is given by* (4.10), *belongs to the set* $\mathbf{u}_2^L + W$ *and satisfies*

$$\left\|\mathbf{v}_2 - \mathbf{u}_2^L\right\|_W \le C\epsilon_2. \tag{4.11}$$

Conversely, we can choose ϵ_2 *such that for* $\mathbf{v}_2$ *that belongs to* $\mathbf{u}_2^L + W$ *and satisfies* (4.4), *there exists a unique solution* $\mathbf{u}_2 = \mathcal{H}_2(\mathbf{v}_2) \in \mathbf{u}_2^L + W$ *of* (4.10) *with*

$$\left\|\mathbf{u}_2 - \mathbf{u}_2^L\right\|_W \le C\epsilon_2. \tag{4.12}$$

The mapping $\mathcal{H}_2$ *is continuous:*

$$\|\mathcal{H}_2(\mathbf{v}_2^1) - \mathcal{H}_2(\mathbf{v}_2^2)\|_W \le C\,\|\mathbf{v}_2^1 - \mathbf{v}_2^2\|_W \quad \forall \mathbf{v}_2^1, \mathbf{v}_2^2 \in \mathbf{u}_2^L + W. \tag{4.13}$$

Proof. The derivation of the estimate (4.11) is straightforward. Conversely, let $\mathbf{v}_2$ in $\mathbf{u}_2^L + W$ satisfy (4.4). We solve (4.10) with a fixed-point method. The proof relies on Lemma 4.2 and is similar to the proof of Lemma 4.1. □

Next, we define the mapping $\mathcal{H}$ by

$$\mathcal{H}(\mathbf{v}_2) = (\mathcal{G}(\mathcal{H}_2(\mathbf{v}_2)), \mathcal{H}_2(\mathbf{v}_2)). \tag{4.14}$$

The mapping $\mathcal{H}$ is, of course, continuous. Furthermore, if $\mathbf{v}_2$ is a solution of the system of differential equations

$$\hat{\mathbf{A}}\big(\sigma, \mathcal{H}(\mathbf{v}_2)\big)\mathbf{v}_2' - \mathbf{v}_2'' = \mathbf{0}, \tag{4.15}$$

where the matrix-valued function $\hat{\mathbf{A}}$ is given by

$$\hat{\mathbf{A}}(\sigma, \mathbf{u}) = \mathbf{B}(\sigma, \mathbf{u})\big(\hat{\mathbf{D}}(\sigma, \mathbf{u})\big)^{-1}, \tag{4.16}$$

the function $\mathbf{u} = \mathcal{H}(\mathbf{v}_2)$ is a solution of (1.1).

Finally, system (1.1) has been replaced by system (4.15), where the matrix in front of the second-order derivative is the identity matrix. Next, if $\mathbf{v}_2$ is a solution of (4.15), then

$$\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)(\mathbf{v}_2 - \mathbf{v}_2^L) - \mathbf{v}_2' = \hat{\Phi}(\sigma, \mathbf{v}_2), \quad \mathbf{v}_2(-\infty) = \mathbf{u}_2^L, \tag{4.17}$$

where

$$\hat{\Phi}(\sigma, \mathbf{v}_2)(x) = \int_{-\infty}^{x} \Big(\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)\mathbf{v}_2' - \mathbf{B}(\sigma, \mathcal{H}(\mathbf{v}_2))\Big(\big(\hat{\mathbf{D}}(\sigma, \mathcal{H}(\mathbf{v}_2))\big)^{-1}\mathbf{v}_2'\Big)'\Big) ds. \tag{4.18}$$

To apply Theorem 3.3, we check that the matrix-valued function $\hat{\mathbf{A}}$ and the function $\hat{\Phi}$ satisfy hypotheses [H1], [H2], and [H3] from §3. We first prove the following proposition.

PROPOSITION 4.4. *The matrix-valued function $\hat{\mathbf{A}}$ satisfies assumption* [H1].

Proof. First, a straightforward computation gives the following lemma.

LEMMA 4.5. *For $\mathbf{u} \in \Omega$ and $\sigma \in \mathbb{R}$, the pair $(\tau, \mathbf{f}) \in \mathbb{R} \times \mathbb{R}^q$ is an eigenpair of the matrix $\hat{\mathbf{A}}(\sigma, \mathbf{u})$ if the vector $\mathbf{h}$, given by*

$$\mathbf{h} = \begin{pmatrix} -\big(\mathbf{A}_1(\mathbf{u}) - \sigma\mathbf{Id}_r\big)^{-1}\mathbf{A}_2(\mathbf{u})\big(\hat{\mathbf{D}}(\sigma, \mathbf{u})\big)^{-1}\mathbf{f} \\ \big(\hat{\mathbf{D}}(\sigma, \mathbf{u})\big)^{-1}\mathbf{f} \end{pmatrix}, \tag{4.19}$$

satisfies

$$\big(\mathbf{A}(\mathbf{u}) - \sigma\mathbf{Id}_p\big)\mathbf{h} = \tau\mathbf{D}(\mathbf{u})\mathbf{h}.$$

Next, we have the following lemma.

LEMMA 4.6. *Assume that the field $\mathbf{r}_i$ is GNL. Choose $\mathbf{u} \in \Omega$. There exists a C^1 curve $\sigma \in (\lambda_i(\mathbf{u}^L) - \hat{\epsilon}, \lambda_i(\mathbf{u}^L) + \hat{\epsilon}) \to (\tau_i(\sigma), \mathbf{h}_i(\sigma))$ such that*

$$\big(\mathbf{A}(\mathbf{u}) - \sigma\mathbf{Id}_p\big)\mathbf{h}_i(\sigma) = \tau_i(\sigma)\mathbf{D}(\mathbf{u})\mathbf{h}_i(\sigma) \tag{4.20}$$

and

$$\begin{gathered} (\mathbf{l}_i(\mathbf{u}), \mathbf{h}_i(\sigma)) = 1, \\ |\mathbf{h}_i(\sigma) - \mathbf{r}_i(\mathbf{u}^L)| \leq C|\lambda_i(\mathbf{u}) - \sigma|, \\ |\tau_i(\sigma) - \lambda_i(\mathbf{u}) + \sigma| \leq C|\lambda_i(\mathbf{u}^L) - \sigma|^2. \end{gathered} \tag{4.21}$$

Proof. First, when $\sigma = \lambda_i(\mathbf{u})$, system (4.20) has the solution $(\tau^0, \mathbf{h}^0) = (0, \mathbf{r}_i(\mathbf{u}))$. Next, we decompose a vector $\mathbf{h}$ on the eigenbasis $\{\mathbf{r}_k(\mathbf{u})\}_{1\leq k\leq p}$ of the matrix $\mathbf{A}(\mathbf{u})$: $\mathbf{h} = \sum_{k=1}^{p} h_k \mathbf{r}_k(\mathbf{u})$. We set $h_i = 1$ and define the mapping ϕ_σ from $\mathbb{R}^{p+1}$ to $\mathbb{R}^p$ by

$$\phi(\sigma, \tau, (h_k)_{k\neq i}) = (\mathbf{A}(\mathbf{u}) - \sigma \mathbf{1}) \sum_{k\neq i} h_k \mathbf{r}_k(\mathbf{u}) - \tau \mathbf{D}(\mathbf{u}) \Big(\mathbf{r}_i(\mathbf{u}) + \sum_{k\neq i} h_k \mathbf{r}_k(\mathbf{u}) \Big).$$

The mapping ϕ is C^∞, and a straightforward computation gives

$$\begin{aligned} &\frac{\partial \phi}{\partial \tau}(\lambda_i(\mathbf{u}), 0, \mathbf{0}) = -\mathbf{D}(\mathbf{u})\mathbf{r}_i(\mathbf{u}), \\ &\frac{\partial \phi}{\partial h_k}(\lambda_i(\mathbf{u}), 0, \mathbf{0}) = \big(\mathbf{A}(\mathbf{u}) - \lambda_i(\mathbf{u})\mathbf{Id}\big)\mathbf{r}_k(\mathbf{u}), \qquad k \neq i. \end{aligned}$$

The vector space spanned by the $p-1$ vectors $\big(\mathbf{A}(\mathbf{u}) - \lambda_i(\mathbf{u})\mathbf{Id}\big)\mathbf{r}_k(\mathbf{u})$, $k \neq i$, is $(\mathbf{l}_i(\mathbf{u}))^\perp$ because the matrix $\mathbf{A}(\mathbf{u})$ has distinct eigenvalues. On the other hand, since the field $\mathbf{r}_i$ is GNL, we have by virtue of (2.7) that $\mathbf{l}_i(\mathbf{u}) \cdot \mathbf{D}(\mathbf{u})\mathbf{r}_i(\mathbf{u}) \neq 0$, and we deduce that the p vectors $\big(\mathbf{A}(\mathbf{u}) - \lambda_i(\mathbf{u})\mathbf{Id}\big)\mathbf{r}_k(\mathbf{u})$, $k \neq i$, and $\mathbf{D}(\mathbf{u})\mathbf{r}_i(\mathbf{u})$ span $\mathbb{R}^p$. Thus we may apply the implicit-function theorem and obtain a C^1 curve $\sigma \to (\tau_i(\sigma), \mathbf{h}_i(\sigma))$ of solutions of (4.20). The derivation of (4.21) is then straightforward. □

We deduce from Lemmas 4.5 and 4.6 that given a GNL field $\mathbf{r}_i$, for ϵ_2 chosen small enough and σ such that $|\sigma - \lambda_i(\mathbf{u}^L)| \leq \epsilon_2$, the matrix $\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)$ has a right eigenvector $\mathbf{f}_i(\sigma)$ associated with the eigenvalue $\tau_i(\sigma)$. By virtue of (4.21), the function $\sigma \to \tau_i(\sigma)$ satisfies (3.15). We denote by $\mathbf{g}_i(\sigma)$ the associated left eigenvector, normalized by $\mathbf{g}_i(\sigma) \cdot \mathbf{f}_i(\sigma) = 1$. The set $\mathcal{I}$ of hypothesis [H1] is thus the set of indices i for which the field $\mathbf{r}_i$ is genuinely nonlinear.

LEMMA 4.7. *Let $\sigma \in \mathbb{R}$ and $\mathbf{u}^L \in \Omega$ such that $(\sigma, \mathbf{u}^L)$ satisfies* (2.4) *and* (2.5). *Assume that the matrix $\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)$ has an eigenpair $(\tau_0, \mathbf{f}_0)$. Denote by $\mathbf{g}_0$ the left eigenvector of $\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)$ associated with τ_0 normalized by $\mathbf{g}_0 \cdot \mathbf{f}_0 = 1$. Denote by $\mathbf{F}_0$ the vector space $\mathbf{F}_0 = \{\mathbf{f},\ \mathbf{g}_0 \cdot \mathbf{f} = 0\}$. We can find a positive number $\hat{\alpha}$ that depends only on σ and $\mathbf{u}^L$ such that if $|\tau_0| \leq \hat{\alpha}$, then*

$$\big|\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)\mathbf{f}\big| \geq \hat{c}|\mathbf{f}| \quad \forall \mathbf{f} \in \mathbf{F}_0$$

for some number $\hat{c}$ independent of $\hat{\alpha}$.

Proof. Let $\mathbf{f} \in \mathbb{R}^q$ with $|\mathbf{f}| = 1$. Set

$$\alpha = \Big|\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)\mathbf{f}\Big|.$$

Next, define next the vector $\mathbf{h}$ by (4.19). By Lemma 4.5,

$$\big(\mathbf{A}(\mathbf{u}^L) - \sigma\mathbf{Id}\big)\mathbf{h} = \begin{pmatrix} \mathbf{0} \\ \hat{\mathbf{A}}(\sigma, \mathbf{u}^L)\mathbf{f} \end{pmatrix}$$

so that

$$|(\mathbf{A}(\mathbf{u}^L) - \sigma\mathbf{Id})\mathbf{h}| \leq \alpha.$$

But the matrix $\mathbf{A}(\mathbf{u}^L)$ has p distinct eigenvalues, and we deduce that

$$|\mathbf{h} - a\mathbf{r}_i(\mathbf{u}^L)| \leq C\alpha$$

for some real number a. On the other hand, the same computation applied to the vector $\mathbf{f}_0$ gives

$$|\mathbf{h}_0 - a_0\mathbf{r}_i(\mathbf{u}^L)| \leq C|\tau_0|,$$

where $\mathbf{h}_0$ is obtained from $\mathbf{f}_0$ according to (4.19). We deduce that

$$|\mathbf{h} - \mathbf{h}_0| \leq C(\alpha + |\tau_0|).$$

However, the matrix $\hat{\mathbf{D}}(\sigma, \mathbf{u}^L)$ is regular, and we obtain

$$|\mathbf{f} - \mathbf{f}_0| \leq C(\alpha + |\tau_0|).$$

Next, assume that the vector $\mathbf{f}$ belongs to the vector space $\mathbf{F}_0$. Then

$$1 = |\mathbf{g}_0 \cdot (\mathbf{f} - \mathbf{f}_0)| \leq C(\alpha + |\tau_0|)$$

so that

$$\alpha \geq \frac{1}{C} - |\tau_0|.$$

This concludes the proof of Lemma 4.7. □

Setting $\mathbf{F}_i(\sigma) = \{\mathbf{f} \in \mathbb{R}^p,\ \mathbf{g}_i(\sigma) \cdot \mathbf{f} = 0\}$, the estimate (3.16) follows from Lemma 4.7. The proof of Proposition 4.4 is complete. □

The function $\hat{\Phi}$ obviously satisfies assumption [H2] since it is quadratic in $\mathbf{v}_2$. Finally, we check [H3].

PROPOSITION 4.8. *Let i be an index such that the field $\mathbf{r}_i$ is genuinely nonlinear. Choose $\mathbf{u}^L \in \Omega$ such that $(\lambda_i(\mathbf{u}^L), \mathbf{u}^L)$ satisfies* (2.4) *and* (2.5). *Then the mapping $\hat{\Phi}$ satisfies the assumption* [H3].

Proof. Let $\mathbf{r}_i$ be a GNL field of $\mathbf{A}$ and assume that $(\lambda_i(\mathbf{u}^L), \mathbf{u}^L)$ satisfies (2.4) and (2.5). Let $\mathbf{v}_2 \in \mathbf{u}_2^L + W$. Set $\mathbf{u} = \mathcal{H}(\mathbf{v}_2)$. Then a straightforward computation gives

$$\begin{pmatrix} \mathbf{0} \\ \hat{\Phi}(\sigma, \mathbf{v}_2)(x) \end{pmatrix} = \Phi(\mathbf{u})(x) = \int_{-\infty}^{x} (\mathbf{A}(\mathbf{u}^L) - \mathbf{A}(\mathbf{u}(s)))\mathbf{u}'(s)ds. \tag{4.22}$$

Recalling that the properties of the mapping Φ are given by Lemmas 3.1 and 3.2, we must connect the functions $u_i = \mathbf{l}_i(\mathbf{u}^L) \cdot (\mathbf{u} - \mathbf{u}^L)$ and $v_i = \mathbf{g}_i^0 \cdot (\mathbf{v}_2 - \mathbf{u}_2^L)$, where $\mathbf{g}_i^0 = \mathbf{g}_i(\lambda_i(\mathbf{u}^L))$, to conclude the proof of Proposition 4.8.

LEMMA 4.9. *Assume that the field $\mathbf{r}_i$ is GNL and that $(\lambda_i(\mathbf{u}^L), \mathbf{u}^L)$ satisfies* (2.4) *and* (2.5). *Let $(\sigma, \mathbf{v}_2) \in \mathbb{R} \times (\mathbf{u}_2^L + W)$ such that*

$$|\sigma - \lambda_i(\mathbf{u}^L)| \leq \epsilon_2, \quad \|\mathbf{v}_2 - \mathbf{u}_2^L\|_W \leq \epsilon_2. \tag{4.23}$$

Let $\mathbf{u}_2 = \mathcal{H}_2(\mathbf{v}_2)$ denote the solution of (4.10), *set $\mathbf{u} = (\mathcal{G}(\mathbf{u}_2), \mathbf{u}_2)$, and let*

$$u_i(x) = \mathbf{l}_i(\mathbf{u}^L) \cdot (\mathbf{u}(x) - \mathbf{u}^L), \tag{4.24a}$$

$$v_i(x) = \mathbf{g}_i^0 \cdot (\mathbf{v}_2(x) - \mathbf{u}_2^L). \tag{4.24b}$$

Then, provided that ϵ_2 is small enough,

$$u_i = v_i + \varpi_i(\mathbf{v}_2), \tag{4.25a}$$

where the function ϖ *satisfies*

$$\|\varpi_i(\mathbf{v}_2)\|_W \le C(\epsilon^2 + \|\mathbf{Q}_i^0(\mathbf{v}_2 - \mathbf{u}_2^L\|_W) \quad \forall \mathbf{v}_2 \in \hat{\mathcal{K}}_i^\epsilon, \tag{4.25b}$$

$$\|\varpi(\mathbf{v}_2^1) - \varpi(\mathbf{v}_2^2)\|_W \le C\|\mathbf{v}_2^1 - \mathbf{v}_2^2\|_W \quad \forall \mathbf{v}_2^1,\, \mathbf{v}_2^2 \in \hat{\mathcal{K}}_i^\epsilon. \tag{4.25c}$$

Furthermore,

$$\|\mathbf{u} - \mathbf{u}^L - u_i \mathbf{r}_i(\mathbf{u}^L)\| \le C\left(\epsilon^2 + \|\mathbf{Q}_i^0(\mathbf{v}_2 - \mathbf{u}_2^L)\|\right). \tag{4.26}$$

Proof. Let $(\sigma, \mathbf{v}_2)$ satisfy (4.23) and $\mathbf{u}_2$ denote the solution of (4.10). Then

$$\mathbf{v}_2(x) - \mathbf{u}_2^L = \hat{\mathbf{D}}^0(\mathbf{u}_2(x) - \mathbf{u}_2^L) + \int_{-\infty}^{x} (\hat{\mathbf{D}}(\mathbf{u}(s)) - \hat{\mathbf{D}}^0)\mathbf{u}_2'(s)s,$$

where $\hat{\mathbf{D}}^0 = \hat{\mathbf{D}}(\lambda_i(\mathbf{u}^L), \mathbf{u}^L)$. Next, set $\mathbf{u}_1 = \mathcal{G}(\mathbf{u}_2)$. By the definition of $\mathbf{u}_1$, a straightforward computation gives

$$\left\|\mathbf{u}_1(x) - \mathbf{u}_1^L + \left(\mathbf{A}_1(\mathbf{u}^L) - \lambda_i(\mathbf{u}^L)\mathbf{Id}_r\right)\mathbf{A}_2(\mathbf{u}^L)(\mathbf{u}_2(x) - \mathbf{u}_2^L)\right\|_W \le C\epsilon^2,$$

and we deduce that

$$\left\|\mathbf{u} - \mathbf{u}^L - \mathbf{C}^0\left(\mathbf{v}_2 - \mathbf{u}_2^L\right)\right\| \le C\epsilon^2, \tag{4.27}$$

where the matrix $\mathbf{C}^0$ is given by

$$\mathbf{C}^0 = \begin{pmatrix} -\left(\mathbf{A}_1(\mathbf{u}^L) - \lambda_i(\mathbf{u}^L)\mathbf{Id}_r\right)^{-1}\mathbf{A}_2(\mathbf{u}^L)\left(\hat{\mathbf{D}}^0\right)^{-1} \\ \left(\hat{\mathbf{D}}^0\right)^{-1} \end{pmatrix}. \tag{4.28}$$

On the other hand, apply Lemma 4.5 with $\sigma = \lambda_i(\mathbf{u}^L)$. We obtain the following expression of the eigenvector $\mathbf{r}_i(\mathbf{u}^L)$ of $\mathbf{A}(\mathbf{u}^L)$ in function of $\mathbf{f}_i^0 = \mathbf{f}_i\left(\lambda_i(\mathbf{u}^L)\right)$:

$$\mathbf{r}_i(\mathbf{u}^L) = \mathbf{C}^0\mathbf{f}_i^0.$$

However, the function $\mathbf{v}_2$ is written

$$\mathbf{v}_2(x)\mathbf{u}_2^L = v_i(x)\mathbf{f}_i^0 + \mathbf{Q}_i^0(\mathbf{v}_2(x) - \mathbf{v}_2^L)$$

so that

$$\mathbf{C}^0\left(\mathbf{v}_2(x) - \mathbf{u}_2^L\right) = v_i(x)\mathbf{r}_i(\mathbf{u}^L) + \mathbf{C}^0\mathbf{Q}_i^0(\mathbf{v}_2(x) - \mathbf{u}_2^L). \tag{4.29}$$

Inserting (4.27) in this expression gives

$$\|\mathbf{u} - \mathbf{u}^L - v_i\mathbf{r}_i(\mathbf{u}^L)\|_W \le C(\epsilon^2 + \|\mathbf{Q}_i^0(\mathbf{v}_2 - \mathbf{u}_2^L)\|_W). \tag{4.30}$$

By the definition of the functions u_i and $\varpi(\mathbf{v}_2)$, we deduce (4.25b) from this estimate. On the other hand, the derivation of (4.24c) is straightforward and the estimate (4.26) follows from (4.30) and (4.25b). □

Lemma 4.9 shows that we can choose the constants in the definitions of the two sets $\mathcal{K}_i^\epsilon$ and $\hat{\mathcal{K}}_i^\epsilon$ such that if $\mathbf{v}_2$ belongs to $\hat{\mathcal{K}}_i^\epsilon$, the function $\mathbf{u} = \mathcal{H}(\mathbf{v}_2)$ belongs to $\mathcal{K}_i^\epsilon$. However, for $\mathbf{u}$ belonging to $\mathcal{K}_i^\epsilon$, Lemmas 3.1 and 3.2 apply and—thanks to the identity (4.22)—it is a simple matter to check that the function $\hat{\Phi}$ satisfies assumption [H3]. This completes the proof of Proposition 4.8. $\square$

We can now conclude the proof of Theorem 2.1. Assume that the field $\mathbf{r}_i$ is GNL and choose $\mathbf{u}^L \in \Omega$ such that $(\lambda_i(\mathbf{u}^L), \mathbf{u}^L)$ satisfies (2.4) and (2.5). Then we can find a positive ϵ_0 such that for $\sigma \in \mathbb{R}$ and $\mathbf{u} \in \mathbf{u}^L + W$ that satisfy $|\sigma - \lambda_i(\mathbf{u}^L)| \leq \epsilon_0$ and $\|\mathbf{u} - \mathbf{u}^L\|_W \leq \epsilon_0$, the pair $(\sigma, \mathbf{u})$ is a solution of (1.1) when the pair $(\sigma, \mathbf{v}_2)$ is a solution of (4.17). Next, since hypotheses [H1], [H2], and [H3] hold, we may apply Theorem 3.3. When σ belongs to Σ_i^ϵ, we obtain a solution of (4.17), from which we deduce the existence of a traveling wave $\mathbf{u}_\sigma$ with speed σ and $\mathbf{u}_\sigma(-\infty) = \mathbf{u}^L$, a solution of (1.1). We obtain next the estimate (2.8) from (4.25), and the proof of Theorem 2.1 is complete. $\square$

5. Uniqueness.

This section is devoted to the proof of Theorem 2.2. We assume in this section that every characteristic field $\mathbf{r}_i$ is GNL. Let $(\sigma, \mathbf{u}^L) \in \mathbb{R} \times \Omega$ satisfy (2.4) and (2.5). Let $\mathbf{u}$ be a traveling wave with speed σ and $\mathbf{u}(-\infty) = \mathbf{u}^L$, a solution of (1.1). Set $\delta = |\mathbf{u}(+\infty) - \mathbf{u}^L|$ and assume that

$$\|\mathbf{u} - \mathbf{u}^L\|_W \leq \hat{c}\delta. \tag{5.1}$$

The transformation of system (1.1) into the fixed-point problem (4.17) relies only on (2.4) and (2.5). Provided that δ is small enough, the function $\mathbf{u}$ is written $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2)$ with $\mathbf{u}_1 = \mathcal{G}(\mathbf{u}_2)$. Next, the system of differential equations (4.10) has a unique bounded solution $\mathbf{v}_2$ that is a solution of (4.17). Note that only [H2] holds a priori, i.e.,

$$\left\|\hat{\Phi}(\sigma, \mathbf{v}_2)\right\|_W \leq C\delta^2. \tag{5.2}$$

The function $\hat{\Phi}$ is indeed quadratic in $\mathbf{v}_2$. On the contrary, the proofs of [H1] and [H3] rely on the fact that σ is close to one of the eigenvalues of $\mathbf{A}(\mathbf{u}^L)$, which we do not suppose.

To begin, we estimate σ.

LEMMA 5.1. *We can find two positive numbers c_7 and δ_1 such that if $\delta \leq \delta_1$, the matrix $\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)$ has a small eigenvalue τ_0:*

$$|\tau_0| \leq c_7\delta. \tag{5.3}$$

Proof. Assume to the contrary that

$$\left|\left(\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)\right)^{-1}\right| \leq C \tag{5.4}$$

for some positive number C. Denote by $\mathbf{E}^+$ (resp., $\mathbf{E}^-$) the sum of the eigenspaces associated with the positive (resp., negative) eigenvalues of $\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)$. Next, denote by $\mathbf{R}^+$ (resp., $\mathbf{R}^-$) the projector on $\mathbf{E}^+$ (resp., $\mathbf{E}^-$) with kernel $\mathbf{E}^-$ (resp. $\mathbf{E}^+$). The unique bounded solution of (4.17) is given by

$$\mathbf{R}^+\mathbf{v}_2(x) = \mathbf{R}^+\mathbf{u}_2^L - \int_x^{+\infty} \exp\bigl(\mathbf{R}^+\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)(x-s)\bigr)\hat{\Phi}(\sigma, \mathbf{v}_2)(s)s,$$

$$\mathbf{R}^-\mathbf{v}_2(x) = \mathbf{R}^-\mathbf{u}_2^L + \int_{-\infty}^{x} \exp\bigl(\mathbf{R}^-\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)(x-s)\bigr)\hat{\Phi}(\sigma, \mathbf{v}_2)(s)s.$$

By virtue of (5.4), we deduce that

$$\|\mathbf{v}_2 - \mathbf{u}_2^L\| \leq C\|\hat{\Phi}(\sigma, \mathbf{v}_2)\|_W$$

and, taking into account (5.2), we deduce that

$$\|\mathbf{v}_2 - \mathbf{u}_2^L\|_W \leq C\delta^2.$$

For δ small enough, this contradicts (5.1), and we deduce that the matrix $\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)$ has a small eigenvalue τ_0 that satisfies (5.3) for some positive number c_7. □

Henceforth, we assume that $\delta \leq \delta_1$ so that the matrix $\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)$ has an eigenvalue τ_0 that satisfies (5.3). Denote by $\mathbf{f}_0$ (resp., $\mathbf{g}_0$) the right (resp., left) eigenvector of $\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)$ associated with τ_0. We normalize $\mathbf{g}_0$ by $\mathbf{g}_0 \cdot \mathbf{f}_0 = 1$.

LEMMA 5.2. *We can choose δ_1 such that if $\delta \leq \delta_1$,*

$$|\sigma - \lambda_i(\mathbf{u}^L)| \leq C\delta. \tag{5.5}$$

Proof. Let $\mathbf{h}_0$ denote the vector defined by (4.19) with $\mathbf{f} = \mathbf{f}_0$. According to Lemma 4.5, the pair $(\tau_0, \mathbf{h}_0)$ satisfies

$$\big(\mathbf{A}(\mathbf{u}^L) - \sigma\mathbf{Id}_p\big)\mathbf{h}_0 = \tau\mathbf{D}(\mathbf{u})\mathbf{h}_0,$$

and we deduce that

$$\big|(\mathbf{A}(\mathbf{u}^L) - \sigma\mathbf{Id}_p)\mathbf{h}_0\big| \leq C\delta.$$

However, the matrix $\mathbf{A}(\mathbf{u}^L)$ has p distinct eigenvalues, and if $\delta \leq \delta_1$ with δ_1 chosen small enough, we get (5.5) for some index $i \in \{1, \ldots, p\}$. □

Since $(\sigma, \mathbf{u}^L)$ satisfies (2.4) and (2.5), we obtain that provided that δ is small enough, the pair $(\lambda_i(\mathbf{u}^L), \mathbf{u}^L)$ also satisfies (2.4) and (2.5). The field $\mathbf{r}_i$ is by assumption GNL and we may apply Lemma 4.6. We obtain a C^1 curve $\sigma \rightarrow (\tau_i(\sigma), \mathbf{f}_i(\sigma))$ defined in a neighborhood of $\sigma = \lambda_i(\mathbf{u}^L)$ such that for any σ, the pair $(\tau_i(\sigma), \mathbf{f}_i(\sigma))$ is an eigenpair of $\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)$. Then, obviously, the eigenvector $\mathbf{f}_0$ of $\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)$ is $\mathbf{f}_0 = \mathbf{f}_i(\sigma)$. In the same manner, $\mathbf{g}_0 = \mathbf{g}_i(\sigma)$ and $\tau_0 = \tau_i(\sigma)$.

Let $\mathbf{F}_i(\sigma)$ denote the vector space defined by $\mathbf{F}_i(\sigma) = \{\mathbf{f},\ \mathbf{g}_i(\sigma) \cdot \mathbf{f} = 0\}$. The assumptions of Lemma 4.7 are satisfied and provided that δ_1 is chosen small enough, we deduce that

$$\big|\hat{\mathbf{A}}(\sigma, \mathbf{u}^L)\mathbf{f}\big| \geq c_8|\mathbf{f}| \quad \forall \mathbf{f} \in \mathbf{F}_i(\sigma) \tag{5.6}$$

for some number c_8 independent of δ. The behavior of σ is precisely described with the following lemma.

LEMMA 5.3. *The number σ satisfies*

$$|\sigma - \lambda_i(\mathbf{u}^L) - \epsilon/2| \leq C\epsilon^2, \tag{5.7}$$

where

$$\epsilon = \mathbf{l}_i(\mathbf{u}^L) \cdot (\mathbf{u}(+\infty) - \mathbf{u}^L). \tag{5.8}$$

Proof. The function $\mathbf{u}$ is a solution of (1.1). We integrate this system over $\mathbb{R}$ to obtain

$$\big(\mathbf{A}(\mathbf{u}^L) - \sigma\mathbf{Id}\big)(\mathbf{u}(+\infty) - \mathbf{u}^L) = \Phi(\mathbf{u})(+\infty),$$

where $\Phi(\mathbf{u})$ is defined by (3.4). Next, we multiply this system on the left by the eigenvector $\mathbf{l}_i(\mathbf{u}^L)$:

$$\big(\lambda_i(\mathbf{u}^L) - \sigma\big)\mathbf{l}_i(\mathbf{u}^L) \cdot \big(\mathbf{u}(+\infty) - \mathbf{u}^L\big) = \mathbf{l}_i(\mathbf{u}^L) \cdot \hat{\Phi}(\mathbf{u})(+\infty).$$

We have assumed that every characteristic field of $\mathbf{A}$ is GNL, and we may apply Lemma 3.2. We obtain

$$\left|(\lambda_i(\mathbf{u}^L) - \sigma)u_i(+\infty) - \frac{(u_i(+\infty))^2}{2}\right| \leq C\delta^3, \tag{5.9}$$

where $u_i(x) = \mathbf{l}_i(\mathbf{u}^L) \cdot (\mathbf{u}(x) - \mathbf{u}^L)$.

On the other hand, define ϵ by (5.8). Then

$$\big||\epsilon| - \delta\big| \leq C\delta^2. \tag{5.10}$$

Indeed, since the function $\mathbf{Q}_i(\sigma)(\mathbf{v}_2 - \mathbf{u}_2^L)$ is a solution of (3.25), we obtain by virtue of Lemma 3.4 and the estimate (5.6) the following estimate:

$$\big\|\mathbf{Q}_i(\sigma)(\mathbf{v}_2 - \mathbf{u}_2^L)\big\|_W \leq C\delta^2.$$

Next, by the continuity of the mapping $\sigma \to \mathbf{Q}_i(\sigma)$, we deduce that

$$\big\|\mathbf{Q}_i^0(\mathbf{v}_2 - \mathbf{u}_2^L)\big\|_W \leq C\delta^2, \tag{5.11}$$

where $\mathbf{Q}_i^0 = \mathbf{Q}_i(\lambda_i(\mathbf{u}^L))$. Finally, we apply Lemma 4.9. The estimates (5.11) and (4.25) give

$$\|\mathbf{u} - \mathbf{u}^L - u_i\mathbf{r}_i(\mathbf{u}^L)\|_W \leq C\delta^2, \tag{5.12}$$

from which we deduce (5.10).

We insert (5.10) in (5.9) to obtain (5.7).

Thus far, we have proved that all the properties listed in assumptions [H1] and [H2] in §3 hold true. Furthermore, the properties of the function $\hat{\Phi}$ listed in assumption [H3] rely on the algebraic computation (4.22) and on Lemma 4.9 so that [H3] is valid. Next, by virtue of the estimates (5.7) and (5.12), σ belongs to Σ_i^ϵ defined by (3.5) and the function $\mathbf{u}$ belongs to $\mathcal{K}_i^\epsilon$ defined by (3.6) provided that c_0 is chosen large enough. In the same manner, since the function $\mathbf{v}_2$ satisfies (5.11), the pair $(\sigma, \mathbf{v}_2)$ belongs to $\hat{\Sigma}_i^\epsilon \times \hat{\mathcal{K}}_i^\epsilon$ defined by (3.18) and (3.19) provided that c_2 is chosen large enough.

Next, we prove that ϵ defined by (5.8) is negative. By virtue of [H3], the function $v_i = \mathbf{g}_i(\sigma) \cdot (\mathbf{v}_2 - \mathbf{u}_2^L)$ is a solution of the differential equation

$$\tau_i(\sigma)v_i + \frac{(v_i)^2}{2} - v_i' = \hat{\theta}_i(\sigma, \mathbf{v}_2),$$

where $\hat{\theta}$ satisfies (3.23) and

$$v_i(-\infty) = 0, \quad |v_i(+\infty) - \epsilon| \leq C\epsilon^2.$$

However, since σ satisfies (5.7), $\tau_i(\sigma)$ satisfies (3.35) so that $v_i(a) = -\tau_i(\sigma)/2$ for some real number a. The function $w_i = \mathbf{g}_i(\sigma) \cdot (\mathbf{w}_2 - \mathbf{u}_2^L)$, where $\mathbf{w}_2(x) = \mathbf{v}_2(x - a)$, $x \in \mathbb{R}$, is thus a solution of

$$\tau_i(\sigma)w_i + \frac{(w_i)^2}{2} - w_i' = \hat{\theta}_i(\sigma, \mathbf{w}_2), \quad w_i(0) = -\frac{\tau_i(\sigma)}{2}.$$

We can then apply Lemma 3.5:

$$\left\| w_i + \frac{2\tau_i(\sigma)}{1+\exp(-\tau_i(\sigma)x/2)} \right\|_W \le C\epsilon^2. \tag{5.13}$$

By virtue of (3.35), $\tau_i(\sigma)$ has the opposite sign of ϵ. If ϵ is positive, $\tau_i(\sigma)$ is negative, and we deduce from (5.13) that in this case

$$|w_i(-\infty) + \epsilon| \le C\epsilon^2,$$

which contradicts the fact that $w_i(-\infty) = 0$. Hence ϵ is negative.

We can now conclude the proof of Theorem 2.2. The functions $\mathbf{w}_2$ and $\mathbf{v}_2$ are solutions of (4.17). By construction of the mapping $\mathcal{F}$, we have

$$\mathbf{w}_2 = \mathcal{F}(\sigma, \mathbf{w}_2).$$

However, we proved that σ belongs to $\hat{\Sigma}_i^\epsilon$ and $\mathbf{w}_2$ belongs to $\hat{\mathcal{K}}_i^\epsilon$ for some index $i \in \{1, \ldots, p\}$ provided that c_3 in the definitions (3.18) and (3.19) is chosen large enough. On the other hand, we proved that ϵ is negative, and we proved in §3 that if $|\epsilon|$ is small enough, for $\sigma \in \hat{\Sigma}_i^\epsilon$, the mapping $\mathbf{v}_2 \to \mathcal{F}(\sigma, \mathbf{v}_2)$ has a unique fixed point in $\hat{\mathcal{K}}_i^\epsilon$. We deduce that if δ_0 is chosen small enough and if the function $\mathbf{u}$ satisfies (2.9), then $\mathbf{u}(x-a) = \mathbf{u}_\sigma(x)$, where $\mathbf{u}_\sigma$ is given by Theorem 2.1. This concludes the proof of Theorem 2.2. □

Appendix. A nonlinear differential equation. This appendix is devoted to the proof of Lemmas 3.5 and 3.8.

PROPOSITION A.1. *Let $\zeta \in W$ and d_0 be a positive number. We can find a positive δ_0 such that for any $\delta \in (-\delta_0, \delta_0)$, if*

$$\|\zeta\|_W \le d_0|\delta|^3, \tag{A.1}$$

the unique solution of the Cauchy problem

$$-\frac{\delta}{2}w(x) + \frac{(w(x))^2}{2} - w'(x) = \zeta(x), \quad w(0) = \frac{\delta}{2} \tag{A.2}$$

is bounded. $w' \in L^1(\mathbb{R})$ and satisfies the estimate

$$w(x) = \frac{\delta}{1+\exp(\delta x/2)} + z \quad \textit{with} \quad z \in W \quad \textit{and} \quad \|z\|_W \le d_1\delta^2. \tag{A.3}$$

Next, let ζ^m, $m = 1, 2$, satisfy (A.1). *Denote by w^m, $m = 1, 2$, the solution of the differential equation* (A.2), *where ζ is replaced by ζ^m, $m = 1, 2$. We have the estimate*

$$\|w^1 - w^2\|_W \le \frac{d_2}{|\delta|}\|\zeta^1 - \zeta^2\|_W. \tag{A.4}$$

Proof. When $\zeta = 0$, the unique solution of the Cauchy problem (A.2) is

$$w(x) = \frac{\delta}{1+\exp(\delta x/2)}.$$

Next, a function w in the form of (A.3) is a solution of (A.2) if the function z is a solution of the following differential equation:

$$\frac{\delta}{2}\text{th}\left(\frac{\delta x}{4}\right)z(x)+z'(x)=-\zeta(x)+\frac{(z(x))^2}{2},\quad z(0)=0. \tag{A.5}$$

We need the following lemma.

Lemma A.2. *Let $\zeta_1\in W$. Then the unique solution of the linear Cauchy problem,*

$$\frac{\delta}{2}\text{th}\left(\frac{\delta x}{4}\right)z(x)+z'(x)=\zeta(x),\quad z(0)=0, \tag{A.6}$$

lies in W. Furthermore,

$$\|z\|_W\le\frac{d_3}{|\delta|}\|\zeta\|_W. \tag{A.7}$$

Proof. Set $\hat{z}(x)=z(4x/\delta)$ and $\hat{\zeta}(x)=\zeta(4x/\delta)$. Then z is a solution of (A.6) if $\hat{z}$ satisfies

$$2\text{th}(x)\hat{z}(x)+\hat{z}'(x)=\frac{4}{\delta}\hat{\zeta}(x).$$

However, $\|\hat{\zeta}\|_W=\|\zeta\|_W$ and it suffices to prove Lemma A.2 when $\delta=4$. The solution of (A.6) is then

$$z(x)=\frac{-1}{\text{ch}^2(x)}\int_0^x\zeta(s)\text{ch}^2(s)ds,$$

and a straightforward integration by parts gives

$$z(x)=\frac{\zeta(x)}{\text{ch}^2(x)}\left(\frac{\text{sh}(2x)}{4}+\frac{x}{2}\right)-\frac{1}{\text{ch}^2(x)}\int_0^x\zeta'(s)\left(\frac{\text{sh}(2s)}{4}+\frac{s}{2}\right)ds.$$

We deduce that

$$\|z\|_W\le 7\|\zeta\|_W.$$

This concludes the proof of Lemma A.2. □

We solve equation (A.5) by using a fixed-point method. Choose $d_4=4d_3d_0$. Next, let K denote the following closed convex subset of W:

$$K=\left\{z\in W,\ z(0)=0,\ \|z\|_W\le d_4\delta^2\right\}.$$

Choose $\delta_0=1/2d_3d_4$. By Lemma A.2, for $\delta\in(-\delta_0,\delta_0)$ and $y\in K$, the unique solution of the Cauchy problem

$$\frac{\delta}{2}\text{th}\left(\frac{\delta x}{4}\right)z(x)+z'(x)=-\zeta(x)+\frac{(y(x))^2}{2},\quad z(0)=0 \tag{A.8}$$

lies in K. This defines a mapping $\mathcal{G}$ from K into itself. The mapping $\mathcal{G}$ is contracting. Indeed, let δ belong to $(-\delta_0,\delta_0)$, ζ^m, $m=1$, 2, satisfy (A.1), and y^m, $m=1$, 2, belong to K. For $m=1$, 2, denote by z^m the unique solution of the Cauchy problem (A.8), where the functions ζ and y are replaced by ζ^m and y^m, respectively. A straightforward computation gives

$$\frac{\delta}{2}\text{th}\left(\frac{\delta x}{4}\right)(z^1-z^2)(x)+(z^1-z^2)'(x)=(\zeta^2-\zeta^1)(x)+(y^1-y^2)(x)\frac{(y^1+y^2)(x)}{2}.$$

Applying Lemma A.2 again gives

$$\|z^1 - z^2\|_W \le d_3\left(\frac{1}{|\delta|}\|\zeta^1 - \zeta^2\|_W + d_4|\delta|\|y^1 - y^2\|_W\right). \tag{A.9}$$

However, $d_3 d_4|\delta| \le 1/2$ and the mapping $\mathcal{G}$ is thus a contraction. It has a unique fixed point in K, which means that the Cauchy problem (A.5) has a unique solution z. Then the function w defined by (A.3) is a solution of (A.2).

Finally, it remains to prove the estimate (A.4). Let ζ^m, $m = 1, 2$, satisfy (A.1) and denote by w^m, $m = 1, 2$, the solution of the Cauchy problem (A.3) when the function ζ is ζ^m. The estimate (A.9) is written

$$\|z^1 - z^2\|_W \le d_3\left(\frac{1}{|\delta|}\|\zeta^1 - \zeta^2\|_W + d_4|\delta|\,\|z^1 - z^2\|_W\right).$$

However, by construction, $d_3 d_4|\delta| \le 1/2$ and

$$\|z^1 - z^2\|_W \le \frac{2d_3}{|\delta|}\|\zeta^1 - \zeta^2\|_W.$$

This concludes the proof of Proposition A.1. □

Acknowledgments. I am indebted to P.-A. Raviart, who encouraged my interest in nonconservative hyperbolic systems and maintained a constant interest in my work. I also wish to thank D. Serre for many interesting discussions about systems in nonconservation form.

REFERENCES

[1] R. Courant and K. O. Friedrichs, *Supersonic Flows and Shock Waves*, Wiley–Interscience, New York, 1948; reprinted by Springer-Verlag, New York, 1985.

[2] L. Sainsaulieu, *A Euler system modeling vaporizing sprays*, in Dynamics of Heterogeneous Combustion and Reacting Systems, Progress Series in Astronautics and Aeronautics 152, A. L. Kuhl et al., eds., American Institute of Aeronautics and Astronautics, Washington, DC, 1993, pp. 280–305.

[3] P. Embid and M. Baer, *Mathematical analysis of a two-phase model for reactive granular material*, Technical report SAND88-3302, Sandia National Laboratory, Albuquerque, NM, 1989.

[4] E. F. Toro, *Riemann-problem based techniques for computing reactive two-phase flows*, in Numerical Combustion, A. Dervieux and B. Larrouturou, eds., Lectures Notes in Phys. 351, Springer-Verlag, Heidelberg, 1989, pp. 472–481.

[5] S. K. Godunov and E. I. Romenskii, *Nonstationary equations of nonlinear elasticity theory in Eulerian coordinates*, Zh. Prikl. Mekh. Tekh., 6 (1972), pp. 124–144.

[6] A. I. Volpert, *The space BV and quasilinear equations*, Mat. Sb., 73 (1967), pp. 225–267.

[7] G. Dal Maso, P. Le Floch, and F. Murat, *Definition and weak stability of a nonconservative product*, J. Math. Pures Appl., to appear.

[8] J.-F. Colombeau, *Multiplication of distributions,* Bull. Amer. Math. Soc., 23 (1990), pp. 251–268.

[9] J.-F. Colombeau and A. Heibig, *Nonconservative products in bounded variation functions,* SIAM J. Math. Anal., 23 (1992), pp. 941–949.

[10] P. Le Floch, *Shock waves for nonlinear hyperbolic systems in nonconservative form,* preprint series 593, Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, MN, 1989.

[11] P.-A. RAVIART AND L. SAINSAULIEU, *A nonconservative convection-diffusion system describing spray dynamics, part* 1: *Solution of Riemann problem,* Math. Models Methods Appl. Sci., 5 (1995), pp. 297–333.

[12] A. MAJDA AND R. PEGO, *Stable viscosity matrices for systems of conservation laws,* J. Differential Equations, 56 (1985), pp. 229–262.

[13] D. WAGNER, *The existence and behavior of viscous structure for plane detonation waves*, SIAM J. Math. Anal., 20 (1989), pp. 1035–1054.

[14] B. L. KEYFITZ AND G. G. WARNECKE, *The existence of viscous profiles and admissibility for transonic shocks*, Comm. Partial Differential Equations, 16 (1991), pp. 1197–1221.

[15] J. SMOLLER, *Shock waves and reaction diffusion equations*, in Grundlehren der mathematischen Wissenschaft 258, Springer-Verlag, New York, 1982.

[16] E. GODLEWSKI AND P.-A. RAVIART, *Hyperbolic Systems of Conservation Laws*, Ellipses, Paris, 1991.

[17] T. P. LIU, private communication, 1992.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1311–1335, September 1996

THE STABILITY OF ROLL SOLUTIONS OF THE TWO-DIMENSIONAL SWIFT–HOHENBERG EQUATION AND THE PHASE-DIFFUSION EQUATION*

MASATAKA KUWAMURA†

Abstract. A stability criterion of roll solutions of the two-dimensional Swift–Hohenberg equation is presented. It clarifies the effect of the system size on the primary instabilty of rolls. An interpretation of the phase-diffusion equation is also given from the viewpoint of spectral analysis. The key to carrying out the spectral analysis is that the infinite-dimensional system of linear equations naturally induced by the Fourier decomposition for the linearized eigenvalue problem of the roll solution can be reduced to the three-dimensional system.

Key words. phase-diffusion equation, Eckhauss instability, zigzag instability, Swift–Hohenberg equation

AMS subject classifications. 35B35, 35Q35, 76E15

1. Introduction. Let us consider a fluid contained in a rectangular cell whose aspect ratio of the depth of the fluid to the horizontal width is sufficiently small. For a critical temperature gradient between the upper and lower plates, buoyancy forces overcome the dissipative effects of viscous shear and thermal conduction, and the motionless fluid spontaneously breaks up into convective rolls of upward- and downward-moving regions of fluid as in Figure 1. In order to study this phenomena, the following simple model equation, which was first derived by Swift and Hohenberg [14], is proposed:

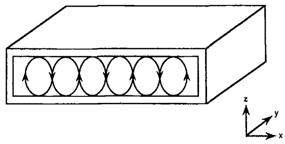


FIG. 1.

$$u_t = (\alpha - (1 + \partial_x^2 + \partial_y^2)^2)u - u^3, \tag{1}$$

where $u(x, y, t)$ represents the rescaled fluid field in a given horizontal plane, e.g., the vertical velocity component in the midplane of the convective rolls, and α is the reduced Rayleigh number, i.e.,

$$\alpha = \frac{R}{R_c} - 1,$$

where R_c is the critical Rayleigh number.

One can heuristically argue that the Swift–Hohenberg equation describes the onset of thermal convection, which forms roll patterns. Let $u \equiv 0$ be the trivial stationary solution of (1), which represents the rest state of the fluid. We immediately find that the eigenvalue of the linearized operator of the right-hand side of (1) at $u \equiv 0$ associated to the Fourier mode $\exp(i(kx + ly))$ is given by

* Received by the editors June 27, 1994; accepted for publication (in revised form) April 19, 1995.

† Hiroshima National College of Maritime Technology, Higashino, Hiroshima 725-02, Japan. Current address: Department of Computer and Communication Sciences, Wakayama University, Wakayama 640, Japan.

$$\mu_{k,l} = \alpha - (1 - (k^2 + l^2))^2.$$

If $\alpha < 0$, then we see that $\mu_{k,l} < 0$ holds for all k and l, so $u \equiv 0$ is stable. This shows that convection does not occur provided $R < R_c$. If $\alpha > 0$, then we see that the largest eigenvalues are $\mu_{k,l} > 0$ $(k^2 + l^2 = 1)$, so $u \equiv 0$ is unstable. The instability in the direction of the Fourier mode $\exp(\pm ix)$ $(k = \pm 1,\, l = 0)$ grows up, and another equilibrium appears which has a spatially periodic structure in the x-direction and uniform in the y-direction. This shows that the convective rolls occur provided $R > R_c$.

The reader should consult Cross and Hohenberg [2], Greenside and Coughran, Jr. [3], Manneville [9], and Newell [10] for the physical background of the Swift–Hohenberg equation.

The above discussion suggests the possibility that there exist stationary solutions of (1) with $\alpha > 0$ which have a spatially periodic structure in the x-direction and are uniform in the y-direction. In fact, Collet and Eckmann [1] showed the following.

EXISTENCE OF ROLL SOLUTIONS. *Suppose that ω satisfies*

$$2/5 < \omega^2 < 2.$$

Then there exists a positive constant ε_0 independent of ω such that for $0 < \varepsilon < \varepsilon_0$, the equation

$$(\alpha - (1 + \partial_x^2 + \partial_y^2)^2)u - u^3 = 0$$

has a unique solution of the form

$$\alpha = 3\varepsilon^2 + (1 - \omega^2)^2, \tag{2}$$

and

$$u_0(x) = \overline{u}(\omega x), \tag{3}$$

where

$$\overline{u}(z) = \varepsilon 2\cos(z) + \sum_{n \geq 3,\, n:\text{odd}} \eta_n 2\cos(nz) \tag{4}$$

with

$$\mid \eta_n \mid \leq C\varepsilon^{1+2n/3}, \tag{5}$$

where η_n depends only on ω and ε.

Notice that that the roll solutions (3) are also stationary solutions of the one-dimensional Swift–Hohenberg equation

$$u_t = (\alpha - (1 + \partial_x^2)^2)u - u^3. \tag{6}$$

The linear stability criterion of the roll solutions (3) of the one-dimensional Swift–Hohenberg equation (6) was also given in [1] as follows.

LINEAR STABILITY CRITERION. *Let $A : L^2(\mathbf{R}) \longrightarrow L^2(\mathbf{R})$ be the linearized operator of the right-hand side of the above equation at u_0 defined by*

$$Av = (\alpha - (1 + \partial_x^2)^2)v - 3u_0^2 v, \quad v \in H^4(\mathbf{R}).$$

Let

$$W = \frac{\omega^2 - 1}{\sqrt{3}\varepsilon}. \tag{7}$$

If $|W| < 1/\sqrt{2}$, then for sufficiently small $\varepsilon > 0$, the spectrum of A lies in the closed left half-plane in $\mathbf{C}$. On the other hand, if $|W| > 1/\sqrt{2}$, then for sufficiently small $\varepsilon > 0$, the spectrum of A intersects the right half-plane.

On the other hand, Kuramoto [7] and Pomeau and Manneville [12] studied the slow dynamics of (1) near the stationary solutions u_0 from the viewpoint of physics. When one observes a roll pattern in large-aspect-ratio Rayleigh–Bénard convection, the pattern may look almost regular everywhere locally in space, whereas globally, the contours of equal phase may deviate largely from straight lines. Based upon the above observation, they proposed (A) to study the effect of perturbations to the stationary solutions u_0 as the phase modulation of $\overline{u}$ in terms of a longer space scale and a slower time scale than the original one, and (B) to regard as virtually negligible the deformations which are not absorbed in the phase modulation when the new scales are used. They studied the dynamics of the phase modulation of $\overline{u}$ by using the formal perturbation method. They succeeded to show at a formal level that the dynamics near u_0 is given by

$$\begin{cases} \overline{u}(\omega x + \phi(X, Y, T)), \\ \phi_T = D_{//}\phi_{XX} + D_{\perp}\phi_{YY}, \quad X = \nu x, \quad Y = \nu y, \quad T = \nu^2 t, \end{cases} \tag{8}$$

where

$$D_{//} = 4 - 8W^2 + O(\varepsilon), \tag{9}$$

$$D_{\perp} = 2\sqrt{3}\varepsilon W, \tag{10}$$

and W is given by (7). Here ν is an artificial parameter for performing a formal perturbation method. This parameter ν determines the new scales X, Y, and T. The diffusion equation in (8) is called the *phase-diffusion equation* which describes the dynamics near the roll solutions.

Thus u_0 is stable if $D_{//} > 0$ and $D_{\perp} > 0$. On the other hand, u_0 is unstable if either $D_{//} < 0$ (the Eckhauss instability) or $D_{\perp} < 0$ (the zigzag instability). See Figure 2.

According to the results of [7] and [12], we know that when the system size is sufficiently large, there are only two types of instabilities, Eckhauss and zigzag. Our question is now as follows: What happens when the system size is not large? For example, fluid mechanics tells us that the zigzag instability is not observed when the length of the axis of rolls is small. Hence the stability of rolls must decrease as the length of rolls increases. What is the critical length where the stability changes? How does the instability of rolls depend on the system size? The aim of this paper is to answer these questions and give a mathematical justification for the physicist's ideas. Before presenting our main results, we formulate our problem precisely.

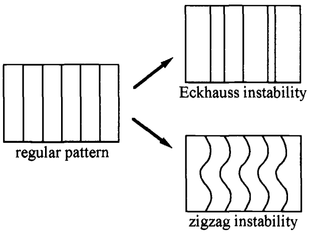


FIG. 2.

We restrict (1), which is originally defined on the infinite spatial domain $\mathbf{R^2}$, to the rectangle domain

$$\Omega = (-L/2, L/2) \times (-M/2, M/2), \quad 0 < L < \infty, \quad 0 < M < \infty$$

with periodic boundary conditions, i.e.,

$$u_t = (\alpha - (1 + \partial_x^2 + \partial_y^2)^2)u - u^3 \quad \text{for} \quad (x, y, t) \in \Omega \times (0, \infty) \tag{11}$$

with

$$\partial_x^j u(-L/2, y, t) = \partial_x^j u(L/2, y, t) \quad (j = 0, 1, 2, 3) \\ \text{for} \quad (y, t) \in (-M/2, M/2) \times (0, \infty)$$

and

$$\partial_y^j u(x, -M/2, t) = \partial_y^j u(x, M/2, t) \quad (j = 0, 1, 2, 3) \\ \text{for} \quad (x, t) \in (-L/2, L/2) \times (0, \infty).$$

This domain Ω serves as a window through which we can practically observe objects of infinite size. Here we assume

$$L = (2N)\lambda,$$

where $\lambda = 2\pi/\omega$ is the wavelength of $u_0(x)$ and N is a positive integer which corresponds to the number of rolls. In other words, the length of the side in the x-direction is an integer multiple of the basic wavelength of the roll pattern. Notice that M is the length of the axis of rolls.

We know that (11) generates a semiflow on $H_{per}^{4\beta}(\Omega)$ for $0 \le \beta < 1$, where

$$H_{per}^{4\beta}(\Omega) = \left\{ u(x, y) = \sum_{m,n=-\infty}^{\infty} u_{m,n} \exp(2\pi i n x/L) \exp(2\pi i m y/M); \right. \\ \left. \sum_{m,n=-\infty}^{\infty} (1 + |m|^2 + |n|^2)^{4\beta} |u_{m,n}|^2 < \infty \right\}.$$

For more details, see Henry [4] and Temam [15].

Notice that $u_0(x)$ is also a stationary solution of (11).

By (2) and (3), u_0 is determined by two parameters ε and ω, which represent the amplitude and wavenumber, respectively. Therefore, the stability of u_0 is determined by these parameters. However, we use a new parameter W instead of ω to investigate the stability of the equilibrium u_0, which is defined by (7). In what follows, we regard ε and W as independent parameters and consider that ω is determined by ε and W in terms of

$$\omega^2 = 1 + \sqrt{3}\varepsilon W. \tag{12}$$

The linear stability criterion of roll solutions is as follows.

THEOREM 1.1. *Let $A : L^2(\Omega) \longrightarrow L^2(\Omega)$ be the linearized operator of the right-hand side of* (1) *at u_0 defined by*

$$Av = (\alpha - (1 + \partial_x^2 + \partial_y^2)^2)v - 3u_0^2 v, \quad v \in H^4(\Omega) \tag{13}$$

with periodic boundary conditions

$$\begin{array}{ll} \partial_x^j v(-L/2, y) = \partial_x^j v(L/2, y) & \text{for} \quad y \in (-M/2, M/2), \\ \partial_y^j v(x, -M/2) = \partial_y^j v(x, M/2) & \text{for} \quad x \in (-L/2, L/2) \\ & (j = 0, 1, 2, 3). \end{array} \tag{14}$$

Then we have the following:

(A) *If $0 \leq W < 1/\sqrt{2}$, then for sufficiently small $\varepsilon > 0$, the spectrum of A lies in the closed left half-plane in* **C**. *This is independent of L and M.*

(B) *If $-1/\sqrt{2} < W < 0$, then for sufficiently small $\varepsilon > 0$,*

(i) *the spectrum of A lies in the closed left half-plane in* **C** *provided*

$$0 < \varepsilon < \frac{2\pi^2}{\sqrt{3}|W|M^2},$$

(ii) *the spectrum of A intersects the right half-plane in* **C** *provided*

$$\varepsilon > \frac{2\pi^2}{\sqrt{3}|W|M^2},$$

where $\varepsilon M^2 = O(1)$ as $\varepsilon \downarrow 0$ and $M \to \infty$.

(C) *If $|W| > 1/\sqrt{2}$, then for sufficiently small $\varepsilon > 0$ and large L, the spectrum of A intersects the right half-plane in* **C**.

When the rectangle domain Ω is sufficiently large and the roll pattern is stable, we can give an accurate characterization of the critical eigenvalues and the associated eigenfunctions.

THEOREM 1.2. *When L and M are sufficiently large, for $0 \leq W < 1/\sqrt{2}$ and sufficiently small $\varepsilon > 0$, there exist $\delta > 0$ which depend only on ε and W* (*independent of L and M*) *such that the following hold:*

(i) $\lim_{\varepsilon\downarrow 0} \delta(\varepsilon, W) = 0$.

(ii) *The eigenvalues of A which belong to the interval $[-\delta, 0]$ are given by*

$$\mu_{mn} = -D_\perp \nu_m - \nu_m^2 - D_{//}\kappa_n^2 + O((\kappa_n + \nu_m)^3)$$

$$\text{for} \quad 0 \leq \nu_m < \sqrt{3}\varepsilon\rho \quad \text{and} \quad |\kappa_n| \leq \sqrt{3}\varepsilon\rho/2,$$

where $D_{//}$ and $D_{\perp}$ are given by (9) *and* (10)*, respectively, and*

$$\nu_m = \left(\frac{2\pi m}{M}\right)^2,$$
$$\kappa_n = \frac{2\pi n}{L},$$

and $\rho > 0$ is a constant independent of ε, W, L, and M. The associated eigenfunctions are given by

$$\psi_{mn} = \partial_x u_0 \exp(2\pi i m y/M) \exp(2\pi i n x/L) + O(\kappa_n) + O(\nu_m).$$

(iii) *The other eigenvalues belong to the interval* $(-\infty, -\delta)$.

Theorem 1.2 says that there are many eigenvalues near zero when the system size is sufficiently large. However, these eigenvalues are discrete because L and M are finite. Therefore, we can take an eigenspace whose dimension is finite but sufficiently large as follows.

THEOREM 1.3. *When $0 < W < 1/\sqrt{2}$ and ε is sufficiently small, for sufficiently large L and M, we can choose $\beta > 0$ and $\gamma > 0$ which depend on ε, W, L, and M such that the following hold:*

(i) *β and γ satisfy*

$$0 < \beta < \gamma,$$
$$\lim_{L,M\to\infty} \beta(\varepsilon, W, L, M) = 0,$$
$$\lim_{L,M\to\infty} (\gamma(\varepsilon, W, L, M) - \beta(\varepsilon, W, L, M)) = 0.$$

(ii) *The eigenvalues of A which belong to the interval $[-\beta, 0]$ are given by*

$$\mu_{mn} = -D_{\perp}\left(\frac{2\pi m}{M}\right)^2 - D_{//}\left(\frac{2\pi n}{L}\right)^2 + o\left(\left(\frac{1}{M} + \frac{1}{L}\right)^2\right)$$

for $|m| \le \rho_1(M)$ and $|n| \le \rho_2(L)$, where $\rho_1(M)$ and $\rho_2(L)$ are integers such that

$$\lim_{M\to\infty} \rho_1(M) = \infty, \quad \lim_{M\to\infty} \frac{\rho_1(M)}{\sqrt[3]{M}} = 0,$$
$$\lim_{L\to\infty} \rho_2(L) = \infty, \quad \lim_{L\to\infty} \frac{\rho_2(L)}{\sqrt[3]{L}} = 0$$

and the associated eigenfunctions are given by

$$\psi_{mn} = \partial_x u_0 \exp(2\pi i m y/M) \exp(2\pi i n x/L) + O(1/M) + O(1/L).$$

(iii) *The eigenvalues μ which belong to the interval $(-\infty, -\beta)$ satisfy*

$$\mu < -\gamma.$$

The choice of the eigenspace in Theorem 1.3 is not unique because it depends on the choice of β and γ. Using an argument in the same spirit as that of inertial-manifold theory, the dynamics near the roll solutions can be well approximated by the dynamics projected on this space [8].

When the domain is square (i.e., $L = M$), we determine a scaling parameter ν by

$$\nu = \frac{\lambda}{L} \left(= \frac{\lambda}{M} \right), \tag{15}$$

where λ is the basic wavelength of roll patterns. Recalling the Fourier series expansion of the solutions of the phase-diffusion equation in (8), we find that under the scaling (15), the dynamics given by (8) are the same as the dynamics on the above eigenspace, i.e., the dynamics defined by the following system of the ordinary differential equations:

$$\frac{da_{mn}}{dT} = -(D_\perp (m\omega)^2 + D_{//}(n\omega)^2) a_{mn},$$

where a_{mn} is the coefficient of the eigenfunction

$$\psi_{mn} = \partial_x u_0 \exp(im\omega Y) \exp(in\omega X) + O(1/M) + O(1/L)$$

and $X = \nu x$, $Y = \nu y$, and $T = \nu^2 t$. Thus we know that the phase-diffusion equation describes the dynamics near the roll solutions and that Theorem 1.3 gives an interpretation of the phase-diffusion equation (8) from the viewpoint of spectral analysis.

The organization of this paper is as follows. Section 2 is devoted to the proof of Theorems 1.1 and 1.2. Our strategy for the proof is as follows. We apply the separation of the variables to the eigenvalue problem corresponding to (13). The y-component of the eigenvalue problem is easily solved. In order to solve the x-component, we apply the Bloch transformation introduced by Collet and Eckmann [1] to study the one-dimensional case. This technique converts the eigenvalue problem in $L^2(-L/2, L/2)$ into the one in $L^2(0, \lambda)$. Next, we deal with the system of linear equations naturally induced by the Fourier decomposition of the eigenvalue problem in the same line of arguments as in [1]. At first glance, it seems to be difficult to solve our problem since the dimension of the system is infinite. However, our system, can be reduced to the three-dimensional system which consists of the Fourier components with wavenumbers $\pm\omega$ and 0. This is the most outstanding property of our system; it enables us to carry out the spectral analysis precisely. Section 3 contains several concluding remarks.

2. Proof of Theorems 1.1 and 1.2. We consider the following eigenvalue problem:

$$Aw = \mu w \quad \text{in} \quad L^2(\Omega), \tag{16}$$

i.e.,

$$\alpha w - (1 + \partial_x^2)^2 w - 3u_0^2 w - 2(1 + \partial_x^2)\partial_y^2 w - \partial_y^4 w = \mu w \quad \text{in} \quad L^2(\Omega) \tag{17}$$

with the periodic boundary conditions in (14). We apply the separation of variables to (17). Let

$$w(x, y) = v(x)\varphi(y).$$

Then it follows from (17) that

$$(\alpha v - (1 + \partial_x^2)^2 v - 3u_0^2 v - \mu v)\varphi = 2(1 + \partial_x^2) v \partial_y^2 \varphi + v \partial_y^4 \varphi. \tag{18}$$

Now we consider the following eigenvalue problem:

$$-\partial_y^2\varphi = \nu\varphi \quad \text{in} \quad L^2(-M/2, M/2) \tag{19}$$

with periodic boundary conditions

$$\partial_y^j\varphi(-M/2) = \partial_y^j\varphi(M/2) \qquad (j = 0, 1). \tag{20}$$

Noting $-\partial_y^4\varphi = \nu\partial_y^2\varphi$ by (19), it follows from (18) that

$$\alpha v - (1 + \partial_x^2)^2 v - 3u_0^2 v + 2\nu(1 + \partial_x^2)v - \nu^2 v = \mu v.$$

For $\nu \geq 0$, let A_ν be the linear operator which maps $L^2(-L/2, L/2)$ into itself defined by

$$A_\nu v = \alpha v - (1 + \partial_x^2)^2 v - 3u_0^2 v + 2\nu(1 + \partial_x^2)v - \nu^2 v \tag{21}$$

with periodic boundary conditions

$$\partial_x^j v(-L/2) = \partial_x^j v(L/2) \qquad (j = 0, 1, 2, 3). \tag{22}$$

Then we obtain the following eigenvalue problem:

$$A_\nu v = \mu v \quad \text{in} \quad L^2(-L/2, L/2) \tag{23}$$

with the periodic boundary conditions in (22). Since (21) is a self-adjoint operator which has a compact resolvent, we denote the eigenvalues of (23) by

$$\mu_n(\nu) \quad \text{for} \quad n = 0, \pm 1, \pm 2, \ldots$$

and the associated eigenfunctions by

$$v_n(\nu) \quad \text{for} \quad n = 0, \pm 1, \pm 2, \ldots.$$

On the other hand, the eigenvalue problem (19) is easily solved. We denote the eigenvalues of (19) by

$$\nu_l = \left(\frac{2\pi l}{M}\right)^2 \quad \text{for} \quad l = 0, \pm 1, \pm 2, \ldots \tag{24}$$

and the associated eigenfunctions by

$$\varphi_l = \exp(2\pi i l y/M) \quad \text{for} \quad l = 0, \pm 1, \pm 2, \ldots. \tag{25}$$

Hence the eigenvalues of (16) are given by

$$\mu_n(\nu_l) \quad \text{for} \quad l, n = 0, \pm 1, \pm 2, \ldots$$

and the associated eigenfunctions by

$$v_n(\nu_l)\varphi_l \quad \text{for} \quad l, n = 0, \pm 1, \pm 2, \ldots.$$

Since $\{v_n(\nu)\}_{n\in\mathbf{Z}}$ and $\{\varphi_l\}_{l\in\mathbf{Z}}$ are complete in $L^2(-L/2, L/2)$ and $L^2(-M/2, M/2)$, respectively, we know that $\{v_n(\nu_l)\varphi_l\}_{l,n\in\mathbf{Z}}$ is complete in $L^2(\Omega)$, namely, all the eigenvalues of (16) are given by $\mu = \mu_n(\nu_l)$. Therefore, we consider the eigenvalue problem (23).

In order to study the eigenvalue problem (23), we use the Bloch technique, which is available for the analysis of Schrödinger operators with periodic potentials. For more details, see Reed and Simon [13].

First, we consider a direct decomposition of the space $L^2(-L/2, L/2)$. For each $n \in [-N, N-1]$, let $X_n = L^2(0, \lambda)$. We denote by X the direct sum $\bigoplus_{n=-N}^{N-1} X_n$ equipped with the inner product

$$\langle x, x'\rangle_X = \sum_{n=-N}^{N-1} \langle x_n, x'_n\rangle_{L^2(0,\lambda)}$$
$$\text{for} \quad x = (x_n) \in X, \quad x_n \in X_n, \quad \text{and} \quad x' = (x'_n) \in X, \quad x'_n \in X_n,$$

and the norm $\| \; x \; \|_X^2 = \langle x, x\rangle_X$. The follwing lemma shows that $X = \bigoplus X_n$ can be identified with $L^2(-L/2, L/2)$.

LEMMA 2.1 (Kuwamura [8, Lem. 1]). *Let U be the linear operator which maps $L^2(-L/2, L/2)$ into X defined by*

$$Uf = ((Uf)_n), \quad (Uf)_n \in X_n,$$
$$(Uf)_n(x) = \frac{1}{\sqrt{2N}} \sum_{m=-N}^{N-1} f(x + \lambda m)e^{-i\kappa_n(x+\lambda m)} \quad \text{for} \quad f \in L^2(-L/2, L/2)$$

and U^* be the linear operator which maps X into $L^2(-L/2, L/2)$ defined by

$$(U^*g)(x + \lambda m) = \frac{1}{\sqrt{2N}} \sum_{n=-N}^{N-1} g_n(x)e^{i\kappa_n(x+\lambda m)}, \quad -N \leq m \leq N-1,$$
$$\text{for} \quad g = (g_n) \in X, \quad g_n \in X_n = L^2(0, \lambda),$$

where $x \in [0, \lambda]$ and $\kappa_n = \omega n/2N = (2\pi n/L)(-N \leq n \leq N-1)$. Then we have

$$\|Uf\|_X^2 = \|f\|_{L^2(-L/2,L/2)}^2,$$

$$\|U^*g\|_{L^2(-L/2,L/2)}^2 = \|g\|_X^2,$$

$$\langle Uf, g\rangle_X = \langle f, U^*g\rangle_{L^2(-L/2,L/2)},$$

$$U^*U = id_{L^2(-L/2,L/2)},$$

$$UU^* = id_X.$$

The operator $A : X \longrightarrow X$ is called decomposable if and only if there exists a family of operators $A_n : X_n \longrightarrow X_n$ such that for each $x = (x_n) \in X$,

$$(Ax)_n = A_n x_n$$

holds. Then we write

$$A = \bigoplus_n A_n.$$

The following theorem shows that the linear operator (21) is decomposable.

THEOREM 2.2. *The operator $A_\nu : L^2(-L/2, L/2) \longrightarrow L^2(-L/2, L/2)$ defined by* (21) *with the periodic boundary conditions in* (22) *is decomposable. More precisely, let $A_{\nu,\kappa_n} : L^2(0,\lambda) \longrightarrow L^2(0,\lambda)$ be the family of operators defined by*

$$A_{\nu,\kappa_n}\psi = \alpha\psi - (1-(\kappa_n - i\partial_x)^2)^2\psi - 3u_0^2\psi + 2\nu(1-(\kappa_n - i\partial_x)^2)\psi - \nu^2\psi, \quad \psi \in H^4(0,\lambda), \tag{26}$$

with periodic boundary conditions

$$\partial_x^j\psi(0) = \partial_x^j\psi(\lambda) \qquad (j = 0,1,2,3), \tag{27}$$

where

$$\kappa_n = \frac{\omega n}{2N} = \frac{2\pi n}{L}, \qquad n \in [-N, N-1]. \tag{28}$$

Then

$$UA_\nu U^* = \bigoplus_n A_{\nu,\kappa_n} \tag{29}$$

holds, where $U^* : X \longrightarrow L^2(-L/2,L/2)$ and $U : L^2(-L/2,L/2) \longrightarrow X$ are given in Lemma 2.1.

Remark. We can obtain (26) as follows: for each $\psi \in L^2(0,\lambda)$, let

$$v(x) = e^{i\kappa_n x}\psi(x), \tag{30}$$

where κ_n is given by (28). Notice that v satisfies the periodic boundary conditions in (22) if and only if ψ satisfies the periodic boundary conditions in (27). Substituting (30) into (23), we have the following eigenvalue problem:

$$A_{\nu,\kappa_n}\psi = \mu\psi \quad \text{in} \quad L^2(0,\lambda) \tag{31}$$

with the periodic boundary conditions (27), where A_{ν,κ_n} is defined by (26). Notice that A_{ν,κ_n} is self-adjoint. The technique which transforms the eigenvalue problem (23) into (31) is called the Bloch technique. The relation (29) is essential for understanding the mathematical background of this technique.

We find that v is an eigenvector of A_ν associated to an eigenvalue μ if and only if there exists some κ_n such that $v(x) = e^{i\kappa_n x}\psi(x)$ holds, where ψ is an eigenvector of A_{ν,κ_n} associated to the eigenvalue μ. Therefore, we consider the eigenvalue problem (31) instead of (23).

Theorem 2.2 can be proved by similar argument to [8, Thm. 2]. Theorems 1.1 and 1.2 are the direct consequences of the following.

THEOREM 2.3. (A) *When $0 \le W < 1/\sqrt{2}$, for sufficiently small ε and each $n \in [-N, N-1]$ and $\nu = \nu_m \ge 0$, all the eigenvalues of A_{ν_m,κ_n} are negative except for the simple zero eigenvalue of A_{ν_0,κ_0}.*

(B) *When $-1/\sqrt{2} < W < 0$, for sufficiently small ε, we have the following:*

(i) *If* $0 < \varepsilon < 2\pi^2/\sqrt{3}|W|M^2$, *then for each* $n \in [-N, N-1]$ *and* $\nu = \nu_m \geq 0$, *all the eigenvalues of* A_{ν_m,κ_n} *are negative except for the simple zero eigenvalue of* A_{ν_0,κ_0}. *Here* $\varepsilon M^2 = O(1)$ *as* $\varepsilon \downarrow 0$ *and* $M \to \infty$.

(ii) *If* $\varepsilon > 2\pi^2/\sqrt{3}|W|M^2$, *then* $A_{\nu_{\pm 1},\kappa_0}$ *has a positive eigenvalue. Here,* $\varepsilon M^2 = O(1)$ *as* $\varepsilon \downarrow 0$ *and* $M \to \infty$.

(C) *When* $|W| > 1/\sqrt{2}$, *for sufficiently small* $\varepsilon > 0$ *and large* L, *there exists some* $n \in [-N, N-1]$ *such that* A_{ν_0,κ_n} *has a positive eigenvalue.*

THEOREM 2.4. *When* L *and* M *are sufficiently large, under the conditions of Theorem* 2.3(A), *there exists* $\rho > 0$ *independent of* ε, W, L, *and* M *such that the following hold:*

(i) *For each* $\kappa_n \in [-\sqrt{3}\rho\varepsilon/2, \sqrt{3}\rho\varepsilon/2]$ *and* $\nu_m \in [0, \sqrt{3}\rho\varepsilon]$, *the principal eigenvalue of* A_{ν_m,κ_n}*—say* μ_{mn}*—satisfies*

$$\mu_{mn} = -D_{\perp}\nu_m - \nu_m^2 - D_{//}\kappa_n^2 + O((\kappa_n + \nu_m)^3),$$

where

$$D_{//} = 4 - 8W^2 + O(\varepsilon),$$
$$D_{\perp} = 2\sqrt{3}\varepsilon W,$$

and

$$\nu_m = \left(\frac{2\pi m}{M}\right)^2,$$
$$\kappa_n = \frac{2\pi n}{L},$$

and the eigenfunction associated to the eigenvalue μ_{mn}*—say* ψ_{mn}*—satisfies*

$$\psi_{mn} = \partial_x u_0 + O(\kappa_n) + O(\nu_m).$$

The other eigenvalues of A_{ν_m,κ_n} *satisfy*

$$\mu < -\delta_1,$$

where $\delta_1 > 0$ *depends only on* ε *and* W.

(ii) *For each* $\kappa_n \notin [-\sqrt{3}\rho\varepsilon/2, \sqrt{3}\rho\varepsilon/2]$ *or* $\nu_m \notin [0, \sqrt{3}\rho\varepsilon]$, *all the eigenvalues of* A_{ν_m,κ_n} *satisfy*

$$\mu < -\delta_2,$$

where $\delta_2 > 0$ *depends only on* ε *and* W.

Proof of Theorems 2.3 *and* 2.4. We consider the eigenvalue problem (31), i.e.,

$$A_{\nu,\kappa_n}\psi = \mu\psi \quad \text{in} \quad L^2(0, \lambda)$$

with the periodic boundary conditions in (27), where

$$A_{\nu,\kappa_n}\psi = \alpha\psi - (1 - (\kappa_n - i\partial_x)^2)^2\psi - 3u_0^2\psi + 2\nu(1 - (\kappa_n - i\partial_x)^2)\psi - \nu^2\psi.$$

We decompose ψ and u_0 into their Fourier components

$$\psi(x) = \sum_m a_m e^{im\omega x}$$

and

$$u_0(x) = \sum_m b_m e^{im\omega x},$$

respectively. By (4) and (5), we know that

$$b_m = \begin{cases} \varepsilon & \text{for} \quad |m| = 1, \\ O(\varepsilon^{1+2|m|/3}) & \text{for} \quad |m| \geq 3, m : \text{odd}, \\ 0 & \text{for} \quad m : \text{even}. \end{cases} \tag{32}$$

Since $\psi \in L^2(0, \lambda)$ is naturally identfied with $(a_m)_{m \in \mathbf{Z}} \in l^2$, (31) is equivalent to the following system of equations:

$$\begin{aligned} &\alpha a_m - (1 - (\kappa_n + m\omega)^2)^2 a_m + 2\nu(1 - (\kappa_n + m\omega)^2) a_m \\ &\quad - \nu^2 a_m - \mu a_m - T_0 a_m = \sum_{r \neq m} T_{m-r} a_r \quad \text{for} \quad m \in \mathbf{Z}, \end{aligned} \tag{33}$$

where

$$T_r = 3 \sum_{p+q=r} b_p b_q.$$

By (32), we have

$$T_m = \begin{cases} 6\varepsilon^2 + O(\varepsilon^6) & \text{for} \quad m = 0, \\ 3\varepsilon^2 + O(\varepsilon^4) & \text{for} \quad |m| = 2, \\ O(|m|\varepsilon^{2(|m|+2)/3}) & \text{for} \quad |m| \geq 4, m : \text{even}, \\ 0 & \text{for} \quad m : \text{odd}. \end{cases} \tag{34}$$

Let

$$\begin{aligned} S_m &= \alpha - (1 - (\kappa_n + m\omega)^2)^2 + 2\nu(1 - (\kappa_n + m\omega)^2) - \nu^2 - T_0 - \mu \\ &= \alpha - \{1 - ((\kappa_n + m\omega)^2 + \nu)\}^2 - T_0 - \mu \end{aligned} \tag{35}$$

and

$$B_{m,j}(\psi) = \sum_{|r-m| \geq j} T_{m-r} a_r.$$

The system of equations (33) is rewritten as follows:

$$S_m a_m = B_{m,1}(\psi) \quad \text{for} \quad m \in \mathbf{Z}. \tag{36}$$

Notice that (36) has a trivial solution $\psi \equiv 0$. We will study (36) as follows. We solve the system of equations

$$S_m a_m = B_{m,1}(\psi) \quad \text{for} \quad |m| \geq 2. \tag{37}$$

for $(\ldots, a_{-3}, a_{-2}, a_2, a_3, \ldots) \in l^2$ as a function of a_{-1}, a_0 and a_1. (37) is solved by using the contraction-mapping theorem. We denote the solution of (37) by

$$a_m = a_m(a_{-1}, a_0, a_1), \quad \text{for} \quad |m| \geq 2. \tag{38}$$

Substituting (38) into

$$\left\{ \begin{aligned} S_{-1}a_{-1} &= B_{-1,1}(\psi), \\ S_0 a_0 &= B_{0,1}(\psi), \\ S_1 a_1 &= B_{1,1}(\psi), \end{aligned} \right. \tag{39}$$

we solve (39) for (a_{-1}, a_0, a_1). Now, we start to solve (37).

The following lemma is the key to reduce the infinite-dimensional system of linear equations (36) to the three-dimensional system (39).

LEMMA 2.5. *For $\mu \geq -1/4$ and sufficiently small $\varepsilon > 0$, the system of equations* (37) *has a unique solution of the following form:*

$$a_m = S_m^{-1} R_m a_0 \quad \text{for} \quad m: \text{ even} \tag{40}$$

and

$$a_m = S_m^{-1} R'_m a_{-1} + S_m^{-1} R''_m a_1 \quad \text{for} \quad m: \text{ odd}, \tag{41}$$

where

$$\begin{aligned} R_{\pm 2} &= T_{\pm 2} + O(\varepsilon^4), \\ R'_{-3} &= T_{-2} + O(\varepsilon^4), \\ R''_{-3} &= O(\varepsilon^4), \\ R'_3 &= O(\varepsilon^4), \\ R''_3 &= T_2 + O(\varepsilon^4), \\ R_m &= O(|m|\varepsilon^{2(|m|+2)/3}) \quad \text{for} \quad |m| \geq 4,\ m: \text{even}, \\ R'_m &= O(|m|\varepsilon^{2(|m|+1)/3}) \quad \text{for} \quad |m| \geq 4,\ m: \text{odd}, \\ R''_m &= O(|m|\varepsilon^{2(|m|+1)/3}) \quad \text{for} \quad |m| \geq 4,\ m: \text{odd}, \end{aligned}$$

Proof. Let $\phi = (\ldots, a_{-3}, a_{-2}, a_2, a_3, \ldots) \in l^2$. Then the system of equations (37) is rewritten as follows:

$$a_m = S_m^{-1} q_m + S_m^{-1} p_m(\phi) \quad \text{for} \quad |m| \geq 2, \tag{42}$$

where q_m is an inhomogeneous term, i.e.,

$$q_m = T_{m+1} a_{-1} + T_m a_0 + T_{m-1} a_1,$$

and $p_m(\phi)$ is a linear operator with respect to ϕ, i.e.,

$$p_m(\phi) = \sum_{|j| \geq 2, j \neq m} T_{m-j} a_j.$$

Noting (2), (12), and

$$\nu \geq 0, \quad |\kappa_n| = \left| \frac{\omega n}{2N} \right| \leq \omega/2,$$

we find

$$\begin{aligned} S_m &= \alpha - (1-(\kappa_n+m\omega)^2)^2 + 2\nu(1-(\kappa_n+m\omega)^2) - \nu^2 - T_0 - \mu \\ &\leq -Cm^4 \end{aligned} \tag{43}$$

for $\mu \geq -1/4$ and $|m| \geq 2$. Moreover, by (34), we have

$$|p_m(\phi)| \leq \left| \sum_{|j|\geq 2, j\neq m} T_{m-j}a_j \right| \leq C\varepsilon^2 \|\phi\|,$$

where

$$\|\phi\| = \left(\sum_{|j|\geq 2} |a_j|^2 \right)^{1/2} .$$

Therefore, (42) is solved by using the contraction-mapping theorem. In fact, the successive approximation starting from $\phi \equiv 0$ can be applied to (42). Then we have

$$a_m = S_m^{-1} R_m a_0 \quad \text{for} \quad m: \text{ even} \tag{44}$$

and

$$a_m = S_m^{-1} R_m' a_{-1} + S_m^{-1} R_m'' a_1 \quad \text{for} \quad m: \text{odd}, \tag{45}$$

where

$$\begin{aligned} R_m = T_m \quad &+ \sum_{|j|\geq 2, j\neq m} T_{m-j}T_jS_j^{-1} \\ &+ \sum_{|j|,|k|\geq 2, j\neq m, k\neq j} T_{m-j}T_{j-k}T_kS_j^{-1}S_k^{-1} \\ &+ \cdots, \end{aligned}$$

$$\begin{aligned} R_m' = T_{m+1} \quad &+ \sum_{|j|\geq 2, j\neq m} T_{m-j}T_{j+1}S_j^{-1} \\ &+ \sum_{|j|,|k|\geq 2, j\neq m, k\neq j} T_{m-j}T_{j-k}T_{k+1}S_j^{-1}S_k^{-1} \\ &+ \cdots, \end{aligned}$$

and

$$\begin{aligned} R_m'' = T_{m-1} \quad &+ \sum_{|j|\geq 2, j\neq m} T_{m-j}T_{j-1}S_j^{-1} \\ &+ \sum_{|j|,|k|\geq 2, j\neq m, k\neq j} T_{m-j}T_{j-k}T_{k-1}S_j^{-1}S_k^{-1} \\ &+ \cdots. \end{aligned}$$

Therefore, (40) and (41) follow from (34) and (43)–(45). □

Now we substitute (40) and (41) into (39). By (34) and (45), we have

$$\begin{aligned} B_{-1,1}(\psi) &= \textstyle\sum_{|r+1|\geq 1} T_{-1-r}a_r \\ &=: P_{-1}a_{-1} + P_1a_1, \end{aligned}$$

where

$$\begin{aligned} P_{-1} &= \textstyle\sum_{|r|\geq 2, r:\mathrm{odd}} T_{-1-r} S_r^{-1} R_r' \\ &= T_{-2} S_{-3}^{-1} T_2 + O(\varepsilon^6) \end{aligned} \tag{46}$$

and

$$\begin{aligned} P_1 &= T_{-2} + \textstyle\sum_{|r|\geq 2, r:\mathrm{odd}} T_{-1-r} S_r^{-1} R_r'' \\ &= T_{-2} + O(\varepsilon^6). \end{aligned} \tag{47}$$

Similarly, we have

$$B_{0,1}(\psi) = V_0 a_0$$

and

$$B_{-1,1}(\psi) = Q_{-1} a_{-1} + Q_1 a_1,$$

where

$$V_0 = T_{-2} S_{-2}^{-1} T_2 + T_{-2} S_2^{-1} T_2 + O(\varepsilon^6), \tag{48}$$

$$Q_{-1} = T_2 + O(\varepsilon^6), \tag{49}$$

and

$$Q_1 = T_{-2} S_3^{-1} T_2 + O(\varepsilon^6). \tag{50}$$

Thus we have the following equations:

$$(S_0 - V_0) a_0 = 0 \tag{51}$$

and

$$\begin{pmatrix} S_{-1} - P_{-1} & -P_1 \\ -Q_{-1} & S_1 - Q_1 \end{pmatrix} \begin{pmatrix} a_{-1} \\ a_1 \end{pmatrix} = 0. \tag{52}$$

First, we may ask whether equation (51) has a nontrivial solution. Suppose that (51) has a nontrivial solution. Then it follows from (35), (48), and (51) that

$$\alpha - \{1 - (\kappa_n^2 + \nu)\}^2 - T_0 - \mu - T_{-2} S_{-2}^{-1} T_2 - T_{-2} S_2^{-1} T_2 + O(\varepsilon^6) = 0.$$

Noting (2), (12), and (34), we have

$$\mu = -3\varepsilon^2 (1 - W^2) - \{1 - (\kappa_n^2 + \nu)\}^2 + O(\varepsilon^4).$$

Hence we see that when $|W| < 1/\sqrt{2}$, for sufficiently small ε,

$$\mu < -\varepsilon^2 (1 - W^2)$$

holds. Thus when $|W| < 1/\sqrt{2}$, there exists $\delta_0 > 0$ independent of ε such that for $\mu > -\delta_0 \varepsilon^2$, the equation (51) has no solution other than the trivial solution.

Next, we may ask whether equation (52) has a nontrivial solution. Noting (2), (12), (34), and (35), we have

$$\begin{pmatrix} S_{-1}-P_{-1} & -P_1 \\ -Q_{-1} & S_1-Q_1 \end{pmatrix} =$$

$$\begin{pmatrix} -3\varepsilon^2(1-W^2)-\{1-((\kappa_n-\omega)^2+\nu)\}^2 & -3\varepsilon^2 \\ -3\varepsilon^2 & -3\varepsilon^2(1-W^2)-\{1-((\kappa_n+\omega)^2+\nu)\}^2 \end{pmatrix}$$
$$-\mu I + O(\varepsilon^4).$$

Hence we consider the eigenvalues of

$$B = \begin{pmatrix} -3\varepsilon^2(1-W^2)-\{1-((\kappa_n-\omega)^2+\nu)\}^2 & -3\varepsilon^2 \\ -3\varepsilon^2 & -3\varepsilon^2(1-W^2)-\{1-((\kappa_n+\omega)^2+\nu)\}^2 \end{pmatrix}.$$

In order to investigate B, it is useful to parametrize ν and κ_n as follows:

$$\nu = \sqrt{3}\varepsilon\nu' \quad \text{for} \quad \nu' \geq 0 \tag{53}$$

and

$$\kappa_n = \sqrt{3}\varepsilon K_n/2. \tag{54}$$

Notice that

$$|K_n| \leq \frac{\omega}{\sqrt{3}\varepsilon}$$

by $|\kappa_n| \leq \omega/2$. Moreover, we set

$$x = \nu' + \sqrt{3}\varepsilon K_n^2/4 \geq 0 \tag{55}$$

and

$$y = \omega K_n. \tag{56}$$

Then, it follows from (53)–(56) and (12) that

$$\{1-((\omega-\kappa_n)^2+\nu)\}^2 = 3\varepsilon^2(W-y+x)^2$$

and

$$\{1-((\omega+\kappa_n)^2+\nu)\}^2 = 3\varepsilon^2(W+y+x)^2.$$

Hence we have

$$\begin{aligned} &-3\varepsilon^2(1-W^2)-\{1-((\kappa_n-\omega)^2+\nu)\}^2 \\ &\quad = -3\varepsilon^2(1+y^2+x^2+2Wx-2Wy-2yx) \end{aligned}$$

and

$$\begin{aligned} &-3\varepsilon^2(1-W^2)-\{1-((\kappa_n+\omega)^2+\nu)\}^2 \\ &\quad = -3\varepsilon^2(1+y^2+x^2+2Wx+2Wy+2yx). \end{aligned}$$

Therefore, the characteristic equation of the matrix $(1/3\varepsilon^2)B$ is given by

$$\begin{aligned}&\lambda^2 + 2(1 + y^2 + x^2 + 2Wx)\lambda \\ &\quad + (1 + y^2 + x^2 + 2Wx)^2 - (2Wy + 2yx)^2 - 1 = 0.\end{aligned}$$

We find that

$$\begin{aligned}&-(1 + y^2 + x^2 + 2Wx) \\ &\quad = -1 + W^2 - y^2 - (x + W)^2 < 0\end{aligned}$$

provided $|W| < 1/\sqrt{2}$. Moreover, we know that

$$\begin{aligned}&(1 + y^2 + x^2 + 2Wx)^2 - (2Wy + 2yx)^2 - 1 \\ &\quad = 2y^2 + 2x^2 - 4W^2(x + W)^2 + 4Wx + \{(x^2 - y^2) + 2W(x + W)\}^2 \\ &\quad \geq \omega^2 + 2x^2 - 4W^2(x + W)^2 + 4Wx\end{aligned}$$

holds provided $|K_n| \geq 1/\sqrt{2}$. Let

$$F(x) = \omega^2 + 2x^2 - 4W^2(x + W)^2 + 4Wx.$$

It is easy to check that when $|W| < 1/\sqrt{2}$,

$$\begin{aligned}F(x) &\geq \omega^2 - 2W^2 \\ &= 1 - 2W^2 + \sqrt{3}\varepsilon W > 0\end{aligned}$$

holds for sufficiently small $\varepsilon > 0$. Therefore, we find that when $|K_n| \geq 1/\sqrt{2}$ and $|W| < 1/\sqrt{2}$,

$$(1 + y^2 + x^2 + 2Wx)^2 - (2Wy + 2yx)^2 - 1 > 0$$

holds for sufficiently small $\varepsilon > 0$. Hence we know that the eigenvalues of $(1/3\varepsilon^2)B$ are negative. Thus there exists $\delta_1 > 0$ independent of ε such that when $|K_n| \geq 1/\sqrt{2}$ and $|W| < 1/\sqrt{2}$, the matrix $B - \mu I$ has no zero eigenvalue, i.e., $|B - \mu I| \neq 0$ for sufficiently small ε and $\mu > -\delta_1\varepsilon^2$. Thus we find that when $|K_n| \geq 1/\sqrt{2}$ and $|W| < 1/\sqrt{2}$,

$$\begin{vmatrix} S_{-1} - P_{-1} & -P_1 \\ -Q_{-1} & S_1 - Q_1 \end{vmatrix} \neq 0$$

holds for $\mu > -\delta_1\varepsilon^2$ and sufficiently small ε, so equation (52) has no solution other than the trivial solution.

Now we consider the case $|K_n| < 1/\sqrt{2}$. In this case, it follows from (53), (54), and (12) that

$$\{1 - ((\omega - \kappa_n)^2 + \nu)\}^2 = 3\varepsilon^2(W - K_n + \nu')^2 + O(\varepsilon^3)$$

and

$$\{1 - ((\omega + \kappa_n)^2 + \nu)\}^2 = 3\varepsilon^2(W + K_n + \nu')^2 + O(\varepsilon^3).$$

Hence we have

$$B = 3\varepsilon^2 \begin{pmatrix} W^2 - 1 - (W - K_n + \nu')^2 & -1 \\ -1 & W^2 - 1 - (W + K_n + \nu')^2 \end{pmatrix} + O(\varepsilon^3)$$

for sufficiently small ε. We consider the eigenvalues of

$$B' = \begin{pmatrix} W^2 - 1 - (W - K_n + \nu')^2 & -1 \\ -1 & W^2 - 1 - (W + K_n + \nu')^2 \end{pmatrix}.$$

We find that the characteristic equation of B' is

$$\begin{aligned}&\lambda^2 + 2(1 + K_n^2 + \nu'^2 + 2W\nu')\lambda \\ &\quad + (1 + K_n^2 + \nu'^2 + 2W\nu')^2 - 4K_n^2(W + \nu')^2 - 1 = 0,\end{aligned}$$

so the eigenvalues of B' are given by

$$\lambda_- = -1 - \nu'^2 - K_n^2 - 2W\nu' - \sqrt{1 + 4K_n^2(W + \nu')^2}$$

and

$$\lambda_+ = -1 - \nu'^2 - K_n^2 - 2W\nu' + \sqrt{1 + 4K_n^2(W + \nu')^2}.$$

We find that when $|W| < 1/\sqrt{2}$,

$$\lambda_- \;=\; -1 + W^2 - (\nu' + W)^2 - K_n^2 - \sqrt{1 + 4K_n^2(W + \nu')^2} \;<\; -1.$$

Using the inequality

$$\sqrt{1 + x} \leq 1 + x/2 \quad \text{for} \quad x \geq 0,$$

we obtain

$$\begin{aligned}\lambda_+ &\leq -1 - \nu'^2 - K_n^2 - 2W\nu' + 1 + 2K_n^2(W + \nu')^2 \\ &= -K_n^2(1 - 2W^2) - (\nu'^2 + 2W\nu')(1 - 2K_n^2).\end{aligned} \tag{57}$$

First, we consider the case $-1/\sqrt{2} < W < 0$. In this case, we know that

$$\nu'^2 + 2W\nu' < 0 \quad \text{for} \quad 0 < \nu' < -2W$$

and

$$\nu'^2 + 2W\nu' > 0 \quad \text{for} \quad \nu' > -2W.$$

By (53) and (24), we recall that

$$\nu'_m = \frac{\nu_m}{\sqrt{3}\varepsilon} = \frac{1}{\sqrt{3}\varepsilon}\left(\frac{2\pi m}{M}\right)^2.$$

Hence if

$$\frac{1}{\sqrt{3}\varepsilon}\left(\frac{2\pi}{M}\right)^2 > -2W,$$

i.e.,

$$0 < \varepsilon < \frac{2\pi^2}{\sqrt{3}|W|M^2},$$

then there exists $\delta_2 > 0$ independent of ε and M such that

$$\lambda_+ < -\delta_2$$

holds provided $\varepsilon M^2 = O(1)$ as $\varepsilon \downarrow 0$ and $M \to \infty$.

On the other hand, if

$$\frac{1}{\sqrt{3}\varepsilon}\left(\frac{2\pi}{M}\right)^2 < -2W,$$

i.e.,

$$\varepsilon > \frac{2\pi^2}{\sqrt{3}|W|M^2},$$

then there exists $\delta_3 > 0$ independent of ε and M such that for $K_n = 0$ and $\nu' = \nu'_{\pm 1}$,

$$\lambda_+ = -(\nu'^2 + 2W\nu') > \delta_3$$

holds provided $\varepsilon M^2 = O(1)$ as $\varepsilon \downarrow 0$ and $M \to \infty$.

Now we consider the case when $0 \le W < 1/\sqrt{2}$. Noting (57), we can choose $\rho > 0$ independent of ε, which will be spcified later such that

$$\lambda_+ \le -\min(\rho^2(1-2W^2), (\rho^2 + 2W\rho)/2)$$

holds for $|K_n| \ge \rho$ or $\nu' \ge \rho$. Hence there exists $\delta_4 > 0$ independent of ε such that when $|K_n| \ge \rho$ or $\nu' \ge \rho$, the matrix $B - \mu I$ has no zero eigenvalue, i.e., $|B - \mu I| \ne 0$ for $\mu > -\delta_4\varepsilon^2$ and sufficiently small ε. Thus we find that when $0 \le W < 1/\sqrt{2}$ and $\max(|K_n|, \nu') \ge \rho$,

$$\begin{vmatrix} S_{-1} - P_{-1} & -P_1 \\ -Q_{-1} & S_1 - Q_1 \end{vmatrix} \ne 0$$

holds for $\mu > -\delta_4\varepsilon^2$ and sufficiently small ε, so the equation (52) has no solution other than the trivial solution.

Finally, we consider the case where $0 \le W < 1/\sqrt{2}$, $|K_n| \le \rho$, and $0 \le \nu' \le \rho$. By (35), (46), (47), (49), and (50), we have

$$\begin{aligned}
&\begin{pmatrix} S_{-1} - P_{-1} & -P_1 \\ -Q_{-1} & S_1 - Q_1 \end{pmatrix} = \\
&\begin{pmatrix} \alpha - T_0 - \mu - \{1 - ((\kappa_n - \omega)^2 + \nu)\}^2 & -3\varepsilon^2 \\ -3\varepsilon^2 & \alpha - T_0 - \mu - \{1 - ((\kappa_n + \omega)^2 + \nu)\}^2 \end{pmatrix} \\
&\qquad +O(\varepsilon^4) \\
&\qquad =: B'' + O(\varepsilon^4)
\end{aligned}$$

for sufficiently small $\varepsilon > 0$. Noting

$$|\kappa_n| \le \sqrt{3}\rho\varepsilon/2 \tag{58}$$

and

$$0 \le \nu \le \sqrt{3}\rho\varepsilon, \tag{59}$$

it follows from (2) and (34) that the eigenvalue of B'' is given by

$$\lambda_\pm = -3\varepsilon^2 \pm 3\varepsilon^2 - \mu + O((\kappa_n + \nu)^2) + o(\varepsilon^2)$$

for sufficiently small ε. Hence we can choose an appropriate $\rho > 0$ independent of ε such that

$$|\lambda_+| > \varepsilon^2/4 \quad \text{for} \quad |\mu| > \varepsilon^2/2$$

and

$$|\lambda_-| > \varepsilon^2/4 \quad \text{for} \quad |\mu + 6\varepsilon^2| > \varepsilon^2/2.$$

Therefore, we find that for $|\mu| > \varepsilon^2/2$ and $|\mu + 6\varepsilon^2| > \varepsilon^2/2$,

$$\begin{vmatrix} S_{-1} - P_{-1} & -P_1 \\ -Q_{-1} & S_1 - Q_1 \end{vmatrix} \neq 0$$

holds, so equation (52) has no solution other than the trivial solution.

Finally, we consider the case $|\mu| < \varepsilon^2/2$ by using analytic-perturbation theory (Kato [5]). Noting (58) and (59), we expand the operator

$$A_{\nu_m,\kappa_n} = \alpha - (1 - (\kappa_n - i\partial_x)^2)^2 - 3u_0^2 + 2\nu_m(1 - (\kappa_n - i\partial_x)^2) - \nu_m^2 \quad \text{in} \quad L^2(0,\lambda)$$

with the periodic boundary conditions in (27) with respect to ν_m and κ_n near $\nu_m = \kappa_n = 0$. In what follows, for simplicity, we suppress the subscripts m and n of ν_m and κ_n, respectively. $A_{\nu,\kappa}$ is expanded in powers of ν and κ as follows:

$$A_{\nu,\kappa} = \sum_{m,n=0}^{4} A^{(m,n)} \nu^m \kappa^n,$$

where

$$\begin{aligned}
A^{(0,0)} &= \alpha - (1 + \partial_x^2)^2 - 3u_0^2, \\
A^{(0,1)} &= -4i(\partial_x + \partial_x^3), \quad A^{(1,0)} = 2(1 + \partial_x^2), \\
A^{(0,2)} &= 2 + 6\partial_x^2, \quad A^{(1,1)} = 4i\partial_x, \quad A^{(2,0)} = -1, \\
A^{(0,3)} &= 4i\partial_x, \quad A^{(1,2)} = -2, \quad A^{(2,1)} = A^{(3,0)} = 0, \\
A^{(0,4)} &= -1, \quad A^{(1,3)} = A^{(2,2)} = A^{(3,1)} = A^{(4,0)} = 0.
\end{aligned}$$

Since $A^{(m,n)}$ is relatively bounded with respect to $A^{(0,0)}$—that is, $\|A^{(m,n)}f\| \leq C(\|A^{(0,0)}f\| + \|f\|)$ holds for some $C > 0$—we can apply analytic-perturbation theory [5, Chap. 7, Thm. 2.6, and Remark 2.7].

We denote by $\mu_{\nu,\kappa}$ the pricipal eigenvalue of $A_{\nu,\kappa}$ and by $\psi_{\nu,\kappa}$ the associated eigenfunction of $\mu_{\nu,\kappa}$. Here we notice that $\mu_{0,0}$ is the simple zero eigenvalue of $A_{0,0}$, and its associated eigenfunction is $\psi_{0,0} = \partial_x u_0 = -2\varepsilon\omega \sin \omega x + O(\varepsilon^3)$. In fact, in a manner similar to the previous argument, we see that $A_{0,0}\psi = 0$ has no solution other than the trivial solution in $\langle \partial_x u_0 \rangle^\perp = \{v \in L^2(0,\lambda); \langle v, \partial_x u_0 \rangle = 0\}$. Hence the geometric multiplicity of $\mu_{0,0}$ is equal to one. Moreover, noting that $A_{0,0}$ is self-adjoint, we find that

$$R(A_{0,0}) = N(A_{0,0}^*)^\perp = N(A_{0,0})^\perp = \langle \partial_x u_0 \rangle^\perp$$

by the closed-range theorem, where R and N denote the range space and null space, respectively. Hence we see that $\partial_x u_0 \notin R(A_{0,0})$, so the algebraic multiplicity of $\mu_{0,0}$ is also equal to one.

Let $\mu_{\nu,\kappa}$ and $\psi_{\nu,\kappa}$ be expanded in powers of ν and κ as follows:

$$\mu_{\nu,\kappa} = \sum_{m,n=0}^{\infty} \mu^{(m,n)} \nu^m \kappa^n$$

and

$$\psi_{\nu,\kappa} = \sum_{m,n=0}^{\infty} \psi^{(m,n)} \nu^m \kappa^n.$$

Substituting this into $A_{\nu,\kappa}\psi_{\nu,\kappa} = \mu_{\nu,\kappa}\psi_{\nu,\kappa}$ and comparing the coefficient of each power of ν and κ, we obtain

$$A^{(0,0)}\psi^{(0,0)} = \mu^{(0,0)}\psi^{(0,0)}, \tag{60}$$

$$A^{(0,0)}\psi^{(0,1)} + A^{(0,1)}\psi^{(0,0)} = \mu^{(0,0)}\psi^{(0,1)} + \mu^{(0,1)}\psi^{(0,0)}, \tag{61}$$

$$A^{(0,0)}\psi^{(1,0)} + A^{(1,0)}\psi^{(0,0)} = \mu^{(0,0)}\psi^{(1,0)} + \mu^{(1,0)}\psi^{(0,0)}, \tag{62}$$

$$\begin{aligned} &A^{(0,0)}\psi^{(0,2)} + A^{(0,1)}\psi^{(0,1)} + A^{(0,2)}\psi^{(0,0)} \\ &= \mu^{(0,0)}\psi^{(0,2)} + \mu^{(0,1)}\psi^{(0,1)} + \mu^{(0,2)}\psi^{(0,0)}, \end{aligned} \tag{63}$$

$$\begin{aligned} &A^{(1,1)}\psi^{(0,0)} + A^{(1,0)}\psi^{(0,1)} + A^{(0,1)}\psi^{(1,0)} + A^{(0,0)}\psi^{(1,1)} \\ &= \mu^{(0,0)}\psi^{(1,1)} + \mu^{(0,1)}\psi^{(1,0)} + \mu^{(1,0)}\psi^{(0,1)} + \mu^{(1,1)}\psi^{(0,0)}, \end{aligned} \tag{64}$$

$$\begin{aligned} &A^{(0,0)}\psi^{(2,0)} + A^{(1,0)}\psi^{(1,0)} + A^{(2,0)}\psi^{(0,0)} \\ &= \mu^{(0,0)}\psi^{(2,0)} + \mu^{(1,0)}\psi^{(1,0)} + \mu^{(2,0)}\psi^{(0,0)}. \end{aligned} \tag{65}$$

Since $A^{(0,0)}$ is the linearized operator of the right-hand side of (6) at u_0, it follows from (60) that

$$\mu^{(0,0)} = 0 \tag{66}$$

and

$$\psi^{(0,0)} = \partial_x u_0 = -2\varepsilon\omega \sin \omega x + O(\varepsilon^3). \tag{67}$$

Hence, by (61) and (66), we have

$$A^{(0,0)}\psi^{(0,1)} = -A^{(0,1)}\psi^{(0,0)} + \mu^{(0,1)}\psi^{(0,0)}. \tag{68}$$

By the solvability condition for $\psi^{(0,1)}$, we have

$$\mu^{(0,1)} = \frac{\langle \psi^{(0,0)}, A^{(0,1)}\psi^{(0,0)} \rangle}{\langle \psi^{(0,0)}, \psi^{(0,0)} \rangle},$$

where $\langle \, , \rangle$ stands for the inner product in $L^2(0, \lambda)$. Noting (67) and the fact that $A^{(0,1)}$ is an odd-order differential operator, we see that

$$\langle \psi^{(0,0)}, A^{(0,1)}\psi^{(0,0)} \rangle = 0,$$

which leads to

$$\mu^{(0,1)} = 0. \tag{69}$$

Hence we can solve (68), i.e.,

$$A^{(0,0)}\psi^{(0,1)} = -A^{(0,1)}\psi^{(0,0)}. \tag{70}$$

Using (12) and (67), we decompose $\psi^{(0,1)}$ and $A^{(0,1)}\psi^{(0,0)}$ into their Fourier components as follows:

$$\psi^{(0,1)} = \sum_m c_m e^{im\omega x}, \quad \overline{c_m} = -c_{-m}$$

and

$$\begin{aligned} A^{(0,1)}\psi^{(0,0)} &= 4i\varepsilon(\omega^2 - \omega^4)(e^{i\omega x} + e^{-i\omega x}) + O(\varepsilon^3) \\ &= -4i\sqrt{3}W\varepsilon^2(e^{i\omega x} + e^{-i\omega x}) + O(\varepsilon^3). \end{aligned}$$

To solve equation (70), we must repeat the same line of arguments as applied to (33). Here we calculate the essential part of equation (70). Recalling (4), we have

$$\begin{aligned} A^{(0,0)}\psi^{(0,1)} &= (\alpha - (1 + \partial_x^2)^2 - 3u_0^2)\psi^{(0,1)} \\ &= \sum_m (\alpha - (1 - m^2\omega^2)^2 - 6\varepsilon^2) c_m e^{im\omega x} \\ &\quad -3\varepsilon^2 \sum_m c_m e^{i(m+2)\omega x} - 3\varepsilon^2 \sum_m c_m e^{i(m-2)\omega x} + O(\varepsilon^3). \end{aligned}$$

Comparing each coefficient of $e^{i\omega x}$ and $e^{-i\omega x}$ in (70), we obtain the following equations for c_1 and c_{-1}:

$$\left\{ \begin{array}{rcl} (\alpha - (1 - \omega^2)^2 - 6\varepsilon^2)c_1 - 3\varepsilon^2 c_{-1} &=& 4i\sqrt{3}W\varepsilon^2, \\ -3\varepsilon^2 c_1 + (\alpha - (1 - \omega^2)^2 - 6\varepsilon^2)c_{-1} &=& 4i\sqrt{3}W\varepsilon^2. \end{array} \right.$$

Noting (2), we have

$$-3\varepsilon^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_{-1} \end{pmatrix} = 4i\sqrt{3}W\varepsilon^2 \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Recalling the condition $\overline{c_1} = -c_{-1}$, we obtain

$$\mathrm{Im}c_1 = \mathrm{Im}c_{-1} = -\frac{2W}{\sqrt{3}}.$$

Therefore, we have

$$\psi^{(0,1)} = -\frac{4iW}{\sqrt{3}}\cos\omega x + C\sin\omega x + O(\varepsilon). \tag{71}$$

Similarly, it follows from (62) and (12) that

$$\begin{aligned} \mu^{(1,0)} &= \frac{\langle\psi^{(0,0)}, A^{(1,0)}\psi^{(0,0)}\rangle}{\langle\psi^{(0,0)}, \psi^{(0,0)}\rangle} \\ &= 2(1-\omega^2) \\ &= -2\sqrt{3}\varepsilon W \end{aligned}$$

and

$$\psi^{(1,0)} = \psi^{(0,0)} = \partial_x u_0.$$

Next, we consider (63). By (66) and (69), we have

$$A^{(0,0)}\psi^{(0,2)} = -A^{(0,1)}\psi^{(0,1)} - A^{(0,2)}\psi^{(0,0)} + \mu^{(0,2)}\psi^{(0,0)}.$$

By the solvability condition for $\psi^{(0,2)}$, we have

$$\mu^{(0,2)} = \frac{\langle\psi^{(0,0)}, A^{(0,1)}\psi^{(0,1)}\rangle + \langle\psi^{(0,0)}, A^{(0,2)}\psi^{(0,0)}\rangle}{\langle\psi^{(0,0)}, \psi^{(0,0)}\rangle}.$$

By (12), (67), and (71), we obtain

$$\begin{aligned} \langle\psi^{(0,0)}, \psi^{(0,0)}\rangle &= 4\varepsilon^2\pi + O(\varepsilon^3), \\ \langle\psi^{(0,0)}, A^{(0,1)}\psi^{(0,1)}\rangle &= 32\varepsilon^2 W^2\pi + O(\varepsilon^3), \\ \langle\psi^{(0,0)}, A^{(0,2)}\psi^{(0,0)}\rangle &= -16\varepsilon^2\pi + O(\varepsilon^3). \end{aligned}$$

Therefore, we find that

$$\mu^{(0,2)} = -4 + 8W^2 + O(\varepsilon).$$

Similarly, it follows from (64) and (65) that

$$\begin{aligned} \mu^{(1,1)} &= 0, \\ \mu^{(2,0)} &= -1. \end{aligned}$$

Thus the proof of Theorems 2.3 and 2.4 is complete. □

3. Concluding remarks. We studied the linear stability of roll solutions of the two-dimensional Swift–Hohenberg equation. We found that the system size affects the stability of the roll solutions. Moreover, we gave an interpretation of the phase-diffusion equation from the viewpoint of spectral analysis when the roll solutions are stable.

A natural question is as follows: What happens when the roll solution is unstable, namely, $D_\perp < 0$ or $D_{//} < 0$? From the viewpoint of spectral analysis, we expect that the primary instability of the roll solutions can be determined by the behavior of the principal eigenvalues. We proceed to perform a higher-order expansion for the eigenvalue problem

$$A_{\nu_m,\kappa_n}\psi_{\nu_m,\kappa_n} = \mu_{\nu_m,\kappa_n}\psi_{\nu_m,\kappa_n}$$

in the proof of Theorems 1.1 and 1.2. We have that for sufficiently small $2\pi m/M$ and $2\pi n/L$, the principal eigenvalues are given by

$$\begin{aligned}\mu_{mn} = &-D_\perp\left(\frac{2\pi m}{M}\right)^2 - D_{//}\left(\frac{2\pi n}{L}\right)^2\\ &-E\left(\frac{2\pi m}{M}\right)^4 - F\left(\frac{2\pi m}{M}\right)^2\left(\frac{2\pi n}{L}\right)^2 - G\left(\frac{2\pi n}{L}\right)^4 + o\left(\left(\frac{1}{M^2}+\frac{1}{L^2}\right)^3\right),\end{aligned}$$

where o is with respect to $1/M$ and $1/L$. E, F, and G are given by

$$E = 1, \quad F = -\frac{16W}{\sqrt{3}\varepsilon} + O(1), \quad G = \frac{32W^4}{3\varepsilon^2} + O\left(\frac{1}{\varepsilon}\right),$$

where O is with respect to ε. It follows that the wavenumbers of fastest growth are

$$m = \pm\frac{M}{2\pi}\sqrt{\frac{-D_\perp}{2E}} = \pm\frac{\sqrt[4]{3}M\sqrt{\varepsilon|W|}}{2\pi}$$

and

$$n = \pm\frac{L}{2\pi}\sqrt{\frac{-D_{//}}{2G}} = \pm\frac{\varepsilon L\sqrt{3(2W^2-1)}}{8\pi W^2}$$

provided $D_\perp < 0$ and $D_{//} < 0$, respectively. Unfortunately, we cannot rigorously prove that there are no other eigenvalues in the right half-plane in **C**. However, we suspect that the primary instability can be determined by the estimate above.

The phase-diffusion equation is ill posed when the roll solutions are unstable. The dynamics near the roll solutions cannot be described by only the dynamics of phase variables. In fact, various complex dynamics such as the wavenumber-changing process, the nucleation of dislocation are observed. Although many researchers have studied these phenomena, nothing has yet been proved by either mathematics or physics. (for instance, see [11] and the references therein).

Acknowledgments. The author expresses his sincere gratitude to the staff of Hiroshima National College of Maritime Technology, especially Professors Mamoru Sakamoto and Mitsuo Funaki, for their kind hospitality.

REFERENCES

[1] P. Collet and J. P. Eckmann, *Instabilities and Fronts in Extended Systems*, Princeton University Press, Princeton, NJ, 1990.

[2] M. C. Cross and P. C. Hohenberg, *Pattern formation outside of equilibrium*, Rev. Modern Phys., 65 (1993), pp. 851–1112.

[3] H. S. Greenside and W. M. Coughran, Jr., *Nonlinear pattern formation near the onset of Rayleigh–Bénard convection*, Phys. Rev. A, 30 (1984), pp. 398–428.

[4] D. Henry, *Geometric theory of semilinear parabolic equations*, in Lecture Notes in Math. 840, Springer-Verlag, New York, 1981.

[5] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[6] Y. Kuramoto, *Chemical Oscilations, Waves and Turbulence*, Synergetics, vol. 19, Springer-Verlag, New York, 1984.

[7] ———, *Phase dynamics of weakly unstable periodic structures*, Progr. Theoret. Phys., 71 (1984), pp. 1182–1196.

[8] M. Kuwamura, *The phase dynamics method with applications to the Swift–Hohenberg equation*, J. Dynamics Differential Equations, 6 (1994), pp. 185–225.

[9] P. Manneville, *Dissipative Structures and Weak Turbulence*, Academic Press, New York, 1990.

[10] A. C. Newell, *The dynamics and analysis of patterns,* in Complex Systems, SFI Studies in the Sciences of Complexity, Addison–Wesley and Longman Press, Reading, MA, and London, 1989.

[11] A. C. Newell, T. Passot, and J. Lega, *Order parameter equations for patterns*, Ann. Rev. Fluid Mech., 25 (1993), pp. 399–453.

[12] Y. Pomeau and P. Manneville, *Stability and fluctuations of a spatially periodic convective flow*, J. Phys. Lett. 40 (1979), pp. 609–612.

[13] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, vol. 4, Academic Press, New York, 1972.

[14] J. Swift and P. C. Hohenberg, *Hydrodynamic fluctuations at the convective instability*, Phys. Rev. A, 15 (1977), pp. 319–328.

[15] R. Temam, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci. 68, Springer-Verlag, New York, 1988.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1336–1340, September 1996

008

UNIQUE DETERMINATION OF A COLLECTION OF A FINITE NUMBER OF CRACKS FROM TWO BOUNDARY MEASUREMENTS*

HYUNSEOK KIM[†] AND JIN KEUN SEO[‡]

Abstract. We consider the problem of identification of a collection of a finite number of cracks in a planar domain. It is proved that the location and shape of any finite number of cracks can be determined from boundary-voltage measurements corresponding to two boundary-current fluxes.

Key words. cracks, boundary electric measurements, connected components, Jordan curve theorem

AMS subject classifications. 35R30, 35J25

1. Introduction. In [1], A. Friedman and M. Vogelius proved that the location and shape of a single crack (a curve) inside a planar domain can be uniquely determined from boundary-voltage measurements (Dirchlet data) corresponding to assigning two specific boundary-current fluxes (Neumann data). In [2], K. Bryan and M. Vogelius extended the above result by showing that if one knows a priori that the collection of cracks consists of at most n cracks, then it can be determined by voltage measurements corresponding to $n+1$ specific fluxes. In this paper, we prove that only two measurements are sufficient to determine a collection of cracks (see the Main Theorem below). Since it was proved in [1] that one flux is not sufficient to determine even a single crack, our result may be regarded as an optimal extension of [1] and [2]. We introduce some notations and definitions to precisely describe our result.

Let Ω be a simply connected bounded domain in R^2 with a smooth boundary $\partial\Omega$ and γ a positive real analytic function on $\overline{\Omega}$. By a crack, we mean a C^2 simple curve σ in Ω, i.e., a one-to-one twice continuously differentiable map $\sigma : [0,1] \to \Omega$ with nonvanishing derivative, and by a collection of cracks, we mean a collection of cracks consisting of a finite number of mutually disjoint cracks σ_k, $k = 1, \dots, n$ (possibly $n = 0$).

Given a function $\psi \in L^2(\partial\Omega)$ with average zero, i.e., $\int_{\partial\Omega} \psi\, ds = 0$ and a collection Σ of cracks in Ω, let us denote by $P(\psi, \Sigma)$ the following minimization problem:

$$P(\psi,\Sigma) : \left|\begin{array}{l} \text{Find a function u in } H^1(\Omega) \text{ that minimizes the functional} \\ \displaystyle J(v) = \frac{1}{2}\int\limits_{\Omega\Sigma} \gamma|\nabla v|^2 dx - \int\limits_{\partial\Omega} \psi v\, ds \\ \text{in the class } K = \{v \in H^1(\Omega)\text{: } v \text{ is constant on each } \sigma_k \text{ in } \Sigma\}. \end{array}\right.$$

This minimization problem physically corresponds to minimizing the total energy required to sustain the specified boundary-current flux ψ and the requirement that the potential u is constant on the cracks means that the cracks are perfectly conducting.

* Received by the editors October 7, 1994; accepted for publication (in revised form) April 14, 1995.

[†] Department of Mathematics, Pohang University of Science and Technology, Pohang, 790-784, Republic of Korea.

[‡] Department of Mathematics, Yonsei University, Seoul 120-740, Republic of Korea. The research of this author was supported in part by the GARC–KOSEF nondirected fund, the Korean Research Foundation (1993), and the BSRI (1994).

A solution u to the problem $P(\psi, \Sigma)$ is continuous on Ω (see Lemma 2.1 in [1]) and satisfies the following boundary value problem:

$$\begin{cases} \nabla \cdot (\gamma \nabla u) = 0 \text{ in } \Omega \backslash \Sigma, \\ u = \ \textit{constant} \ \text{ on each } \sigma_k \in \Sigma, \\ \gamma \dfrac{\partial u}{\partial \nu} = \psi \text{ on } \partial\Omega, \end{cases}$$

where $\frac{\partial}{\partial \nu}$ denotes the outward normal derivative on $\partial\Omega$.

Let us state our main result.

MAIN THEOREM. *Let ψ_1 and ψ_2 be two nonvanishing piecewise-continuous functions on $\partial\Omega$ with average zero such that for each real α, the set $\{z \in \partial\Omega : \psi_1(z) - \alpha\psi_2(z) \geq 0\}$ is connected and ψ_1 is not identical to $\alpha\psi_2$. Suppose that Σ and $\tilde{\Sigma}$ are collections of cracks in Ω and for each $i = 1, 2, u_i$ and $\tilde{u}_i$ are solutions to the problems $P(\psi_i, \Sigma)$ and $P(\psi_i, \tilde{\Sigma})$, respectively. Then $u_i = \tilde{u}_i$ for $i = 1, 2$ on $\partial\Omega$ implies that $\Sigma = \tilde{\Sigma}$.*

Remarks. (1) It is easy to construct functions ψ_i satisfying the hypotheses of the Main Theorem. For completeness, we give an example:

Imagine $\partial\Omega$ as the interval [0, 8] with endpoints 0 and 8 identified.

Define $\psi_i \in C^0([0, 8])$ as follows:

$$\psi_1(x) = \begin{cases} -1 & \text{for all } x \in [0,1] \cup [7,8], \\ 1 & \text{for all } x \in [3,5], \end{cases}$$

$$\psi_2(x) = \begin{cases} 0 & \text{for all } x \in [0,1] \cup [3,5] \cup [7,8], \\ 1 & \text{for } x = 2, \\ -1 & \text{for } x = 6, \end{cases}$$

and

ψ_i's are linear in the remaining domains.

Then the ψ_i's satisfy the hypotheses of the Main Theorem.

(2) Our results extend Theorem 1.1 of [1] for a special case. Indeed, we could remove the restriction $\epsilon \leq \epsilon_0$ of that theorem.

Our proof of the Main Theorem depends heavily on the maximum principle and topological properties of R^2.

In §2 we establish some preliminary lemmas, and in §3 we prove Main Theorem.

2. Preliminary lemmas. Throughout this section, we assume that u is a solution to the minimization problem $P(\psi, \Sigma)$, where ψ is a nonvanishing piecewise-continuous function on $\partial\Omega$ with $\int_{\partial\Omega} \psi \, ds = 0$ and Σ is a collection of cracks in Ω. Clearly, u is nonconstant.

LEMMA 2.1. *Let $\sigma \in \Sigma$, and let Ω' be a subdomain of Ω with $\sigma \subset \Omega'$. Then*

$$\inf_{\Omega'} u < u|_\sigma < \sup_{\Omega'} u.$$

Proof. Let $c = u|_\sigma$. To obtain a contradiction, assume that $c = \sup_{\Omega'} u$.

For $s > 0$, let V_s denote the open set

$$V_s = \{z \in \Omega' : \text{ dist } (z, \sigma) < s\}.$$

Choose $\epsilon > 0$ so small that $\overline{V_{2\epsilon}} \subset \Omega'$ and $(V_{2\epsilon} \backslash \sigma) \cap \Sigma = \emptyset$. If we set $a = \sup_{\partial V_\epsilon} u$; then it follows from the strong-maximum principle that $a < c$.

Let $b = (a + c)/2$, and define

$$\Omega'' = \{z \in V_\epsilon : u(z) > b\}.$$

Then

$$\sigma \subset \Omega'', \qquad \overline{\Omega''} \subset V_\epsilon, \quad \text{and} \quad \partial\Omega'' = \{z \in V_\epsilon : u(z) = b\}.$$

Define

$$\tilde{u} = \begin{cases} b & \text{for all } z \text{ in } \Omega'', \\ u(z) & \text{for all } z \text{ in } \Omega \backslash \Omega''. \end{cases}$$

Then $\tilde{u}$ belongs to the class K and $J(\tilde{u}) \leq J(u)$. Since u is a solution to the minimization problem $P(\psi, \Sigma)$, we obtain $J(\tilde{u}) = J(u)$ and, therefore, $\nabla u = 0$ in Ω''. Hence u is constant in Ω'', and by the analytic continuation, u is constant in Ω—a contradiction. The assumption that $c = \inf_{\Omega'} u$ also leads to the same contradiction. This completes the proof.

LEMMA 2.2. *If Ω' is a subdomain of Ω, then*

$$\inf_{\partial\Omega'} u \leq u(z) \leq \sup_{\partial\Omega'} u \quad \textit{for all } z \in \Omega'.$$

Proof. By the maximum principle, u cannot have a local maximum in $\Omega \backslash \Sigma$. The result now follows from Lemma 2.1.

We now state the key lemma for our proof of the Main Theorem.

LEMMA 2.3. *If the set $\{z \in \partial\Omega : \psi(z) \geq 0\}$ is connected, then we have $\nabla u(z) \neq 0$ for every z in $\Omega \backslash \Sigma$.*

Proof. To obtain a contradiction, assume that $\nabla u(z_0) = 0$ for some z_0 in $\Omega \backslash \Sigma$. Assume for simplicity that $z_0 = 0$ and $u(0) = 0$. Let (r, θ) denote polar coordinates near 0. Since $\nabla u(0) = 0$ and u is analytic near 0, we know that $\frac{\partial}{\partial r} u(0, \theta) = 0$, and by expanding in a Taylor series in r, we obtain

$$u(z) = r^n (a \sin(n\theta) + b \cos(n\theta) + rA(r, \theta))$$

for some a and b (not both zero) and some $n \geq 2$.

Here $A(r, \theta)$ is a smooth function near 0.

Since $a \sin(n\theta) + b \cos(n\theta) = \sqrt{a^2 + b^2} \sin(n\theta + \alpha)$ for some $\alpha \in [0, 2\pi]$, without loss of generality, we may assume by a rotation about 0 that $b = 0$. Since $a \neq 0$, we may also assume that $a > 0$.

Then we have

$$u(z) = u(r, \theta) = r^n (a \sin(n\theta) + rA(r, \theta))$$

and

$$\lim_{r \to 0} \left\{ \frac{u(r, \theta)}{r^n} - a \sin(n\theta) \right\} = 0 \quad \text{uniformly in } \theta.$$

Hence there exists an open disc $B = B(0,\delta)$ in $\Omega\backslash\Sigma$ with center 0 and radius δ such that u is positive in $S_1^+ \cup S_2^+$ and u is negative in $S_1^- \cup S_2^-$, where

$$S_i^+ = \left\{(r,\theta) \in B : \frac{(2i-2)\pi}{n} + \frac{\pi}{4n} < \theta < \frac{(2i-1)\pi}{n} - \frac{\pi}{4n}\right\},$$

and

$$S_i^- = \left\{(r,\theta) \in B : \frac{(2i-1)\pi}{n} + \frac{\pi}{4n} < \theta < \frac{2i\pi}{n} - \frac{\pi}{4n}\right\} \quad (i = 1,2).$$

Set

$$\Omega^+ = \{z \in \Omega : u(z) > 0\} \quad \text{and} \quad \Omega^- = \{z \in \Omega : u(z) < 0\}.$$

For each $i = 1,2$, let Ω_i^+ (respectively, Ω_i^-) denote the connected component of Ω^+ (respectively, Ω^-) containing S_i^+ (respectively, S_i^-). Then the sets Ω_i^+ and Ω_i^- are subdomains of Ω.

CLAIM 1. $\Omega_1^+ \cap \Omega_2^+ = \emptyset$ *and* $\Omega_1^- \cap \Omega_2^- = \emptyset$.

Proof. Suppose that $\Omega_1^+ \cap \Omega_2^+ \neq \emptyset$. Then by the definition of connected components, we see that $\Omega_1^+ = \Omega_2^+$. For each $i = 1,2$, choose a point z_i in S_i^+, and let ρ_i be a line segment in $S_i^+ \cup \{0\}$ whose endpoints are 0 and z_i. Since Ω_i^+ is an open connected subset of R^2, there exists a simple closed curve ρ in $\Omega_i^+ \cup \{0\}$ containing $\rho_1 \cup \rho_2$. Note that from the Jordan curve theorem, the interior of ρ contains either S_1^- or S_2^-. Hence by Lemma 2.2, u is nonnegative in either S_1^- or S_2^-—a contradiction. The assumption that $\Omega_1^- \cap \Omega_2^- \neq \emptyset$ also leads to a similar contradiction.

Set

$$\Gamma_i^+ = \partial\Omega_i^+ \cap \partial\Omega \quad \text{and} \quad \Gamma_i^- = \partial\Omega_i^- \cap \partial\Omega.$$

Then all sets Γ_i^+ and Γ_i^- are nonempty (if $\Gamma_i^+ = \emptyset$, then $\overline{\Omega_i^+} \subset \Omega$ and from the definition of connected components, $u = 0$ on $\partial\Omega_i^+$, which implies by Lemma 2.2 that $u = 0$ in Ω_i^+—a contradiction).

Let z_i^+ and z_i^- be points on $\partial\Omega$ such that

$$u(z_i^+) = \max_{\Gamma_i^+} u \quad \text{and} \quad u(z_i^-) = \min_{\Gamma_i^-} u \quad (i = 1,2).$$

Then by Lemma 2.2,

$$u(z_i^+) = \sup_{\Omega_i^+} u \quad \text{and} \quad u(z_i^-) = \inf_{\Omega_i^-} u \quad (i = 1,2).$$

CLAIM 2. $\psi(z_i^+) > 0$ *and* $\psi(z_i^-) < 0$.

Proof. Since $S_i^- \subset \Omega_i^-$, $u(z_i^-) < 0$ by Lemma 2.2 and since u is continuous on $\overline{\Omega}$, there is an open disc B_i^- centered at z_i^- such that $B_i^- \cap \Omega \subset \Omega_i^-$, and $B_i^- \cap \partial\Omega$ is a smooth portion of $\partial\Omega_i^-$. Hopf's lemma shows that $\psi(z_i^-) = \frac{\partial u}{\partial \nu}(z_i^-) < 0$. The same argument proves that $\psi(z_i^+) > 0$.

For each $i = 1,2$, choose z_i' in S_i^-, and let ρ_i' be a line segment in $S_i^- \cup \{0\}$ whose endpoints are 0 and z_i'. Then since each Ω_i^- is open and connected and since some open ball centered at each z_i^- intersects Ω_i^- as in the proof of Claim 2, there exists a simple curve ρ' in $\Omega_1^- \cup \Omega_2^- \cup \{0, z_1^-, z_2^-\}$ containing $\rho_1' \cup \rho_2'$ whose endpoints are z_1^- and z_2^-. By the Jordan curve theorem, we see that the curve ρ' divides Ω into two subdomains Ω_1 and Ω_2, where $\Omega_1^+ \subset \Omega_1$ and $\Omega_2^+ \subset \Omega_2$.

Note that $\partial\Omega \subset \partial\Omega_1 \cup \partial\Omega_2$, $\partial\Omega \cap \partial\Omega_1 \cap \partial\Omega_2 = \{z_1^-, z_2^-\}$, $z_1^+ \in \partial\Omega_1$, and $z_2^+ \in \partial\Omega_2$. Then it follows from Claim 2 that the set $\{z \in \partial\Omega : \psi(z) \geq 0\}$ consists of at least two disjoint curves, which is contrary to the hypothesis. This completes the proof.

3. Proof of the Main Theorem. It follows from the analytic continuation and continuity of u and $\tilde{u}$ that $u = \tilde{u}$ in Ω; for details, see Bryan and Vogelius [2, §2].

Suppose that $\Sigma \neq \tilde{\Sigma}$. Then we may assume that there is an simple curve ρ in $\Omega \backslash \Sigma$ such that each u_i is constant on ρ. Furthermore, we may assume that ρ is an analytic curve and $\frac{\partial u_2}{\partial \nu}(z_0) \neq 0$ for some z_0 in ρ. Otherwise, u_2 must be constant in Ω by the analytic continuation and $\psi_2 = 0$, which is a contradiction. Set $u = u_1 - \alpha u_2$ and $\psi = \psi_1 - \alpha \psi_2$, where $\alpha = \frac{\partial u_1}{\partial \nu}(z_0) / \frac{\partial u_2}{\partial \nu}(z_0)$. Then $\nabla u(z_0) = 0$, and using a standard argument, we can easily show that u is a solution to the minimization problem $P(\psi, \Sigma)$, which is contrary to Lemma 2.3. This completes the proof.

Remarks. (1) Our technique in this paper does not work in three-dimensional case because we do not have a three-dimensional version of the Jordan curve theorem. Indeed, we do not know how to solve the following interesting problem:

Let B be the unit ball in R^3, σ a curve in B, and ψ a smooth nonzero function on ∂B satisfying $\int_{\partial B} \psi \, ds = 0$. Suppose that the solution u to the Neumann problem

$$\begin{cases} \Delta u = 0 & \text{in } B, \\ \dfrac{\partial u}{\partial \nu} = \psi & \text{on } \partial B \end{cases}$$

satisfies $\nabla u = 0$ on σ. Is the set $\{x \in \partial B : \psi(x) > 0\}$ disconnected?

(2) A referee pointed out that the result in this paper works for the case $\gamma \in C^\infty(\Omega)$.

(3) After this paper was accepted for publication, we learned that a similar result was obtained independently by G. Alessandrini and A. Diaz Valenzuela [*SIAM J. Control. Optim.*, 34 (1996), pp. 913–921].

Acknowledgments. We would like to thank the referees for several valuable comments.

REFERENCES

[1] A. Friedman and M. Vogelius, *Determining cracks by boundary measurements*, Indiana Univ. Math. J., 38 (1989), pp. 527–556.

[2] K. Bryan and M. Vogelius, *A uniqueness result concerning the identification of a collection of cracks from finitely many electrostatic boundary measurements*, SIAM J. Math. Anal., 23 (1992), pp. 950–958.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1341–1359, September 1996

009

PHASE-FIELD THEORY FOR FITZHUGH–NAGUMO-TYPE SYSTEMS*

PIERPAOLO SORAVIA[†] AND PANAGIOTIS E. SOUGANIDIS[‡]

Abstract. In this paper, we study the asymptotics of Fitzhugh–Nagumo-type systems of reaction-diffusion equations with bistable nonlinearity. In the limit, we obtain an interface moving with normal velocity determined by the dynamics and the scaling.

Key words. Fitzhugh–Nagumo-type systems, phase-field theory, front propagation, phase transition, reaction-diffusion systems

AMS subject classifications. 35K47, 35K55, 35K65

Introduction. Several chemical and biological wave phenomena are modeled by systems of reaction-diffusion equations which, in a simplified form, look like

$$
\begin{cases} \text{(i)} \quad u_t^\epsilon - \alpha\beta\Delta u^\epsilon + \alpha f^\epsilon(u^\epsilon, v^\epsilon) = 0, \\ \\ \text{(ii)} \quad v_t^\epsilon - \epsilon\gamma(b\Delta v^\epsilon - g^\epsilon(u^\epsilon, v^\epsilon)) = 0 \end{cases} \quad \text{in } \mathbb{R}^N \times (0,\infty), \tag{0.1}
$$

where the positive constants α, β, γ, and b are related to the particular physical derivation, the positive parameter ϵ is a measure of the ratio of the rates of change of v^ϵ and u^ϵ, and, finally, the vector field $(u,v) \longrightarrow (f^\epsilon(u,v), g^\epsilon(u,v))$ is of *bistable type.* A typical bistable vector field is

$$
\begin{cases} f(u,v) = (u-\mu)(u^2-1)+v, \\ \\ g(u,v) = \sigma v - u \end{cases} \quad \text{for } \mu \in (-1,1) \text{ and } \sigma > 0. \tag{0.2}
$$

One of the best-known examples of (0.1) is the Fitzhugh–Nagumo model (see Fitzhugh [FH] and Nagumo, Arimoto, and Yoshizawa [NAY]), which describes waves in neural activity and the conduction of electric impulses in nerve axons. Other examples are the Belousov–Zhabatinskii chemical reactions (see Tyson and Fife [TF]). For an expanded list of references about physical problems modeled by (0.1) as well as a detailed discussion on the qualitative properties of the solutions of (0.1), we refer to the monographs of Fife [F1, F2], the papers of Hastings [H] and Chen [Chxy], and the references therein.

In this paper, we study the asymptotic behavior of the solutions of (0.1) in the limit $\epsilon \longrightarrow 0$ for either

$$
\alpha = \gamma = \epsilon^{-1} \quad \text{and} \quad \beta = \epsilon^2 \tag{0.3}
$$

or

$$
\alpha = \epsilon^{-2}, \quad \gamma = \epsilon^{-1}, \quad \text{and} \quad \beta = \epsilon^2. \tag{0.4}
$$

*Received by the editors January 10, 1994; accepted for publication (in revised form) March 7, 1995.

[†]Dipartimento di Matematica Pura e Applicata, Universitá di Padova, 35131 Padova Italy. The research of this author was completed while he was visiting the Division of Applied Mathematics, Brown University, Providence, RI 02912 and was partially supported by the Consiglio Nazionale delle Ricerche.

[‡]Department of Mathematics, University of Wisconsin–Madison, Madison, WI 53706. The research of this author was partially supported by the National Science Foundation, the Army Research Office, and the Alfred P. Sloan Foundation.

Below we describe our results in a somewhat informal way and only in a special case. The precise statements and assumptions in full generality are formulated in the main body of the paper.

The behavior in the limit $\epsilon \to 0$ of the solutions of (0.1) with (f, g) given by (0.2) when (0.3) holds is governed by the partial differential equation (PDE)

$$v_t - b\Delta v + g_\pm(v) = 0 \quad \text{in } \bigcup_{t>0}(\Omega^\pm(t) \times \{t\}), \tag{0.5}$$

where

$$g_\pm(v) = \sigma v - h_\pm(v),$$

$h_-(v)$ and $h_+(v)$ being the smallest and largest zeroes, respectively, of the map $u \mapsto f(u, v)$. The set

$$\Gamma(t) = \mathbb{R}^N \backslash (\Omega^+(t) \cup \Omega^-(t))$$

moves with normal velocity

$$V = -c(v(x,t)),$$

where $c(v(x,t))$ is the speed of the increasing traveling wave associated with $u \mapsto f(u, v(x,t))$ which connects $h_\pm(v(x,t))$.

To study the asymptotic limit (0.4) as $\epsilon \to 0$ of (0.1), we need to consider $f^\epsilon(u, v) = f(u, \epsilon v)$, where (f, g) is given by (0.2) and $\mu = 0$. In this case, the asymptotics are again governed by (0.5) but now the normal velocity V of the interface is

$$V = \kappa - \dot{c}(0)v(x,t),$$

where κ denotes the mean curvature of the interface. Finally, in either case, the u^ϵ's converge uniformly to $h_\pm(v)$ in $\bigcup_{t>0}\Omega^\pm(t) \times \{t\}$. Notice that all the above statements are global in time and not only up to the first time the geometric evolution develops singularities!

Local-in-time existence of smooth solutions for the limit problem when the speed of the interface is $c(v) - \epsilon\kappa$ was proved by Chen in [Chxy]. Giga, Goto, and Ishii in [GGI] gave a definition of global solutions to the above limit problem and proved their existence. Results equivalent to ours which, however, hold only as long as the evolution is smooth were obtained by Chen (see [Chx]).

We proceed with a brief review of what it means for a surface Γ_t to move with a prescribed normal velocity

$$V = V(Dn, n, x, t). \tag{0.6}$$

Surfaces in $\mathbb{R}^N$ evolving according to this rule can start out smooth and yet develop singularities at a later time. A great deal of work has been done recently in order to interpret the evolution of surfaces *past singularities.* Here we will be using a combination of the so-called *level-set* and *distance-function* approaches. For a detailed description of all approaches and their relationship as well as their consequences, we refer to the papers cited below and references therein.

The *level-set* approach was introduced for numerical calculations by Osher and Sethian (see [OsS]). (See also Ohta, Jasnow, and Kawasaki [OhJK] in the physics literature and Barles [Ba] for a first-order model for flame propagation.) This approach represents the evolving surface as the level set of an auxiliary function solving an appropriate nonlinear PDE. The level-set approach has been extensively developed by Evans and Spruck in [ESp] for motion by mean curvature and independently by Chen, Giga, and Goto (see [ChGG]) for more general geometric motions. Then Giga, Goto, Ishii, and Sato [GGIS] and later Goto [G] and Ishii and Souganidis [IS] used the level-set approach to study motions, where the velocity depends on space and time and V may depend on Dn in a superlinear way, respectively. All the above works are based on the theory of *viscosity solutions* to fully nonlinear second-order parabolic (possibly degenerate) equations, which were introduced by Crandall and Lions in [CrL] and Lions in [L]. (For a detailed overview of the theory of viscosity solutions as well as a complete list of references (until at least 1991), we refer the reader to Crandall, Ishii, and Lions [CrIL]).

The *distance-function* approach, which was initiated by Soner in [So1] and later extended to very general situations by Barles, Soner, and Souganidis (see [BaSS]), is more intrinsic. It describes the motion in terms of the properties of the distance function to the evolving surface. For the precise relation between the two approaches as well as their consequences, we refer to [BaSS].

The generalized evolution $\{\Gamma_t\}_{t\geq 0}$ governed by (0.6), which starts with a given closed surface $\Gamma_0 \subset \mathbb{R}^N$, exists and is uniquely defined for all $t \geq 0$. Moreover, it agrees with the classical differential-geometric flow as long as the latter exists. The geometric motion may, on the other hand, develop singularities, change topological type, and exhibit various other geometric pathologies.

In spite of these peculiarities, the generalized motion $\{\Gamma_t\}_{t\geq 0}$ has been proven in several occasions to be the *right* way to extend the classical motion past singularities. Some of the most definitive results in this direction were obtained by Evans, Soner, and Souganidis (see [ESoS]) and Barles, Soner, and Souganidis (see [BaSS]), who proved that the generalized evolution governs for all times the asymptotic behavior of the following semilinear reaction-diffusion equation, which was proposed by Allen and Cahn in [AC] to describe the time evolution of an "order parameter" u determining the phase of a polycrystalline material:

$$u_t - \alpha\beta\Delta u + \alpha F'(u) = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty)$$

in the limit $\epsilon \to 0^+$ for

$$\alpha = \epsilon^{-1} \quad \text{and} \quad \beta = \epsilon^2 \quad \text{or} \quad \alpha = \epsilon^{-2} \quad \text{and} \quad \beta = \epsilon^2.$$

Here F is a W-shaped potential and α and β are related to the physical model yielding the equation. The choice of the appropriate asymptotic limit is governed by the difference of the depths of the wells of F. The results of [BaSS] apply to more general situations where F also depends on (x, t) (see also [BaBS]). It is exactly this general dependence on the potential that will allow us to obtain in this paper the results mentioned above.

Formal asymptotic expansions suggesting the relation between generalized evolution and the asymptotics of reaction-diffusion equations have been carried out by Caginalp [Ca], Fife [F1, F2], Rubinstein, Sternberg, and Keller [RSK], and others. The radial case was studied by Bronsard and Kohn in [BrK]. In [DeMS], de Mottoni and Schatzmann gave a complete proof for the case of the classical geometric motion. Chen [Chxy] generalized much of this work and gave simpler proofs but still for smooth geometric evolutions, as has Korevaar in unpublished work. Ilmanen [I] and Soner [So2] refined [ESoS] to overcome the possibility of *interface fattening* and Katsoulakis, Kossioris, and Reitich [KKR] studied the asymptotics in bounded domains. Using the asymptotics of reaction-diffusion equations to build a generalized mean-curvature flow has been suggested by Bronsard and Kohn in [BrK], DeGiorgi in [DeG], and others. This is what is called the *phase-field* approach to mean curvature. An immediate consequence of [BaSS], [ESoS], etc. is the equivalance of the level-set/distance-function and phase-field approaches. Another recent result related to the above was obtained by Katsoulakis and Souganidis (see [KS1] and [KS2]), who proved that the hydrodynamic limit of certain interacting particle systems yields (again for all times) generalized front evolution. Finally, for a general theory about how moving fronts arise in the limit of appropriately scaled systems and/or equations, we refer to Barles and Souganidis [BaS]; see also Souganidis [Sou].

The paper is organized as follows. In §1, we formulate the asymptotic problem, state the assumptions, and recall some basic facts. In §2, we recall the precise definition of the generalized evolution and introduce the necessary tools from [BaSS]. Section 3 is devoted to the statement and the proof of our main result. Finally, in §4, we discuss the asymptotics of (0.1) in the limit (0.4).

1. The Fitzhugh–Nagumo model. As mentioned in the introduction, in this paper, we consider the behavior of system (0.1) with initial conditions

$$u^\epsilon = u_0^\epsilon \quad \text{and} \quad v^\epsilon = v_0^\epsilon \quad \text{on } \mathbb{R}^N \times \{0\} \tag{1.1}$$

in the asymptotic limits (0.3) or (0.4). For technical reasons which we will point out later, we will work in a periodic domain $\Pi \times [0,\infty)$, where Π is an N-dimensional torus.

Throughout the paper, we will assume that

$$f^\epsilon \text{ and } g^\epsilon \text{ are smooth} \quad \text{and} \quad f_v^\epsilon \geq 0, \quad g_u^\epsilon \leq 0 \quad \text{for small } \epsilon. \tag{1.2}$$

The sign conditions in (1.2), which are satisfied by (0.2), are not compatible with the standard maximum principle for systems. Our results also hold with similar proofs if instead of (1.2) we assume

$$f_v^\epsilon \geq 0, \quad g_u^\epsilon \geq 0 \quad \text{or} \quad f_v^\epsilon \leq 0, \quad g_u^\epsilon \leq 0 \quad \text{or} \quad f_v^\epsilon \leq 0, \quad g_u^\epsilon \geq 0. \tag{1.3}$$

The next set of assumptions are about the fact that, for each v fixed, $u \to f^\epsilon(u,v)$ is of *bistable* type, i.e., there exist $\overline{v}_- < \overline{v}_+$ such that for all $[v_-,v_+] \subseteq (\overline{v}_-,\overline{v}_+)$, $v \in [v_-,v_+]$, and ϵ and a sufficently small,

$$\begin{cases} u \mapsto f^\epsilon(u,v) - a \text{ vanishes at only three points} \\ \qquad\qquad h_-^\epsilon(v,a) < h_0^\epsilon(v,a) < h_+^\epsilon(v,a) \\ \text{and} \\ \text{(i)} \ f_u^\epsilon(h_\pm^\epsilon(v,a),v) \geq k > 0 \quad \text{and} \quad f_u^\epsilon(h_0^\epsilon(v,a),v) \leq -k \\ \text{(ii)} \ g^\epsilon, h_\pm^\epsilon, h_0^\epsilon \to g, h_\pm, h_0 \quad \text{as } \epsilon \to 0, \end{cases} \tag{1.4}$$

where $k = k(v_-,v_+)$ is independent of ϵ and all the limits are uniform in $v \in [v_-,v_+]$ and a. In view of (1.4), for each ϵ and a sufficiently small and $v \in (\overline{v}_-,\overline{v}_+)$, there exists a unique $c^\epsilon(v,a)$ and a unique-up-to-translations $q^\epsilon(\cdot,v,a)$ (cf. Aronson and Weinberger [AW], Fife and McLeod [FM], etc.) such that

$$q_{rr}^\epsilon(r,v,a) + c^\epsilon(v,a) q_r^\epsilon(r,v,a) = f^\epsilon(q^\epsilon(r,v,a)) - a \tag{1.5}$$

and

$$q^\epsilon(\pm\infty, v, a) = h^\epsilon_\pm(v, a). \tag{1.6}$$

We continue by listing some technical assumptions we make on (q^ϵ, c^ϵ). We then verify these assumptions for some special vector fields like the one given by (0.2). For more general f^ϵ's, we refer to [BaSS].

To this end, we assume that, as $\epsilon \to 0$,

$$q^\epsilon \text{ and } c^\epsilon \text{ depend smoothly on } (v, a) \tag{1.7}$$

and either

$$c^\epsilon(v, a) \to \alpha(v, a) \tag{1.8}$$

or

$$\epsilon^{-1} c^\epsilon(v, \epsilon a) \to \alpha(v, a) \quad \text{if } c^\epsilon(v, \epsilon a) \to 0 \tag{1.9}$$

with all the limits local uniform in (v, a). Moreover, we assume that there exists $K > 0$ (independent of v) such that

$$|\alpha(v, a) - \alpha(\hat{v}, a)| \leq K|v - \hat{v}| \tag{1.10}$$

for all sufficiently small a. In the case that (1.8) holds, we will also assume that there exists $K > 0$ such that

$$\begin{cases} \text{(i)} \ \lim\limits_{\epsilon \to 0} \sup\limits_{(r,v,a)} [\epsilon[|q^\epsilon_v| + |q^\epsilon_{rv}| + |q^\epsilon_a|] + \epsilon^2 |q^\epsilon_{vv}|] = 0, \\ \text{(ii)} \ |q^\epsilon_{rr}| + |q^\epsilon_r| \leq K e^{-K\delta} \quad \text{for all } |r| > \delta. \end{cases} \tag{1.11}$$

If (1.9) holds, then we assume that there exists $K > 0$ such that

$$\begin{cases} \text{(i)} \ \lim\limits_{\epsilon \to 0} \sup\limits_{(r,v,a)} [\epsilon(|q^\epsilon_v| + |q^\epsilon_{vv}| + |q^\epsilon_a|) + |q^\epsilon_{rv}|] = 0, \\ \text{(ii)} \ \frac{1}{\epsilon^2}|q^\epsilon_{rr}| + \frac{1}{\epsilon}|q^\epsilon_r| \leq K e^{-K\delta} \quad \text{for all } |r| > \delta. \end{cases} \tag{1.12}$$

Finally, for all v and ϵ, a sufficiently small, we assume that there exists $M > 0$ such that

$$q^\epsilon_r \geq 0 \quad \text{and} \quad \|q^\epsilon_v\|_\infty \leq M. \tag{1.13}$$

Next, we present a couple of examples where the above hypotheses hold true. Indeed, let $\mu^\epsilon \in (-1, 1)$ and consider either of the functions

$$f^\epsilon(u, v) = (u - \mu^\epsilon)(u^2 - 1) + v \tag{1.14}$$

or

$$f^\epsilon(u, v) = (u - \mu^\epsilon)(u^2 - 1) + \epsilon v. \tag{1.15}$$

It is immediate that in either case (1.4) holds. Moreover, the pair (q^ϵ, c^ϵ) is given by

$$(1.16)\qquad \begin{cases} q^\epsilon(r,v,a) = h_-^\epsilon(v,a) + m^\epsilon(v,a)(1+\exp(-m^\epsilon(v,a)(r+r_0^\epsilon(v,a))/\sqrt{2}))^{-1} \\ \text{and} \\ c^\epsilon(r,v,a) = (2h_0^\epsilon(v,a) - h_+^\epsilon(v,a) - h_-^\epsilon(v,a))/\sqrt{2}, \end{cases}$$

where

$$m^\epsilon(v,a) = h_+^\epsilon(v,a) - h_-^\epsilon(v,a)$$

and $r_0^\epsilon(v,a)$ is appropriately chosen. If f^ϵ is given by (1.15) and $\mu^\epsilon \to 0$, then (1.9) holds and $\alpha(v,a) = \frac{3}{\sqrt{2}}(v+a)$. The rest of the conditions above can be easily checked to hold.

The next set of assumptions are about the initial conditions (1.1). We assume that the periodic-in-Π functions $u_0^\epsilon, v_0^\epsilon \in L^\infty(\mathbb{R}^N)$ are such that, as $\epsilon \to 0$,

$$(1.17)\qquad v_0^\epsilon \to v_0 \quad \text{uniformly},$$

and that there exist an open set $\Omega_0 \subset \mathbb{R}^N$ and a closed, Π-periodic, $(N-1)$-hypersurface Γ_0 such that

$$(1.18)\qquad \mathbb{R}^N = \Omega_0 \cup \overline{\Omega}_0^c \cup \Gamma_0 \text{ (disjoint union)},$$

and the closed sets

$$(1.19)\qquad \Gamma_0^\epsilon = \{x \in \mathbb{R}^N : u_0^\epsilon(x) = h_0^\epsilon(v_0^\epsilon(x))\}$$

converge to Γ_0 in the Hausdorff metric, where

$$h_0^\epsilon(v) = h_0^\epsilon(v,0) \quad \text{and} \quad h_\pm^\epsilon(v) = h_\pm^\epsilon(v,0).$$

For each $\delta > 0$, there exists $\eta = \eta(\delta) > 0$ such that

$$(1.20)\qquad \begin{cases} \varliminf\limits_{\epsilon\to 0} u_0^\epsilon \ge h_0(v_0) + \eta & \qquad\qquad \Omega_0 \cap \{x : \operatorname{dist}(x,\Gamma_0) \ge \delta\}, \\ & \text{locally uniformly in} \\ \varlimsup\limits_{\epsilon\to 0} u_0^\epsilon \le h_0(v_0) - \eta & \qquad\qquad \overline{\Omega}_0^c \cap \{x : \operatorname{dist}(x,\Gamma_0) \ge \delta\}, \end{cases}$$

where $\operatorname{dist}(x,\Gamma_0)$ is the usual distance function from x to Γ_0. Notice that no smoothness is required on u_0^ϵ, Γ_0^ϵ, and Γ_0!

Our last assumption is about the existence of L^∞-bounds on the solution (u^ϵ, v^ϵ) of (0.1) and (1.2) which are independent of ϵ. We assume that

$$(1.21)\qquad \begin{cases} \text{there exist } M > 0,\ T > 0, \text{ and } \rho > 0 \text{ such that for all } \epsilon \text{ small,} \\ v^\epsilon \in (\overline{v}_- + \rho, \overline{v}_+ - \rho) \quad \text{and} \quad |u^\epsilon| \le M \quad \text{in } Q_T = \mathbb{R}^N \times (0,T). \end{cases}$$

This assumption, which is essential to defining $h_\pm^\epsilon$ and h_0^ϵ in (1.4), can be verified for the vector field (0.2) using the theory of invariant regions developed by Rauch and Smoller (see [RS]) and Chueh, Conley, and Smoller (see [ChCS]). The following lemma gives sufficient conditions so that (1.21) holds for any $T > 0$.

LEMMA 1.1. *Assume that $f^\epsilon(u,v)$ is given by* (1.14) *with $\mu = \mu^\epsilon$, $g(u,v) = \sigma v - u$, and that $\{v_0^\epsilon(x) : x \in \mathbb{R}^N\} \subset [v_-, v_+] \subset (\overline{v}_-, \overline{v}_+)$, where $\overline{v}_\pm$ are as in* (1.4). *Set $u_+ = \max_{[v_-,v_+]} h_+(v)$ and $u_- = \min_{[v_-,v_+]} h_-(v)$ and assume $\sigma v_+ > u_+$ and $\sigma v_- < u_-$. Then*

$$v^\epsilon \in [v_-, v_+] \quad \textit{and} \quad u^\epsilon \in [u_-, u_+] \quad \textit{on } \mathbb{R}^N \times (0, +\infty)$$

and $h^\epsilon_\pm(v^\epsilon, a)$ *and* $h^\epsilon_0(v^\epsilon, a)$ *are well defined for all* ϵ *and* a *sufficiently small. Finally, if* $v_c \in (\overline{v}_-, \overline{v}_+)$ *is the unique point such that*

$$\int_{h_-(v_c)}^{h_+(v_c)} f(u, v_c) du = 0$$

and $v_c \notin (v_-, v_+)$, *then* $\alpha(v, 0)$, *given by* (1.8), *has constant sign.* □

The proof of this lemma is based on the results of [ChCS] and the observation that for all suitably small $\rho > 0$, $[v_-, v_+] \times [\sigma v_- + \rho, \sigma v_+ - \rho]$ is, in the language of [RS], a contracting rectangle.

In the case that f^ϵ is given by (1.15) and $\mu^\epsilon = 0$, we need to make the additional assumption

$$0 \in (v_-, v_+). \tag{1.22}$$

We conclude this section with an easy consequence of the assumptions above.

LEMMA 1.2. *Assume* (1.21). *Then along sequences* $\epsilon_n \to 0$, $u^{\epsilon_n} \rightharpoonup u$ *in* $L^\infty(Q_T)$ *weak* $*$ *and* $v^{\epsilon_n} \to v$ *in* $C(Q_T)$.

Proof. The standard weak $*$ compactness of bounded sequences in $L^\infty(Q_T)$ yields that $u^{\epsilon_n} \rightharpoonup u$, $v^{\epsilon_n} \rightharpoonup v$, and $G^{\epsilon_n} = g(u^{\epsilon_n}, v^{\epsilon_n}) \rightharpoonup G$ in $L^\infty(Q_T)$ weak $*$. On the other hand, v^{ϵ_n} is a classical solution of

$$v^{\epsilon_n}_t - b\Delta v^{\epsilon_n} + G^{\epsilon_n} = 0 \quad \text{in } Q_T\,.$$

Classical parabolic theory (see [LUS]) implies that for any $p \in (1, \infty)$, there exists a subsequence (which we again denote by v^{ϵ_n}) $v^{\epsilon_n} \to v$ in $W^{2,1}_p(Q_T)$, where $W^{2,1}_p(Q_T)$ is the usual Sobolev space restricted to Π-periodic functions. (Incidentally, this is one of the places where the periodicity assumption plays a role.) Therefore, v solves (in the sense of distributions) the problem

$$\begin{cases} v_t - b\Delta v + G = 0 & \text{in } Q_T, \\ v = v_0 & \text{on } \mathbb{R}^N \times \{0\}. \end{cases} \tag{1.23}$$

If $p \in (1, \infty)$ is chosen sufficiently large, then $v^{\epsilon_n} \to v$ in $C(Q_T)$ by the standard Sobolev estimates. Finally, the uniqueness of weak solutions to (1.23) yields that the whole sequence $v^{\epsilon_n} \to v$ in $C(Q_T)$. □

Our goal in this paper is to prove a stronger convergence result for the sequence u^{ϵ_n} to find a relationship between u and v and to determine an expression for G.

2. Generalized front propagation.

We begin by recalling the level-set definition of generalized front propagation according to (0.6). Given a closed set $\Gamma_0 \subset \mathbb{R}^N$, $N \geq 2$, choose $\theta_0 \in UC(\mathbb{R}^N)$, where $UC(\Omega)$ denotes the space of uniformly continuous functions defined on Ω, satisfying

$$\Gamma_0 = \{x \in \mathbb{R}^N : \theta_0(x) = 0\} \tag{2.1}$$

and consider the PDEs

$$\begin{cases} \text{(i)} \;\; \theta_t = F(D^2\theta, D\theta, x, t) & \text{in } \mathbb{R}^N \times (0, \infty), \\ \text{(ii)} \;\; \theta = \theta_0 & \text{on } \mathbb{R}^N \times \{t = 0\}. \end{cases} \tag{2.2}$$

The function F is related to V in (0.6) by

$$F(A,p,x,t) = |p|V\left(-\frac{1}{|p|}\left(I - \frac{p\otimes p}{|p|^2}\right)A, -\frac{p}{|p|}, x, t\right) \tag{2.3}$$

for all $x \in \mathbb{R}^N$, $t \geq 0$, $p \in \mathbb{R}^N\backslash\{0\}$, and $A \in S^N$, the space of $N \times N$ symmetric matrices. For the derivation of F, we refer to [ChGG], [BaSS], [IS], etc. As is proved in [ChGG] (see also [BaSS] and [IS] for a relaxation of the assumptions of [ChGG]), the initial value problem (2.2) admits a unique solution $\theta \in UC(\mathbb{R}^N \times (0,\infty))$. (See [BaSS], [ChGG], [ESp], etc. for the relevant definitions, proofs, comments, etc.). Define the closed sets

$$\Gamma_t = \{x \in \mathbb{R}^N : \theta(x,t) = 0\}, \quad t \geq 0, \tag{2.4}$$

and call $\{\Gamma_t\}_{t\geq 0}$ the (level-set) *generalized evolution* according to (0.6) starting from Γ_0. It follows (consult [ChGG], [ESp] in the case Γ_0 is compact, and [BaSS] and [IS] if not) that the definition (2.3) does not depend on the choice of the particular function θ_0 satisfying (2.1). One of the most intriguing questions related to the generalized propagation described above is whether or not the sets Γ_t have interior. This is a rather complicated issue which we do not want to address here. Instead, we refer to [BaSS] for a detailed discussion.

A consequence of the level-set formulation is that the signed distance function to Γ_t satisfies certain equations which we state below. Of course, as mentioned in the introduction, this property of the distance function can also be taken to be the definition of the weak propagation (see [Sol], [BaSS]). To this end, we define the signed distance function $d(x,t)$ from x to Γ_t by

$$d(x,t) = \begin{cases} \operatorname{dist}(x,\Gamma_t) & \text{if } \theta(x,t) > 0, \\ -\operatorname{dist}(x,\Gamma_t) & \text{if } \theta(x,t) < 0, \end{cases} \tag{2.5}$$

where $\operatorname{dist}(x,\Gamma_t)$ is (again) the usual distance from x to Γ_t and $\theta \in UC(\mathbb{R}^N \times (0,\infty))$ is a solution of (2.2) which yields Γ_t.

THEOREM 2.1. *The signed distance function satisfies*

$$\begin{cases} d_t \geq F(D^2d, Dd, x - dDd, t), \\ -(D^2dDd, Dd) \geq 0 \end{cases} \qquad \textit{in } \{d > 0\} \tag{2.6}$$

and

$$\begin{cases} d_t \leq F(D^2d, Dd, x - dDd, t), \\ -(D^2dDd, Dd) \leq 0 \end{cases} \qquad \textit{in } \{d < 0\}. \tag{2.7}$$

For a discussion of the meaning of (2.6) and (2.7) as well as the proof, we refer to [BaSS].

As mentioned in the introduction, one of the most striking applications of the above is the rigorous asymptotics of

$$u_t - \alpha\beta\Delta u + \beta f(x,t,u) = 0 \quad \text{in } \mathbb{R}^N \times (0,\infty).$$

In the next section, we will use the main ideas involved in the proof of the above asymptotics to obtain our main results.

3. The asymptotics $\alpha = \gamma = \epsilon^{-1}$ and $\beta = \epsilon^2$ of the Fitzhugh–Nagumo system.

We consider the system

$$\begin{cases} u_t^\epsilon - \epsilon\Delta u^\epsilon + \frac{1}{\epsilon} f^\epsilon(u^\epsilon, v^\epsilon) = 0, \\ \\ v_t^\epsilon - b\Delta v^\epsilon + g^\epsilon(u^\epsilon, v^\epsilon) = 0, \end{cases} \quad \text{in } \mathbb{R}^N \times (0,\infty) \tag{3.1}$$

with initial condition $(u_0^\epsilon, v_0^\epsilon)$ and we assume (1.2), (1.4), (1.7), (1.8), (1.10), (1.11), (1.13), and (1.17)–(1.21). Lemma 1.2 yields the existence of sequences $\epsilon_n \to 0$ and functions $(u, v) \in L^\infty(Q_T) \times C(Q_T)$ such that $u^{\epsilon_n} \rightharpoonup u$ in $L^\infty(Q_T)$ weak $*$ and $v^{\epsilon_n} \to v$ in $C(Q_T)$. Let $\alpha(v) = \alpha(v, 0)$ be given by (1.8) and consider the front $\{\Gamma_t\}_{t \geq 0}$ defined by the geometric PDEs

$$\begin{cases} \theta_t + \alpha(v(x,t))|D\theta| = 0 & \text{in } \mathbb{R}^N \times (0,\infty), \\ \\ \theta = d_0 & \text{on } \mathbb{R}^N \times \{0\}, \end{cases} \tag{3.2}$$

where d_0 is the signed distance function from Γ_0, which is given by (1.18)–(1.20). Finally, let d be the signed distance function given by (2.5).

THEOREM 3.1. *Assume* (1.2), (1.4), (1.7), (1.8), (1.10), (1.11), (1.13), *and* (1.17)–(1.21) *and consider the functions* (u, v) *given by Lemma* 1.2 *for a given sequence* $\epsilon_n \to 0$, *and the front* Γ_t *which moves with normal velocity* $-\alpha(v)$. *Then*

$$\begin{cases} u = \begin{cases} h_+(v) & \{d > 0\} \\ & \text{in} \\ h_-(v) & \{d < 0\} \end{cases}, \quad u \in [h_-(v), h_+(v)] \quad \text{a.e. on } \Gamma_t, \\ \text{and} \\ u^{\epsilon_n} \to u \text{ locally uniformly in } \{d \neq 0\}. \end{cases} \tag{3.3}$$

Moreover, v satisfies (in the sense of distributions) the singular limit problem

$$v_t - b\Delta v + G(v) = 0 \quad \text{in } Q_T, \tag{3.4}$$

where

(3.5)

$$G(v) = \begin{cases} g(h_-(v), v) & \text{in } \{d < 0\} \\ \\ g(h_+(v), v) & \text{in } \{d > 0\} \end{cases} \quad \text{and} \quad G(v) \in [g(h_+(v), v), g(h_-(v), v)] \quad \text{a.e. on } \Gamma_t.$$

The existence of solutions of (3.4) and (3.5) was proved by Giga, Goto, and Ishii [GGI] by a method completely different from ours.

The assertion of Theorem 3.1 can be substantially strengthened as indicated in Theorem 3.2 below in the case where the front defined by (3.2) does not develop interior. For example, this can happen (cf. [BaSS]) when $\alpha(v(x,t))$ does not change sign. A sufficient condition for this to happen in the case of (0.2) is given by Lemma 1.1. Theorem 3.2 asserts that if interior does not develop, the conclusions of Theorem 3.1 hold for the whole family (u^ϵ, v^ϵ) and not only along sequences.

THEOREM 3.2. *In addition to the assumptions of Theorem* 3.1, *assume that*

$$\operatorname{meas}\{(x,t) : \theta(x,t) = 0\} = 0, \tag{3.6}$$

where θ *is the solution of* (3.2). *Then*

$$u^\epsilon \to \begin{cases} h_+(v) & \\ & \textit{locally uniformly in} \\ h_-(v) & \end{cases} \quad \begin{matrix} \{d > 0\} \\ \\ \{d < 0\} \end{matrix}$$

and

$$v^\epsilon \to v \quad \textit{in } C(Q_T).$$

In what follows, we will present the proofs of Theorems 3.1 and 3.2 under the additional hypothesis that

$$u_0^\epsilon(x) = q^\epsilon\left(\frac{d_0(x)}{\epsilon}, v_0^\epsilon(x), 0\right), \tag{3.7}$$

where d_0 is the signed distance function from Γ_0, $r \to q^\epsilon(r, v_0^\epsilon(x), 0)$ solves (1.5) and (1.6), and $q^\epsilon(0, v, 0) = h_0^\epsilon(v)$. This assumption is made only in order to simplify the presentation below. The general case follows by combining our arguments with those of Chen [Chx].

The basic step of the proof of Theorem 3.1 is based upon an adaptation of the proof presented in [BaSS] for the asymptotics of the scalar reaction-diffusion equation

$$u_t^\epsilon - \epsilon\Delta u^\epsilon + \frac{1}{\epsilon} f^\epsilon(u^\epsilon, x, t) = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty), \tag{3.8}$$

where f^ϵ is of bistable type. The proof is based upon building super- and subsolutions of (3.8) of the form

$$q^{a,\epsilon}\left(\frac{w^{a,\delta}}{\epsilon}, x, t\right),$$

where $q^{a,\epsilon}$ is a traveling wave corresponding to $u \mapsto f^\epsilon(u, x, t) \pm a$ and $w^{a,\delta}$ is some approximation to the signed distance of the limiting front.

In preparation for the proof of Theorem 3.1, we first regularize the function $v \in C(Q_T)$, which is obtained as the limit of v^{ϵ_n} by Lemma 1.2 by means of a sequence $(v^m)_{m\in\mathbf{N}} \subset C^{2,1}(Q_T)$ such that $v^m \to v$ in $C(Q_T)$. Then we approximate (3.2) by

$$\begin{cases} \theta_t^{a,\delta,m} + \alpha(v^m, a)|D\theta^{a,\delta,m}| = 0 & \text{in } \mathbb{R}^N \times (0, \infty), \\ \theta^{a,\delta,m} = d_0 + \delta & \text{on } \mathbb{R}^N \times \{0\}, \end{cases} \tag{3.9}$$

where $\alpha(v^m, a)$ is the limit of $c^\epsilon(v^m, a)$, which exists by (1.8) and $\delta > 0$. Observe that the assumptions on f^ϵ yield that $\alpha(v^m, a) \to \alpha(v)$ as $m \to \infty$ and $a \to 0$. Finally, the stability results of the geometric PDEs (cf. [BaSS]) yield

$$\theta^{a,\delta,m} \to \theta \quad \text{in } C(Q_T) \quad \text{as } m \to \infty, \quad \delta \to 0, \quad a \to 0. \tag{3.10}$$

As in [ESoS] and [BaSS], we introduce the auxiliary function $\eta_\delta : \mathbb{R} \to \mathbb{R}$ satisfying

$$\begin{cases} \eta_\delta \text{ is smooth}, \quad 0 \le \eta_\delta' \le C, \quad |\eta_\delta''| \le C\delta^{-1}, \quad \text{and} \\ \eta_\delta(z) = -\delta \quad \text{if } z \le \delta/4, \quad \eta_\delta(z) = z - \delta \quad \text{if } z \ge \delta/2, \end{cases} \tag{3.11}$$

and define

$$w^{a,\delta,m}(x,t) = \eta_\delta(d^{a,\delta,m}(x,t)),$$

where $x \mapsto d^{a,\delta,m}(x,t)$ is the signed distance function from the set

$$\{y \in \mathbb{R}^N : \theta^{a,\delta,m}(y,t) = 0\}.$$

A straightforward modification of the proof of Lemma 10.1 of [BaSS] yields the following.

LEMMA 3.1. *We have*

$$w_t^{a,\delta,m} + \alpha(v^m(x - w^{a,\delta,m}Dw^{a,\delta,m},t),a)|Dw^{a,\delta,m}| \geq -o(1) \ \ in\ \mathbb{R}^n \times (0,t^*),$$

where t^* is the extinction time of $\{\theta^{a,\delta,m} = 0\}$ and the $o(1)$, as $\delta \to 0$, only depends on the modulus of continuity of $v \in C(Q_T)$ and not on a, δ and m. Finally,

$$|Dw^{a,\delta,m}| = 1 \ \ in\ \left\{d^{a,\delta,m} > \frac{\delta}{2}\right\}.$$

We can now complete the construction of the supersolution which we need for the proof of Theorem 3.1. Since a subsolution can be built by a straightforward adaptation of the above, we omit the details. To this end, let q^ϵ be a traveling wave which corresponds to $u \mapsto f^\epsilon(u,v) - a$ and define

$$\Phi^{a,\delta,m,\epsilon}(x,t) = q^\epsilon\left(\frac{w^{a,\delta,m}(x,t)}{\epsilon}, v^m(x,t), a\right). \tag{3.12}$$

LEMMA 3.2. *Assume* (3.12) *and the assumptions of Theorem* 3.1. *Given $a > 0$, there exists $m_0 = m_0(a)$ such that for all $m \geq m_0$, there exists $\delta_0(a,m)$ such that for all $\delta \leq \delta_0$ there exists $\epsilon_0(a,m,\delta)$ with the property that for all $\epsilon \leq \epsilon_0(a,m,\delta)$, $\Phi^{a,m,\delta,\epsilon}$ is a supersolution of*

$$\phi_t - \epsilon\Delta\phi + \frac{1}{\epsilon}f^\epsilon(\phi,v^\epsilon) \geq 0 \quad in\ Q_T. \tag{3.13}$$

The proof of Lemma 3.2 follows along the lines of the proof of Theorem 9.1 of [BaSS] with few technical adjustments. For completeness we present it below, assuming for simplicity, however, that $w^{a,m,\delta}$ has derivatives, instead of interpreting all statements in the viscosity sense. This can be done easily following [ESoS] and [BaSS]. Finally, to simplify the notation we drop the superscripts whenever it does not create any confusion.

Proof. 1. The equation for the traveling wave yields

$$\begin{aligned} \Phi_t - \epsilon\Delta\Phi + \frac{1}{\epsilon}f^\epsilon(\Phi,v^\epsilon) &= \frac{1}{\epsilon}q_r^\epsilon(w_t - \epsilon\Delta w + c^\epsilon(v^m,a)) + \frac{a}{\epsilon} \\ &- \frac{1}{\epsilon}q_{rr}^\epsilon(|Dw|^2 - 1) + \frac{1}{\epsilon}[f^\epsilon(\Phi,v^\epsilon) - f^\epsilon(\Phi,v^m)] + J, \end{aligned} \tag{3.14}$$

where

$$J = q_v^\epsilon(v_t^m - \epsilon\Delta v^m) - 2q_{rv}^\epsilon Dw \cdot Dv^m - \epsilon q_{vv}^\epsilon|Dv^m|^2,$$

with the above quantities evaluated at $(w(x,t)/\epsilon, v^m(x,t), a)$.

In view of the assumptions on q^ϵ and the choice v^m, it follows that for each fixed m,

$$|\epsilon J| = o(1) \text{ as } \epsilon \to 0$$

uniformly in δ and a.

2. Choose m large and then ϵ small enough so that

$$\begin{gathered}\|v^\epsilon - v\|_\infty + \|v^m - v\|_\infty \le (4K)^{-1}a, \\ \|c^\epsilon(v^m, a) - \alpha(v^m, a)\|_\infty \le 4^{-1}a,\end{gathered}$$

and

$$|\epsilon J| \le 4^{-1}a,$$

where, by (1.4) and (1.21), $\|f_v^\epsilon\| \le K$ in $[-M, M] \times [\overline{v}_- + \rho_1 \overline{v}_+ - \rho]$. Using all of the above, (3.14) yields

$$\Phi_t - \epsilon\Delta\Phi + \frac{1}{\epsilon}f^\epsilon(\Phi, v^\epsilon) \ge \frac{q_r^\epsilon}{\epsilon}(w_t - \epsilon\Delta w + \alpha(v^m, a)) + \frac{a}{4\epsilon} - \frac{q_{rr}^\epsilon}{\epsilon}(|Dw|^2 - 1), \tag{3.15}$$

where again q_r^ϵ and q_{rr}^ϵ are evaluated at $(w/\epsilon, v^m, a)$.

3. We now use the inequalities for w given by Lemma 3.2. We have the following cases.

Case 1: $\delta/2 < d < 2\delta$. Lemma 3.1 yields

$$\Phi_t - \epsilon\Delta\Phi + \frac{1}{\epsilon}f^\epsilon(\Phi, v^\epsilon) \ge \frac{q_r^\epsilon}{\epsilon}\left(-\frac{\epsilon C}{\delta} - (K+1)Co(1)\right) + \frac{a}{4\epsilon} \ge 0$$

for $\delta \le \delta_0(a, m)$ and $\epsilon \le \epsilon_0(a, m, \delta)$, where K is as in (1.10) and since $\Delta d \le Cd^{-1}$ in $\{d > 0\}$.

Case 2: $d \le \delta/2$ *or* $d \ge 2\delta$. If $d \le \delta/2$, then $w \le -\delta/2$, and if $d \ge 2\delta$, then $w \ge \delta$. In either case, (3.15) and Lemma 3.1 yield

$$\Phi_t - \epsilon\Delta\Phi + \frac{1}{\epsilon}f^\epsilon(\Phi, v^\epsilon) \ge \frac{1}{\epsilon}\left[\frac{a}{4} - C|q_{rr}^\epsilon| + q_r^\epsilon\left(-\frac{C\epsilon}{\delta} - Co(1)\right)\right] > 0$$

for $\epsilon \le \epsilon_0(a, m, \delta)$ since by (1.11)(ii), $[|q_r^\epsilon| + |q_{rr}^\epsilon|] \to 0$ as $\epsilon \to 0$ uniformly in a, m, and δ. □

We are now ready to proceed with the proof of Theorem 3.1.

Proof of Theorem 3.1. 1. It follows from (1.14), (1.5), and (1.6), that for a and ϵ sufficiently small, there exist a constant c, independent of v^ϵ, a, and ϵ, and a traveling wave $r \mapsto q^\epsilon(r, v_0^\epsilon(x), a)$ such that

$$q^\epsilon(r, v_0^\epsilon, a) \ge q^\epsilon(r, v_0^\epsilon, 0) + ca\,,$$

where $q^\epsilon(\cdot, v_0^\epsilon, 0)$ is given in (3.7).

The above claim follows easily from our assumptions on the vector field by an appropriate choice of the initial condition $q^\epsilon(0, v, a)$. Indeed, recall that $q^\epsilon(0, v, 0) = h_0^\epsilon(v)$. In view of (1.4), there is a smooth function $\xi^\epsilon(v, a)$, defined for all $v \in [v, v_+]$ and all sufficiently small a and ϵ, such that

$$q^\epsilon(\xi^\epsilon(v, a), v, 0) = \frac{h_-^\epsilon(v, 0) + h_-^\epsilon(v, a)}{2}\,.$$

Now choose $q^\epsilon(\,\cdot\,, v, a)$ so that

$$q^\epsilon(\xi^\epsilon(v,a), v, a) = \frac{h_+^\epsilon(v,0) + h_+^\epsilon(v,a)}{2}\,.$$

The choices of $w^{a,m,\delta}$ and q^ϵ as well as (1.13) now yield

$$\Phi^{a,m,\delta,\epsilon}(x,0) = q^\epsilon\left(\frac{\eta_\delta(d_0^{a,m,\delta}(x))}{\epsilon}, v_0^m, a\right) \ge q^\epsilon\left(\frac{d_0(x)}{\epsilon}, v_0^\epsilon + \frac{a}{4K}, a\right) \ge u_0^\epsilon(x)$$

for small a and ϵ and large m.

2. Lemmas 3.2 and 3.3 and a comparison principle for viscosity solutions of reaction-diffusion equations (see [ESoS]) yield

$$\Phi^{a,m,\delta,\epsilon} \ge u^\epsilon.$$

Using (1.6) and (1.13) we get

$$\text{(3.16)} \qquad u^{\epsilon_n} \le h_+^{\epsilon_n}(v^m, a) \quad \text{for all } (x,t) \quad \text{and} \quad \overline{\lim}\, u^{\epsilon_n} \le h_-(v^m, a) \quad \text{in } \{d < 0\}.$$

The second inequality follows from the remark that if $d(x_0, t_0) < 0$, then $\theta(x_0, t_0) < 0$ and, therefore, $\theta^{a,m,\delta} \le -\xi < 0$ in a neighborhood of (x_0, t_0) for small a, m, δ, and ϵ. But then $d^{a,m,\delta} \le 0$ and, consequently, $w^{a,m,\delta} \le -\delta/2$ in this neighborhood; hence (3.16) follows.

3. Constructing appropriate subsolutions and arguing as before to obtain inequalities analogous to (3.16), we conclude that

$$\text{(3.17)} \qquad u^{\epsilon_n} \longrightarrow \begin{cases} h_+(v) & \text{in } \{d > 0\}, \\ h_-(v) & \text{in } \{d < 0\} \end{cases} \qquad \text{locally uniformly.}$$

Using that $u^{\epsilon_n} \rightharpoonup u$ in $L^\infty(Q_T)$ weak $*$, it is now easy to derive (3.3). Indeed, let $\phi \in C_c^\infty(B((x_0, t_0), \rho))$ be such that $0 \le \varphi$. Then (3.16) yields

$$\int_{Q_T} u^{\epsilon_n} \varphi\, dx dt \le \int_{Q_T} h_+^{\epsilon_n}(v^m, a) \varphi\, dx dt.$$

Letting $n \to +\infty$, $m \to +\infty$, and $a \to 0$ above, we get

$$\int_{Q_T} \varphi(h_+(v) - u)\, dx dt \ge 0,$$

and thus $h_+(v) \ge u$ a.e. in Q_T.

4. To prove (3.4) and (3.5), observe that, by Lemma 1.2, $g^\epsilon(u^\epsilon, v^\epsilon) \rightharpoonup G$ in $L^\infty(Q_T)$ weak $*$. In view of (3.17) and the fact that $v^{\epsilon_n} \to v$ in $C(Q_T)$, it follows immediately that $g^\epsilon \to G$ locally uniformly in $\{d \ne 0\}$, while to obtain (3.5) on Γ_t, we again use (3.16) and (1.2). □

We now proceed with the proof of Theorem 3.2. Here the construction of the super- and subsolutions will be a bit different. The reason is that we need to look at the whole system (3.1), which, in view of (1.2), does not satisfy the assumptions of the maximum principle. To overcome this difficulty, we introduce the function

$$\text{(3.18)} \qquad a(t) = a \exp(St),$$

where

$$S = (3 + L(K+2))L \quad \text{and} \quad a > 0.$$

Here $K > 0$ is such that

$$\|c_a^\epsilon\|_\infty + \|f_v^\epsilon\|_\infty \le K \text{ in } [-M, M] \times [\overline{v}_- + \rho, \overline{v}_+ - \rho],$$

with M given by (1.21), and L is the Lipschitz constant of g^ϵ in $[-M, M] \times [\overline{v}_- + \rho, \overline{v}_+ - \rho]$ and of $h_\pm^\epsilon$ and h_0^ϵ in $[-M, M] \times [\overline{v}_- + \rho, \overline{v}_+ - \rho] \times [0, A]$; recall that we are assuming (1.21).

As before, we introduce the functions

$$\bar{\Phi}^{a,m,\delta,\epsilon}(x,t) = q^\epsilon \left(\frac{w^{a,m,\delta}(x,t)}{\epsilon}, v^m(x,t), (K+1)a(t) \right) \tag{3.19}$$

and

$$\overline{v}(x,t) = v(x,t) + a(t) \quad \text{and} \quad \underline{v}(x,t) = v(x,t) - a(t), \tag{3.20}$$

where v comes from Lemma 1.2 along a fixed subsequence. Finally, let $\underline{\Phi}^{a,m,\delta,\epsilon}$ be defined as $\overline{\Phi}^{a,m,\delta,\epsilon}$ but using $\zeta_\delta(z) = -\eta_\delta(-z)$ in the definition of $w^{a,m,\delta}$.

LEMMA 3.3. *Let u and v be given by Lemma* 1.2 *and assume all the assumptions of Theorem* 3.2. *Then for any $a > 0$, $m \ge m_0(a)$, $\delta \le \delta_0(a,m)$, and $\epsilon \le \epsilon_0(a,m,\delta)$, the functions $\overline{\Phi}^{a,m,\delta,\epsilon}$ and $\underline{\Phi}^{a,m,\delta,\epsilon}$, $\overline{v}$, and $\underline{v}$ satisfy, in the viscosity sense,*

$$\left\{ \begin{array}{l} \text{(i) } \overline{\Phi}_t^{a,m,\delta,\epsilon} - \epsilon \Delta \overline{\Phi}^{a,m,\delta,\epsilon} + \frac{1}{\epsilon} f^\epsilon(\overline{\Phi}^{a,m,d,\epsilon}, \underline{v}) \ge 0, \\ \text{(ii) } \overline{v}_t - b\Delta \overline{v} + g^\epsilon(u,v) - a(t)S = 0 \end{array} \right. \quad \textit{in } Q_T \tag{3.21}$$

and

$$\left\{ \begin{array}{l} \text{(i) } \underline{\Phi}_t^{a,m,\delta,\epsilon} - \epsilon \Delta \underline{\Phi}^{a,m,\delta,\epsilon} + \frac{1}{\epsilon} f^\epsilon(\underline{\Phi}^{a,m,d,\epsilon}, \overline{v}) \le 0, \\ \text{(ii) } \underline{v}_t - b\Delta \underline{v} + g^\epsilon(u,v) + a(t)S = 0 \end{array} \right. \quad \textit{in } Q_T. \tag{3.22}$$

Proof. The proofs of (3.21)(i) and (3.22)(i) follow along the lines of the proof of Lemma 3.2. The main difference is that $\underline{v}$ does not depend on ϵ. Therefore, the proof is not restricted on the particular choice of the subsequence. On the other hand, (3.21)(ii) and (3.22)(ii) hold in the sense of distributions and follow trivially using (3.6). □

Next, we show that the pairs $(\overline{\Phi}^{a,m,\delta,\epsilon}, \overline{v})$ and $(\underline{\Phi}^{a,m,\delta,\epsilon}, \underline{v})$ are some kind of super- and subsolutions of (3.1). To simplify the notation, in what follows, we again drop the explicit dependence on (a, m, δ, ϵ) and write only $(\overline{\Phi}, \overline{v})$ and $(\underline{\Phi}, \underline{v})$.

LEMMA 3.4. *Assume* (1.2) *and* (3.6) *and let $a(t)$ be given by* (3.18). *Then*

$$-g(u,v) + a(t)S \ge_{.} -g(\overline{\Phi}, \overline{v}) \quad \textit{a.e. in } Q_T \tag{3.23}$$

for all $a > 0$, $m \ge m_0(a)$, and $\epsilon \le \epsilon_0(a)$. An analogous inequality holds for $(\underline{\Phi}, \underline{v})$.

Proof. 1. If $d(x,t) > 0$, then $u = h_+(v)$. Choose ϵ small and m large enough so that

$$\|v - v^m\|_\infty \le a \text{ and } |h_\pm^\epsilon - h_\pm| \le a \text{ in } [\overline{v}_- + \rho, \overline{v}_+ - \rho]$$

and compute

$$\begin{aligned}
&- g^\epsilon(u,v) + a(t)S \\
&\quad \geq -g^\epsilon(h_+^\epsilon(v^m,(K+1)a(t)),\overline{v}) - L(|h_+(v) - h_+^\epsilon(v^m,(K+1)a(t))| + a(t)) + a(t)S \\
&\quad \geq -g^\epsilon(h_+^\epsilon(v^m,(K+1)a(t)),\overline{v}) + La \geq -g^\epsilon(\overline{\Phi},\overline{v}).
\end{aligned}$$

2. If $d(x,t) < 0$, then $w^{a,m,\delta}(x,t) \leq -\delta$ for a, δ, and ϵ small and m large enough. Therefore,

$$\overline{\Phi}(x,t) \leq q^\epsilon\left(-\frac{\delta}{\epsilon}, v^m(x,t), (K+1)a(t)\right).$$

Then, by (1.12),

$$\begin{aligned}
&- g(u,v) + a(t)S \\
&= -g(h_-(v),v) + a(t)S \geq -g(h_-^\epsilon(v^m,(K+1)a(t)),\overline{v}) + aL \\
&\geq -g(\overline{\Phi},\overline{v}) + L\left[a - \left(q^\epsilon\left(-\frac{\delta}{\epsilon}, v^m, (K+1)a(t)\right) - h_-^\epsilon(v^m,(K+1)a(t))\right)\right] \\
&\geq -g(\overline{\Phi},\overline{v}) \quad \text{for } \epsilon \leq \epsilon_0(a,m,\delta). \qquad \square
\end{aligned}$$

We now introduce a change of variables in (3.1) in order to control the nonlinear term in the first equation. Indeed, let

$$\lambda^\epsilon = \frac{2}{\epsilon^2} \sup_{[-M,M]^2} |f_u(u,v)|, \tag{3.24}$$

where M is given by (1.21), and define

$$F^\epsilon(t,r,s) = \lambda^\epsilon r + \epsilon^{-1} e^{-\lambda^\epsilon t} f^\epsilon(e^{\lambda^\epsilon t} r, e^{\lambda^\epsilon t} s). \tag{3.25}$$

The following are immediate:

$$\begin{cases}
F_r^\epsilon = \lambda^\epsilon + \epsilon^{-1} f_u^\epsilon(e^{\lambda^\epsilon t} r, e^{\lambda^\epsilon t} s) \geq \lambda^\epsilon/2 \quad \text{if } |r|, |s| \leq e^{-\lambda^\epsilon t} M, \\
F_s^\epsilon = \epsilon^{-1} f_v(e^{\lambda^\epsilon t} r, e^{\lambda^\epsilon t} s) \geq 0.
\end{cases} \tag{3.26}$$

Finally, let

$$\begin{cases}
\hat{u}^\epsilon(x,t) = e^{-\lambda^\epsilon t} u^\epsilon(x,t), \quad \hat{v}^\epsilon(x,t) = e^{-\lambda^\epsilon t} v^\epsilon(x,t), \\
\overline{\varphi}^{a,m,\delta,\epsilon}(x,t) = e^{-\lambda^\epsilon t} \overline{\Phi}^{a,m,\delta,\epsilon}(x,t), \quad \underline{\varphi}^{a,m,\delta,\epsilon}(x,t) = e^{-\lambda^\epsilon t} \underline{\Phi}^{a,m,\delta,\epsilon}(x,t), \\
\tilde{v}(x,t) = e^{-\lambda^\epsilon t} \overline{v}(x,t), \quad \underset{\sim}{v}(x,t) = e^{-\lambda^\epsilon t} \underline{v}(x,t).
\end{cases} \tag{3.27}$$

It is immediate that $(\hat{u}^\epsilon, \hat{v}^\epsilon)$ solves the system

$$\begin{cases}
\text{(i)} \quad \hat{u}_t^\epsilon - \epsilon\Delta\hat{u}^\epsilon + F^\epsilon(t,\hat{u}^\epsilon,\hat{v}^\epsilon) = 0, \\
\text{(ii)} \quad \hat{v}_t^\epsilon - b\Delta\hat{v}^\epsilon + \lambda^\epsilon\hat{v}^\epsilon + e^{-\lambda^\epsilon t} g^\epsilon(u^\epsilon,v^\epsilon) = 0
\end{cases} \quad \text{in } Q_T. \tag{3.28}$$

Moreover, Lemma 3.3 yields that $(\overline{\varphi}^{a,m,\delta,\epsilon}, \underset{\sim}{v})$ satisfies

$$\begin{cases}
\text{(i)} \quad \overline{\varphi}_t - \epsilon\Delta\overline{\varphi} + F^\epsilon(t,\overline{\varphi},\underset{\sim}{v}) \geq 0, \\
\text{(ii)} \quad \underset{\sim}{v}_t - b\Delta\underset{\sim}{v} + \lambda^\epsilon\underset{\sim}{v} = e^{-\lambda^\epsilon t}[-g(u,v) - a(t)S]
\end{cases} \quad \text{in } Q_T, \tag{3.29}$$

where we have once more dropped all of the superscripts. An analogous inequality holds for $(\underline{\varphi}^{a,m,\delta,\epsilon}, \tilde{v})$.

We now proceed as in [Chx] to obtain a comparison result for (3.1), always assuming (3.6).

LEMMA 3.5. *Let the assumptions of Theorem* 3.2 *hold. Then there exists a constant* C, *independent of* ϵ, *such that*

$$\epsilon\lambda^\epsilon \|(\hat{u}^\epsilon - \overline{\varphi}^{a,m,\delta,\epsilon})_+\|_\infty \leq C\|(\underset{\sim}{v} - \hat{v}^\epsilon)_+\|_\infty. \tag{3.30}$$

Proof. The proof is an easy consequence of the assumptions and the above constructions. Indeed, if $(x_0, t_0) \in Q_T$ is a point where $\overline{\varphi} - \hat{u}^\epsilon$ attains a negative minimum—again we drop the superscripts—then, as in the proof of Theorem 3.1, $t_0 > 0$. The assumptions yield that at (x_0, t_0),

$$\begin{aligned} 0 \geq \overline{\varphi}_t - \hat{u}^\epsilon_t &\geq \epsilon(\Delta\overline{\varphi} - \Delta\hat{u}^\epsilon) + F(t_0, \hat{u}^\epsilon, \hat{v}^\epsilon) - F(t_0, \overline{\varphi}, \underset{\sim}{v}) \\ &\geq \frac{\lambda^\epsilon}{2}(\hat{u}^\epsilon - \overline{\varphi}) + F(t_0, \hat{u}^\epsilon, \hat{v}^\epsilon) - F(t_0, \hat{u}^\epsilon, \underset{\sim}{v}) \\ &\geq \frac{\lambda^\epsilon}{2}(\hat{u}^\epsilon - \overline{\varphi}) - \frac{C}{\epsilon}(\underset{\sim}{v} - \hat{v}^\epsilon)_+ , \end{aligned}$$

where $C = \sup_{[-M,M]^2} f_v$. The lemma now easily follows. □

LEMMA 3.6. *Under the assumptions of Theorem* 3.2,

$$\lambda^\epsilon \|(\underset{\sim}{v} - \hat{v}^\epsilon)_+\|_\infty \leq C\|(\underline{\varphi} - \hat{u}^\epsilon)_+\|_\infty \tag{3.31}$$

for some constant C *independent of* ϵ *and for* ϵ *small enough.*

Proof. Here we use (3.28)(ii), (3.29)(ii), and Lemma 3.4. If $V = \underset{\sim}{v} - \hat{v}^\epsilon$, then

$$V_t - b\Delta V + \lambda^\epsilon V = e^{-\lambda^\epsilon t}[a(t)S + g^\epsilon(u^\epsilon, v^\epsilon) - g(u, v)]. \tag{3.32}$$

Multiplying (3.32) by V_+^{p-1} $(p \geq 2)$, integrating over Q_T, and using the periodicity of V, we get

$$\begin{aligned} &\frac{1}{p}\int_\Pi V_+^p dx - \frac{1}{p}\int_\Pi (v_0 - v_0^\epsilon - a)_+^p dx + \int_{Q_T} (b(p-1)V_+^{p-2}|DV_+|^2 + \lambda^\epsilon V_+^p) dxdt \\ &\quad = \int_{Q_T} e^{-\lambda^\epsilon t}[-a(t)S + g^\epsilon(u^\epsilon, v^\epsilon) - g(u, v)]V_+^{p-1} dxdt \\ &\quad \leq \int_{Q_T} e^{-\lambda^\epsilon t}[g^\epsilon(u^\epsilon, v^\epsilon) - g(\underline{\Phi}, \underline{v})]V_+^{p-1} dxdt \\ &\quad = \int_{Q_T} V_+^{p-1}[-I_1(\underline{\Phi}, u^\epsilon)(\underline{\varphi} - \hat{u}^\epsilon) - I_2(\underline{v}, v^\epsilon)(\underset{\sim}{v} - \hat{v}^\epsilon)] dxdt, \end{aligned}$$

where

$$I_1(\underline{\Phi}, u^\epsilon) = \begin{cases} \frac{g(\underline{\Phi},\underline{v}) - g(u^\epsilon,\underline{v})}{\underline{\Phi} - u^\epsilon} & \text{if } \underline{\Phi} \neq u^\epsilon, \\ g_u(u^\epsilon, \underline{v}) & \text{if } u^\epsilon = \underline{\Phi} \end{cases}$$

and

$$I_2(\underline{v}, u^\epsilon) = \begin{cases} \frac{g(u^\epsilon,\underline{v}) - g(u^\epsilon,v^\epsilon)}{\underline{v} - v^\epsilon} & \text{if } \underline{v} \neq v^\epsilon, \\ g_v(u^\epsilon, v^\epsilon) & \text{if } \underline{v} = v^\epsilon. \end{cases}$$

In view of our assumptions, I_1 and I_2 are bounded in Q_T uniformly in ϵ. Finally, if ϵ is so small that $\|v_0^\epsilon - v_0\|_\infty \le a/2$, then

$$\begin{aligned}\lambda^\epsilon \int_{Q_T} V_+^p dxdt &\le C \int_{Q_T} [(\underline{\varphi} - \hat{u}^\epsilon)_+ V_+^{p-1} + V_+^p] dxdt \\ &\le C \left[\|(\underline{\varphi} - \hat{u}^\epsilon)_+\|_p \left(\int_{Q_T} V_+^p dxdt \right)^{1-\frac{1}{p}} + \int_{Q_T} V_+^p dxdt \right].\end{aligned}$$

If ϵ is small enough so that $\lambda^\epsilon - C \ge \lambda^\epsilon/2$, then we conclude the proof, letting $p \to +\infty$. □

Finally, we can state the comparison result for (3.1).

LEMMA 3.7. *Under the assumptions of Theorem* 3.2, *for any $a \le a_0$, $m \ge m_0(a)$, $\delta \le \delta_0(a, m)$, and $\epsilon \le \epsilon_0(a, m, \delta)$,*

$$\underline{\Phi}^{a,m,\delta,\epsilon} \le u^\epsilon \le \overline{\Phi}^{a,m,\delta,\epsilon}$$

and

$$v - a \le v^\epsilon \le v + a.$$

Proof. If, for example, $\overline{\Phi}^{a,m,\delta,\epsilon} - u^\epsilon$ has a negative minimum, then Lemmas 3.5 and 3.6 yield

$$0 < \epsilon(\lambda^\epsilon)^2 \|(\hat{u}^\epsilon - \overline{\varphi})_+\|_\infty \le \lambda^\epsilon C \|(\underline{v} - \hat{v}^\epsilon)_+\|_\infty \le C \|(\underline{\varphi} - \hat{u}^\epsilon)_+\|_\infty,$$

i.e., $\underline{\varphi} - \hat{u}^\epsilon$ has a positive maximum. The comparison for the subsolution pair $(\underline{\Phi}, \underline{v})$ and the definition of λ^ϵ also give

$$0 < \frac{1}{\epsilon^3} \|(\hat{u} - \overline{\varphi})_+\|_\infty \le C \|(\hat{u} - \overline{\varphi})_+\|_\infty,$$

which leads to a contradiction for $\epsilon \to 0$. □

At this point, the proof of Theorem 3.2 follows exactly as the one of Theorem 3.1.

4. The asymptotics $\alpha = \epsilon^{-2}$, $\beta = \epsilon^2$, and $\gamma = \epsilon^{-1}$ of the Fitzhugh–Nagumo system.

Here we present a result about the asymptotics of the system

$$(4.1) \qquad \begin{cases} u_t^\epsilon - \Delta u^\epsilon + \frac{1}{\epsilon^2} f^\epsilon(u^\epsilon, v^\epsilon) = 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{in } \mathbb{R}^N \times (0, \infty), \\ v_t^\epsilon - b\Delta v^\epsilon + g^\epsilon(u^\epsilon, v^\epsilon) = 0 \end{cases}$$

where now we assume that (1.9) holds. The corresponding geometric PDEs is

$$(4.2) \qquad \begin{cases} \theta_t - (\Delta\theta - (D^2\theta D\theta, D\theta)/|D\theta|^2) + \alpha(v)|D\theta| = 0 & \text{in } \mathbb{R}^N \times (0, \infty), \\ \theta = d_0 & \text{on } \mathbb{R}^N \times \{0\}, \end{cases}$$

where v is the limit of the v^ϵ's.

THEOREM 4.1. *Assume* (1.2), (1.4), (1.7), (1.9), (1.10), (1.12), (1.13), *and* (1.17)–(1.21) *and let (u^ϵ, v^ϵ) be the solution of* (4.1). *Then along subsequences, $v^{\epsilon_n} \to v$ in $C(Q_T)$ and $u^{\epsilon_n} \rightharpoonup u$ and $g(u^{\epsilon_n}, v^{\epsilon_n}) \rightharpoonup G$ in $L^\infty(Q_T)$ weak $*$. The pair (u, v) satisfies* (3.3)–(3.5), *where the interface Γ_t now moves with normal velocity equal to its mean* *curvature $\alpha(v)$. Finally, if (3.6) holds for the solution of the corresponding geometric PDE (4.2), then the convergence result holds for the whole family (u^ϵ, v^ϵ).*

The proof of Theorem 4.1 follows along the lines of the proofs of Theorems 3.1 and 3.2 and Proposition 10.2 and Theorem 9.1 of [BaSS]. We therefore omit the details.

REFERENCES

[AC] S. M. Allen and J. W. Cahn, *A macroscopic theory for antiphase boundary motion and its application to antiphase domain coarsening*, Acta Metal., 27 (1979), pp. 1085–1095.

[AW] D. G. Aronson and H. Weinberger, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. Math., 30 (1978), pp. 33–76.

[Ba] G. Barles, *Remark on a flame propagation model*, Rapport 464, Institut National de Recherche Informatique et en Automatique, Le Chesnay, France, 1985.

[BaBS] G. Barles, L. Bronsard, and P. E. Souganidis, *Front propagation for reaction-diffusion equations of bistable type*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 9 (1992), pp. 479–496.

[BaS] G. Barles and P. E. Souganidis, *A new approach to front propagation: Theory and applications,* in preparation.

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

[BrK] L. Bronsard and R. V. Kohn, *Motion by mean curvature as the singular limit of Ginzburgh–Landau model*, J. Differential Equations, 90 (1991), pp. 211–237.

[Ca] G. Caginalp, *Mathematical models of phase boundaries*, in Material Instabilities in Continuum Mechanics and Related Mathematical Problems, J. Ball, ed., Clarendon Press, Oxford, UK, 1988, pp. 35–52.

[Chx] X. Chen, *Generation and propagation of interfaces in reaction diffusion systems*, J. Differential Equations, 96 (1992), pp. 116–141.

[Chxy] X. Y. Chen, *Dynamics of interfaces in reaction diffusion systems*, Hiroshima Math. J., 27 (1991), pp. 47–83.

[ChCS] K. N. Chueh, C. C. Conley, and J. A. Smoller, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373–391.

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

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

[CrIL] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[DeG] E. De Giorgi, *Some conjectures on flow by mean curvature*, in Proc. Capri Workshop, Benevento, Bruno, and Sbordone, eds., Capri, Italy, 1990.

[DeMS] P. De Mottoni and M. Schatzman, *Evolution géometrique d'interfaces*, C. R. Acad. Sci. Paris Sér. I Math., 309 (1989), pp. 453–458.

[ESoS] L. C. Evans, H. M. Soner, and P. E. Souganidis, *Phase transitions and generalized motion by mean curvature*, Comm. Pure Appl. Math., 45 (1992), pp. 1097–1123.

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

[F1] P. C. Fife, *Nonlinear Diffusive Waves*, CBMS Regional Conf. Ser. in Math., Conference Board of the Mathematical Sciences, Washington, DC, 1989.

[F2] ———, *Dynamics of Internal Layers and Diffusive Interfaces*, CBMS–NSF Regional Conf. Ser. in Appl. Math. 53, Society for Industrial and Applied Mathematics, Philadelphia, 1988.

[FH] R. Fitzhugh, *Impulses and physiological states in theoretical models of nerve membranes*, Biophys. J., 1 (1961), pp. 445–466.

[FM] P. C. Fife and B. Mc Leod, *The approach of solutions of nonlinear diffusion equations to travelling solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 335–361.

[GGI] Y. Giga, S. Goto, and H. Ishii, *Global existence of weak solutions for interface equations coupled with diffusion equations*, SIAM J. Math. Anal., 23 (1992), pp. 821–835.

[GGIS] Y. Giga, S. Goto, H. Ishii, and M.-H. Sato, *Comparison principle and convexity properties for singular degenerate parabolic equations on unbounded domains*, Indiana U. Math. J., 40 (1991), pp. 443–470.

[G] S. Goto, *Generalized motion of hypersurfaces with superlinear growth speed in curvature tensors*, Differential Integral Equations, 7 (1994), pp. 323–343.

[H] S. P. Hastings, *Some mathematical problems from neurobiology*, Amer. Math. Monthly, 82 (1975), pp. 881–895.

[I] T. Ilmanen, *Convergence of the Allen–Cahn equation to Brakke's motion by mean curvature*, J. Differential Geom., 38 (1993), pp. 417–461.

[IS] H. Ishii and P. E. Souganidis, *Generalized motion of noncompact hypersurfaces with velocity having arbitrary dependence on the curvature tensors*, Tôhoku Math. J., 47 (1995), pp. 227–250.

[KKR] M. Katsoulakis, G. Kossioris, and F. Reitich, *Generalized motion by mean curvature with Neumann conditions and the Allen–Cahn model for phase transitions*, J. Geom. Anal., to appear.

[KS1] M. Katsoulakis and P. E. Souganidis, *Interacting particle systems and generalized mean curvature evolution*, Arch. Rational Mech. Anal., 127 (1994), pp. 133–157.

[KS2] ———, *Generalized motion by mean curvature as a microscopic limit of stochastic Ising models with long range interactions and Glauber dynamics*, Comm. Math. Phys., 169 (1995), pp. 61–97.

[L] P.-L. Lions, *Optimal control of diffusion processes, part* II: *Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1229–1276.

[LUS] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Uralceva, *Linear and quasilinear equations of parabolic type*, Trans. Amer. Math. Soc., 23 (1968).

[NAY] J. Nagumo, S. Arimoto, and S. Yoshizawa, *An active pulse transmission line simulating nerve axons*, Proc. IRL, 50 (1960), pp. 2061–2070.

[OhJK] T. Ohta, D. Jasnow, and K. Kawasaki, *Universal scaling in the motion of random interfaces*, Phys. Rev. Lett., 49 (1982), pp. 1223–1226.

[OsS] S. Osher and J. Sethian, *Fronts moving with curvature dependent speed: algorithms based on Hamilton–Jacobi equations*, J. Comput. Phys., 79 (1988), pp. 12–49.

[RS] J. Rauch and J. A. Smoller, *Qualitative theory of the Fitzhugh–Nagumo equations*, Adv. Math., 27 (1978), pp. 12–44.

[RSK] J. Rubinstein, P. Sternberg, and J. B. Keller, *Fast reaction, slow diffusion and curve shortening*, SIAM J. Appl. Math., 49 (1989), pp. 116–133.

[So1] H. M. Soner, *Motion of a set by the curvature of its boundary*, J. Differential Equations, 101 (1993), pp. 313–372.

[So2] ———, *Ginzburg–Landau equation and motion by mean curvature* I: *Convergence,* II: *Development of the initial interface*, J. Geom. Anal., to appear.

[Sou] P. E. Souganidis, *Interface dynamics in phase transitions*, in Proc. 1994 ICM, to appear.

[TF] J. Tyson and P. C. Fife, *Target patterns in a realistic model of the Belousov–Zhabatinskii reaction*, J. Chem. Phys., 73 (1980), pp. 2224–2237.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1360–1385, September 1996

010

GINZBURG–LANDAU EQUATIONS AND STABLE SOLUTIONS IN A ROTATIONAL DOMAIN*

SHUICHI JIMBO† AND YOSHIHISA MORITA‡

Abstract. The Ginzburg–Landau (GL) equations, with or without magnetic effect, are studied in the case of a rotational domain in $\mathbb{R}^3$. It can be shown that there exist rotational solutions which describe the physical state of permanent current of electrons in a ring-shaped superconductor. Moreover, if a physical parameter—called the GL parameter—is sufficiently large, then these solutions are stable, that is, they are local minimizers of an energy functional (GL energy). This is proved by the spectral analysis on the linearized equation.

Key words. Ginzburg–Landau equation, stable solutions, rotational domain

AMS subject classifications. 35B35, 35J50, 35B65

1. Introduction.

Ginzburg and Landau [11] proposed a theory of low-temperature superconductivity. In their theory, the (local) energy density $h(\Phi, A)$ inside a superconductor is given by

$$h(\Phi, A) = \frac{1}{2}|(\nabla - iA)\Phi|^2 + \frac{\alpha}{4}(1 - |\Phi|^2)^2 + \frac{1}{2}|\mathrm{rot}A|^2,$$

where Φ ($\mathbb{C}$-valued) is the wave function of the electrons in the material, A is the vector potential of the magnetic field, and $\alpha > 0$ is the Ginzburg–Landau (GL) parameter. We note that this energy density is written in a nondimensional form. A superconducting state (Φ, A) is formulated as a local minimizer of the total energy functional calculated by $h(\Phi, A)$; namely, it is a solution to the variational equation. This variational equation is called the GL equation. After Ginzburg and Landau's work, many mathematical results for the GL equation have arisen and contributed to the study of the existence of solutions and their detailed properties. We are interested in this problem in the case of bounded superconductors. In this direction of study, there are several works concerning various situations (and boundary conditions). Odeh [20] studied the occurrence of a bifurcation from the zero solution as the parameter α varies. Caroll and Glick [6] proved a unique existence of a solution for a certain restricted range of α. Klimov [14] and Bobylev [5] obtained multiple solutions in another range of α. Chen [7] constructed a nonsymmetric solution in a bounded domain in $\mathbb{R}^2$. Yang [25], [26] constructed solutions in a bounded domain in $\mathbb{R}^3$ under an outer magnetic force and analyzed their regularity. Monvel-Berthier, Georgescu, and Pruce [19] gave a detailed characterization of the configuration space with prescribed total vorticity in a bounded domain in $\mathbb{R}^2$. See also [8].

Despite many important works, the stability analysis has not yet been done well. In this paper, we deal with the GL equation in the case where the superconductor is ring-shaped (cf. Fig. 1) and the external (applied) magnetic field is absent. We construct nontrivial solutions and study their stability for large α by analyzing the second variation of the energy functional. We emphasize that the external field is absent but the magnetic field driven by the current of the electrons is taken into the equation. Hence this magnetic effect might influence the stability of the solutions.

* Received by the editors November 30, 1993; accepted for publication (in revised form) March 3, 1995.

† Department of Mathematics, Hokkaido University, Sapporo 060, Japan.

‡ Department of Mathematics and Informatics, Ryukoku University, Ohtsu 520-21, Japan.

However, we can also prove a similar result for the GL equation without the magnetic effect A (cf. equation (1.6) below).

We remark that there are also many works on other important topics. Jaffe and Taubes [12] developed several important methods for analysis of the GL equations, by which they investigated the magnetic screening effect and the quantization phenomena of the total vortices in $\mathbb{R}^2$. Moreover, for the special value of the GL parameter ($\alpha = 1/2$), they solved the prescribed vortices problem in $\mathbb{R}^2$. Berger and Chen [3] constructed a radially symmetric solution (with vortex) in $\mathbb{R}^2$ and showed an elaborate asymptotic behavior for $\alpha \to \infty$. More recently, there have also been extensive studies on the properties of the zero set of the solutions of the equation without the magnetic effect in a domain in $\mathbb{R}^2$ (cf. (1.6) with the boundary condition of the first kind). See [2], [9], [4], and the references therein.

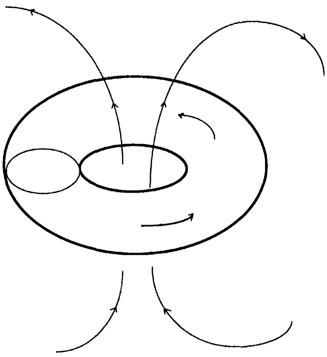


FIG. 1. *Ring-shaped domain* Ω.

We now formulate the problem. Let $\Omega \subset \mathbb{R}^3$ be a bounded domain with a C^3 boundary. We consider the following (GL) functional

$$\mathcal{H}(\Phi, A) = \int_\Omega \left\{ \frac{1}{2}|(\nabla - iA)\Phi|^2 + \frac{\alpha}{4}(1 - |\Phi|^2)^2 \right\} dx + \int_{\mathbb{R}^3} \frac{1}{2}|\mathrm{rot} A|^2 dx. \tag{1.1}$$

Φ is a $\mathbb{C}$-valued function in Ω and A is an $\mathbb{R}^3$-valued function in $\mathbb{R}^3$. The first and second terms correspond to the energy of the electrons confined in Ω and that of the magnetic field caused by the current of electrons, respectively. Note that the magnetic field takes place in the whole space $\mathbb{R}^3$. We suppose $A \in L^2_{\mathrm{loc}}(\mathbb{R}^3; \mathbb{R}^3)$, $\nabla A \in L^2(\mathbb{R}^3; \mathbb{R}^{3\times 3})$, and $\Phi \in H^1(\Omega; \mathbb{C})$. The GL equation is the variational equation of this functional, that is,

$$\begin{cases} (\nabla - iA)^2\Phi + \alpha(1 - |\Phi|^2)\Phi = 0 \quad \text{in} \quad \Omega, \\ \dfrac{\partial \Phi}{\partial \nu} - i\langle A \cdot \nu \rangle \Phi = 0 \quad \text{on} \quad \partial\Omega, \\ \mathrm{rot}\,\mathrm{rot}\, A + \left(i(\overline{\Phi}\nabla\Phi - \Phi\nabla\overline{\Phi})/2 + |\Phi|^2 A \right) \Lambda_\Omega = 0 \quad \text{in} \quad \mathbb{R}^3, \end{cases} \tag{1.2}$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product of vectors in $\mathbb{R}^3$ and $\Lambda_\Omega(x)$ is a discontinuous function such as $\Lambda_\Omega(x) = 1$ for $x \in \Omega$ and $\Lambda_\Omega(x) = 0$ for $x \in \mathbb{R}^3 \setminus \Omega$. We construct nontrivial solutions to (1.2) and prove their stability in the case where Ω is a ring-shaped domain. In the stability analysis, we have to take into account the invariance of gauge transformation. Namely, the transformation $(\Phi, A) \longmapsto (\Phi', A')$ defined by

$$\Phi' = e^{i\rho}\Phi, \qquad A' = A + \nabla\rho \qquad (\rho : \mathbb{R}\text{-valued function in } \mathbb{R}^3) \tag{1.3}$$

leaves $\mathcal{H}$ invariant. Accordingly, if (Φ, A) is a solution to (1.2), (Φ', A') in (1.3) is also a solution of (1.2). Taking various ρ's, we get a continuum of solutions from one solution (Φ, A). We note that all (Φ', A')'s in this continuum correspond to one physical state. In view of this, to see the stability of a solution (Φ, A), we need to study the infinitesimal variation of $\mathcal{H}$ in the direction that is transversal to the continuum at (Φ, A). Consider the second variation of $\mathcal{H}$ obtained by

$$\mathcal{L}(\Phi, A, \Psi, B) = \frac{d^2}{d\epsilon^2}\mathcal{H}(\Phi + \epsilon\Psi, A + \epsilon B)_{|\epsilon=0}. \tag{1.4}$$

We see from the above observation that this quadratic form is degenerate in the tangent space

$$T(\Phi, A) = \{(i\xi\Phi, \nabla\xi) \mid \xi : \mathbb{R}\text{-valued function on } \mathbb{R}^3\}$$

of the continuum at (Φ, A). We take a space $N(\Phi, A)$ which is transversal to $T(\Phi, A)$ and satisfies $T(\Phi, A) \cap N(\Phi, A) = \{(0,0)\}$ and consider whether $\mathcal{L}(\Phi, A, \cdot\,, \cdot\,)$ is positive definite in $N(\Phi, A)$ or not.

In addition to (1.1), we also consider the following functional $\mathcal{H}_0$ (which is given by setting $A = 0$ in (1.1)):

$$\mathcal{H}_0(\Phi) = \int_\Omega \left\{\frac{1}{2}|\nabla\Phi|^2 + \frac{\alpha}{4}(1 - |\Phi|^2)^2\right\} dx. \tag{1.5}$$

In certain situations, the magnetic field is expected to be so small in the interior of Ω that A can be neglected, so this is also a significant model. The variational equation of (1.5) is

$$\begin{cases} \Delta\Phi + \alpha\,(1 - |\Phi|^2)\Phi = 0 & \text{in} \quad \Omega, \\ \dfrac{\partial\Phi}{\partial\nu} = 0 \quad \text{on} \quad \partial\Omega. \end{cases} \tag{1.6}$$

(1.5) and (1.6) are also called the GL functional and the GL equation, respectively. (1.5)–(1.6) has a similar transformation invariance to that of (1.1)–(1.2), that is, if $c \in \mathbb{R}$, the transformation $\Phi \longmapsto \Phi'$ defined by

$$\Phi' = e^{ic}\Phi \tag{1.7}$$

leaves (1.5) and (1.6) invariant. In other words, given any nonzero solution Φ to (1.6), we get a contiuum of solutions $\{e^{ic}\Phi \mid c \in \mathbb{R}\}$ that is a one-dimensional set including Φ. Consider the second variation of $\mathcal{H}_0$ around Φ,

$$\mathcal{L}_0(\Phi, \Psi) = \frac{d^2}{d\epsilon^2}\mathcal{H}_0(\Phi + \epsilon\Psi)_{|\epsilon=0}. \tag{1.8}$$

This quadratic form is degenerate in the direction of the tangent space of the continuum at Φ,

$$T_0(\Phi) = \{ic\Phi \mid c \in \mathbb{R}\}.$$

As in the previous case, we take a space $N_0(\Phi)$ which is transversal to $T_0(\Phi)$ and satisfies $T_0(\Phi) \cap N_0(\Phi) = \{0\}$. We consider $\mathcal{L}_0(\Phi, \cdot\,)$ on $N_0(\Phi)$ and discuss the stability of Φ.

In §2, we formulate several function spaces for the arguments of stability described above. In §3, we specify the rotational domain in $\mathbb{R}^3$ and present our main theorems on the existence of stable solutions. In §4, we construct solutions to (1.2) and (1.6). In §§5 and 6, we prove the stability of the solutions for large α.

2. Formulation. In this section, we formulate $T(\Phi, A)$, $N(\Phi, A)$, $T_0(\Phi)$, and $N_0(\Phi)$ for arbitrary (Φ, A) and Φ. Next, we deduce the concrete expressions of the second variations (1.4) and (1.8) and some of their properties in the above spaces. Let (Φ, A) satisfy $\Phi \in C^1(\overline{\Omega}; \mathbb{C})$, $A \in C^1(\mathbb{R}^3; \mathbb{R}^3)$, and $\nabla A \in L^2(\mathbb{R}^3; \mathbb{R}^{3\times3})$. The solution we construct in §4 satisfies this condition. For convenience, we sometimes deal with the Φ-component in terms of real functions by taking its real and imaginary parts. We set $\Phi = u + vi$ and $\Psi = \phi + \psi i$. Henceforth, we also denote $\mathcal{H}(\Phi, A)$, $\mathcal{L}(\Phi, A, \Psi, B)$, $T(\Phi, A)$, and $N(\Phi, A)$ by $\mathcal{H}(u, v, A)$, $\mathcal{L}(u, v, A, \phi, \psi, B)$, $T(u, v, A)$, and $N(u, v, A)$, respectively. The tangent space $T(\Phi, A)$ is defined as follows:

$$T(\Phi, A) = T(u, v, A) = \{(-v\xi, u\xi, \nabla\xi) \mid \xi \in L^2_{\rm loc}(\mathbb{R}^3),\ \nabla\xi \in H^1(\mathbb{R}^3; \mathbb{R}^3)\}. \tag{2.1}$$

To define a subspace $N(\Phi, A) = N(u, v, A)$ which is transversal to $T(\Phi, A)$, we use the Helmholtz decomposition (cf. [24]). It is known that $L^2(\Omega; \mathbb{R}^3)$ and $L^2(\mathbb{R}^3; \mathbb{R}^3)$ have the following orthogonal decompositions:

$$L^2(\Omega; \mathbb{R}^3) = X_1 \oplus X_2, \quad L^2(\mathbb{R}^3; \mathbb{R}^3) = Y_1 \oplus Y_2,$$

where

$$\begin{aligned} X_1 &= \{\nabla\xi \mid \xi \in L^2(\Omega),\ \nabla\xi \in L^2(\Omega; \mathbb{R}^3)\}, \\ X_2 &= \{B \in L^2(\Omega; \mathbb{R}^3) \mid \operatorname{div} B = 0 \text{ in } H^{-1}(\Omega),\ \langle B \cdot \nu\rangle = 0 \text{ in } H^{-1/2}(\partial\Omega)\}, \\ Y_1 &= \{\nabla\xi \mid \xi \in L^2_{\rm loc}(\mathbb{R}^3),\ \nabla\xi \in L^2(\mathbb{R}^3; \mathbb{R}^3)\}, \\ Y_2 &= \{B \in L^2(\mathbb{R}^3; \mathbb{R}^3) \mid \operatorname{div} B = 0 \text{ in } H^{-1}(\mathbb{R}^3)\}. \end{aligned}$$

Let P and $\widetilde{P}$ be the orthogonal projectors of $L^2(\Omega; \mathbb{R}^3)$ and $L^2(\mathbb{R}^3; \mathbb{R}^3)$ onto X_2 and Y_2, respectively. Let us define

$$\begin{aligned} \overline{N}(u, v, A) &= \left\{(\phi, \psi, B) \in H^1(\Omega)^2 \times H^1(\mathbb{R}^3; \mathbb{R}^3) \mid \int_\Omega (v\phi - u\psi)dx = 0,\ B_{|\Omega} \in X_2\right\}, \\ N(u, v, A) &= \left\{(\phi, \psi, B) \in H^1(\Omega)^2 \times H^1(\mathbb{R}^3; \mathbb{R}^3) \mid \int_\Omega (v\phi - u\psi)dx = 0,\ B \in Y_2\right\}. \end{aligned}$$

For these subspaces, we have the following properties.

Proposition 1.

$$H^1(\Omega) \times H^1(\Omega) \times H^1(\mathbb{R}^3; \mathbb{R}^3) = T(u, v, A) + \overline{N}(u, v, A).$$

Proof. Assume $(u, v) \not\equiv (0, 0)$. Otherwise, the proof is straightforward. Let (ϕ, ψ, B) be any element in the left-hand side. Because $B_{|\Omega} = (I - P)(B_{|\Omega}) + P(B_{|\Omega}) \in X_1 \oplus X_2$, there exists $\xi \in H^2(\Omega)$ such that $\nabla\xi = (I - P)(B_{|\Omega})$. Since $\partial\Omega$ is C^3, there exists $\overline{\xi} \in H^2(\mathbb{R}^3)$ such that $\overline{\xi}_{|\Omega} = \xi$. Set $B_1 = B - \nabla\overline{\xi}$ and

$$c = \int_\Omega (\phi v - u\psi + \overline{\xi}(u^2 + v^2))dx \Big/ \int_\Omega (u^2 + v^2)dx.$$

Setting $\widehat{\xi} = \overline{\xi} - c$, we have

$$(\phi, \psi, B) = \left(-\widehat{\xi}v, \widehat{\xi}u, \nabla\widehat{\xi}\right) + \left(\phi + \widehat{\xi}v, \psi - \widehat{\xi}u, B_1\right) \in T(u, v, A) + \overline{N}(u, v, A). \qquad \square$$

Proposition 2.

$$H^1(\Omega) \times H^1(\Omega) \times H^1(\mathbb{R}^3; \mathbb{R}^3) = T(u, v, A) \oplus N(u, v, A). \tag{2.2}$$

Proof. Assume $(u, v) \not\equiv (0, 0)$. Otherwise, the proof is straightforward. Let (ϕ, ψ, B) be any element in the left-hand side of (2.2). By using the decomposition $B = \nabla\xi + B_1 \in Y_1 \oplus Y_2$, we have

$$(\phi, \psi, B) = (-\xi v, \xi u, \nabla\xi) + (\phi + \xi v, \psi - \xi u, B_1).$$

Modifying ξ by adding an adequate constant to ξ (as in the proof of Proposition 1), we have $(\phi + \xi v, \psi - \xi u, B_1) \in N(u, v, A)$. Hence $(\phi, \psi, B) \in T(u, v, A) + N(u, v, A)$.

If $(-\xi v, \xi u, \nabla \xi)$ belongs to $N(u,v,B)$, ξ is a harmonic function in $\mathbb{R}^3$ and so is $\partial\xi/\partial x_i$ $(1 \le i \le 3)$. Since $\partial\xi/\partial x_i$ is assumed to belong to $L^2(\mathbb{R}^3)$, it is identical to 0 from the Liouville-type theorem. This implies that ξ is a constant function in $\mathbb{R}^3$. On the other hand we have $\int_\Omega (u^2+v^2)\xi dx = 0$ holds and $\xi \equiv 0$ follows. This implies $T(u,v,A) \cap N(u,v,A) = \{(0,0,0)\}$. □

By direct calculation, we can derive a concrete expression of the second variation (1.4).

Formula of the second variation of $\mathcal{H}$.

$$\mathcal{L}(u,v,A,\phi,\psi,B) = \frac{d^2}{d\epsilon^2}\mathcal{H}(u+\epsilon\phi, v+\epsilon\psi, A+\epsilon B)_{|\epsilon=0} \tag{2.3}$$

$$\begin{aligned}
&= \int_\Omega \left\{|\nabla\phi|^2 + |\nabla\psi|^2 - \alpha(1-u^2-v^2)(\phi^2+\psi^2) + 2\alpha(u\phi+v\psi)^2\right\} dx \\
&+ \int_\Omega \left\{A^2(\phi^2+\psi^2) - 2(\phi\langle\nabla\psi\cdot A\rangle - \psi\langle\nabla\phi\cdot A\rangle)\right\} dx \\
&+ \int_{\mathbb{R}^3} |\mathrm{rot}B|^2 dx + \int_\Omega (u^2+v^2)B^2 dx + 4\int_\Omega \langle A\cdot B\rangle(u\phi+v\psi)dx \\
&- 2\int_\Omega \left\{\phi\langle\nabla v\cdot B\rangle - \psi\langle\nabla u\cdot B\rangle + u\langle\nabla\psi\cdot B\rangle - v\langle\nabla\phi\cdot B\rangle\right\} dx.
\end{aligned}$$

Remark. If $\operatorname{div} B = 0$ in Ω and $\langle B\cdot\nu\rangle = 0$ on $\partial\Omega$, then

$$\int_\Omega (u\langle\nabla\psi\cdot B\rangle - v\langle\nabla\phi\cdot B\rangle)\,dx = \int_\Omega (\phi\langle\nabla v\cdot B\rangle - \psi\langle\nabla u\cdot B\rangle)\,dx.$$

This will be used in §6. In the next proposition, we see that the second variation of $\mathcal{H}$ does not depend on the tangential component $T(u,v,A)$. The following property can also be proved by direct calculation.

PROPOSITION 3. *Let* $(\Phi, A) = (u,v,A)$ *be a solution of* (1.2). *Then*

$$\mathcal{L}(u,v,A,\phi,\psi,B) = \mathcal{L}(u,v,A,\phi',\psi',B') \tag{2.4}$$

provided that (ϕ,ψ,B), $(\phi',\psi',B') \in H^1(\Omega)^2 \times H^1(\mathbb{R}^3;\mathbb{R}^3)$, *and* $(\phi-\phi',\psi-\psi',B-B') \in T(u,v,A)$.

We now present a similar formulation for (1.5). Let Φ belong to $C^1(\overline{\Omega};\mathbb{C})$. Similarly to the case of (1.1), we again set $\Phi = u+vi$ and $\Psi = \phi+\psi i$ and denote $\mathcal{H}_0(\Phi)$, $\mathcal{L}_0(\Phi,\Psi)$, $T_0(\Phi)$, and $N_0(\Phi)$ by $\mathcal{H}_0(u,v)$, $\mathcal{L}_0(u,v,\phi,\psi)$, $T_0(u,v)$, and $N_0(u,v)$, respectively. Let us define

$$T_0(\Phi) = T_0(u,v) = \{(-tv,tu) \in H^1(\Omega)\times H^1(\Omega) \mid t\in\mathbb{R}\},$$

$$N_0(\Phi) = N_0(u,v) = \left\{(\phi,\psi)\in H^1(\Omega)\times H^1(\Omega) \mid \int_\Omega (v\phi-u\psi)dx = 0\right\}.$$

We have the following properties.

PROPOSITION 4.

$$H^1(\Omega)\times H^1(\Omega) = T_0(u,v)\oplus N_0(u,v).$$

Proof. The proof is straightforward. □

The following formulas are also proved by a direct calculation.

Formula of the second variation of $\mathcal{H}_0$.

$$\mathcal{L}_0(u,v,\phi,\psi) = \frac{d^2}{d\epsilon^2}\mathcal{H}_0(u+\epsilon\phi, v+\epsilon\psi)_{|\epsilon=0} \tag{2.5}$$

$$= \int_\Omega \left\{ (|\nabla\phi|^2 + |\nabla\psi|^2) - \alpha(1-u^2-v^2)(\phi^2+\psi^2) + 2\alpha(u\phi+v\psi)^2 \right\} dx.$$

PROPOSITION 5. *Let $\Phi = (u,v)$ be a solution of* (1.6). *Then*

$$\mathcal{L}_0(u,v,\phi,\psi) = \mathcal{L}_0(u,v,\phi',\psi')$$

provided that (ϕ,ψ), $(\phi',\psi') \in H^1(\Omega)^2$, and $(\phi-\phi',\psi-\psi') \in T_0(u,v)$.

3. Main results. In this section, we present the main results. Let D be a domain defined by $D \equiv \{(r,z) \in \mathbb{R}^2 \mid r > 0\}$ and let Σ be a bounded domain in D with a C^3 boundary (cf. Fig. 2). Henceforth, we make the following assumption on Σ.

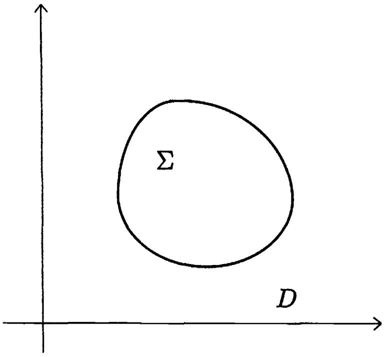


FIG. 2. $\overline{\Sigma} \subset D$.

Assumption. Σ is convex and $\overline{\Sigma} \subset D$.

We now define $\Omega \subset \mathbb{R}^3$ as follows (cf. Fig.1),

$$\Omega = \left\{ (r\cos\theta, r\sin\theta, z) \in \mathbb{R}^3 \mid (r,z) \in \Sigma,\ 0 \le \theta < 2\pi \right\} \simeq \Sigma \times S^1. \tag{3.1}$$

In this section and henceforth, we sometimes use the cylindrical coordinate system (r,θ,z) in $\mathbb{R}^3$ (i.e., $x_1 = r\cos\theta$, $x_2 = r\sin\theta$, $x_3 = z$).

We consider stable rotational solutions to (1.2) and (1.6) for the above domain Ω. In view of the rotationally symmetric situation, we seek for a solution (Φ, A) to (1.2) of the particular form

$$A(r,z,\theta) = Y(r,z)\left(\frac{-\sin\theta}{r}, \frac{\cos\theta}{r}, 0\right), \quad \Phi(r,z,\theta) = W(r,z)e^{im\theta}. \tag{3.2}$$

Here $W = W(r,z) > 0$ and $Y = Y(r,z)$ are real-valued functions in Σ and D, respectively.

The following theorem is the main result of this paper.

THEOREM 6. *Let m be an integer. There exists a $\alpha_0 > 0$ such that for any $\alpha \ge \alpha_0$,* (1.2) *has a solution (Φ_α, A_α) with*

$$\Phi_\alpha(x) = W_\alpha(r,z)e^{im\theta}, \quad A_\alpha(x) = Y_\alpha(r,z)\left(\frac{-\sin\theta}{r}, \frac{\cos\theta}{r}, 0\right),$$

where $W_\alpha \in C^2(\overline{\Sigma})$ and $Y_\alpha \in C^1(D)$, and

$$\lim_{\alpha\to\infty} \sup_{(r,z)\in\Sigma} |W_\alpha(r,z) - 1| = 0. \tag{3.3}$$

Moreover, it is stable in the sense that there exists a constant $\delta > 0$ such that

$$\mathcal{L}(\Phi_\alpha, A_\alpha, \Psi, B) \ge \delta \left(\|\Psi\|^2_{L^2(\Omega;\mathbb{C})} + \|B\|^2_{L^2(\Omega;\mathbb{R}^3)} + \|\nabla B\|^2_{L^2(\mathbb{R}^3;\mathbb{R}^{3\times3})} \right) \tag{3.4}$$

for $(\Psi, B) \in N(\Phi_\alpha, A_\alpha)$ *and* $\alpha \geq \alpha_0$.

We consider a stable solution to (1.6) of the form

$$\Phi(x) = Z(r,z)e^{im\theta} \quad (Z(r,z) > 0), \tag{3.5}$$

and we have the following result.

THEOREM 7. *Let* m *be an integer. Then there exists a constant* $\alpha_1 > 0$ *such that for any* $\alpha \geq \alpha_1$, (1.6) *has a solution* $\Phi_\alpha(x) = Z_\alpha(r,z)e^{im\theta}$ *with*

$$\lim_{\alpha\to\infty} \sup_{(r,z)\in\Sigma} |\, Z_\alpha(r,z) - 1| = 0. \tag{3.6}$$

Moreover, Φ_α *is stable in the sense that there exists a constant* $\delta_0 > 0$ *such that*

$$\mathcal{L}_0(\Phi_\alpha, \Psi) \geq \delta_0 \|\Psi\|^2_{L^2(\Omega;\mathbb{C})} \tag{3.7}$$

for any $\Psi \in N_0(\Phi_\alpha)$ *and* $\alpha \geq \alpha_1$.

We prove these results in §§4–6.

Remark. We can obtain similar theorems for the two-dimensional case, namely for an annulus. The proof can be done in the same manner.

Remark. In our previous paper [13], we constructed a stable solution to (1.6) for a "thin" annulus. Hence Theorem 7 is in an extension of the study in [13].

Remark. With $\Phi = u+vi$ $(u, v:$ real valued$)$, (1.6) becomes a stationary reaction-diffusion system for (u, v). It is interesting to compare our result in Theorem 7 with that obtained by Matano and Mimura [17] for the competition reaction-diffusion system with 2 components. In [17], they constructed nonconstant stable solutions in a special (dumbbell-shaped) domain. Theorem 7 may be regarded as a similar result to theirs. However, there is a significant difference between the competition system (two components) and our case. The competition system does not admit any nonconstant stable solution in an annulus (or ring-shaped domain), whereas the GL equation does.

Remark. Although the range of α in Theorems 6 (i.e., $\alpha \geq \alpha_0$) and 7 (i.e., $\alpha \geq \alpha_1$) might strongly depend on the shape of Ω, it is difficult to specify this range in general. However, it seems to the authors if the ring-shaped domain Ω is very "fat" and its hole is very small, α_0 and α_1 will be taken very large in order that the solutions become stable.

4. Construction of solutions. In this section, we construct solutions to (1.2) and (1.6) of the forms of (3.2) and (3.5), respectively. Moreover, we prove some elaborate asymptotic behaviors of those solutions as $\alpha \to \infty$ (cf. equations (4.2) and (4.4) below), which play essential roles in the stability analysis of the solutions. We begin with (1.6). By direct calculation with (3.5), we get the equation for Z:

$$\begin{cases} L_1 Z - \dfrac{m^2}{r^2} Z + \alpha Z(1-Z^2) = 0 \quad \text{in} \quad \Sigma, \\ \dfrac{\partial Z}{\partial \mathbf{n}} = 0 \quad \text{on} \quad \partial\Sigma, \end{cases} \tag{4.1}$$

where

$$L_1 = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) + \frac{\partial^2}{\partial z^2}$$

and $\mathbf{n}$ is the outward unit normal vector on $\partial\Sigma$ in D.

We have the following result.

PROPOSITION 8. *There exists* $\alpha_0 > 0$ *such that* (4.1) *has a unique solution* $Z =$

$Z_\alpha(r,z) > 0$ *for* $\alpha \geq \alpha_0$ *which satisfies the following asymptotic properties:*

$$(4.2)\qquad \begin{cases} \limsup\limits_{\alpha\to\infty} \sup\limits_{(r,z)\in\Sigma} \alpha|Z_\alpha(r,z)-1| < \infty, \\ \limsup\limits_{\alpha\to\infty} \sup\limits_{(r,z)\in\Sigma} \alpha|\nabla Z_\alpha(r,z)| < \infty, \\ \lim\limits_{\alpha\to\infty} \sup\limits_{(r,z)\in\Sigma} \left|\alpha\,(1 - Z_\alpha(r,z)^2) - \dfrac{m^2}{r^2}\right| = 0. \end{cases}$$

Similarly as above, we get the system of equations for (W,Y) from (1.2) with (3.2):

$$(4.3)\qquad \begin{cases} L_1 W - \dfrac{1}{r^2}(m-Y)^2 W + \alpha W(1-W^2) = 0 \quad \text{in} \quad \Sigma, \\ L_2 Y + (m-Y)W^2 \Lambda_\Sigma = 0 \quad \text{in} \quad D, \\ \dfrac{\partial W}{\partial \mathbf{n}} = 0 \quad \text{on} \quad \partial\Sigma, \quad Y = 0 \quad \text{on} \quad \partial D, \end{cases}$$

where

$$L_2 = \frac{\partial^2}{\partial r^2} - \frac{1}{r}\frac{\partial}{\partial r} + \frac{\partial^2}{\partial z^2}$$

and $\Lambda_\Sigma(r,z) = 1$ for $(r,z) \in \Sigma$ and $\Lambda_\Sigma(r,z) = 0$ for $(r,z) \in D \setminus \Sigma$.

We have the following result.

PROPOSITION 9. *There exists* $\alpha_1 > 0$ *such that* (4.3) *has a solution* $(W,Y) = (W_\alpha(r,z), Y_\alpha(r,z))$ *for* $\alpha > \alpha_1$ *which satisfies the following asymptotic properties:*

$$(4.4)\qquad \begin{cases} \limsup\limits_{\alpha\to\infty} \sup\limits_{(r,z)\in\Sigma} \alpha|W_\alpha(r,z)-1| < \infty, \\ \limsup\limits_{\alpha\to\infty} \sup\limits_{(r,z)\in\Sigma} \alpha|\nabla W_\alpha(r,z)| < \infty, \\ \lim\limits_{\alpha\to\infty} \sup\limits_{(r,z)\in\Sigma} \left|\alpha\,(1 - W_\alpha(r,z)^2) - \dfrac{1}{r^2}(m - Y_\alpha(r,z))^2\right| = 0. \end{cases}$$

We prove these propositions after presenting several auxiliary lemmas.

LEMMA 10. *Let* ρ *be a real-valued function which is* C^3 *in a open set* $E \subset \mathbb{R}^n$. *Then we have*

$$(4.5)\qquad |\nabla\rho|\,\Delta\,|\nabla\rho| \geq \operatorname{grad}\rho\,(\Delta\rho) \quad \textit{in} \quad \{x \in E \mid |\nabla\rho(x)| \neq 0\},$$

where grad ρ *is the following differential operator:*

$$\operatorname{grad}\rho = \sum_{k=1}^n \frac{\partial\rho}{\partial x_k}\frac{\partial}{\partial x_k}.$$

Proof. By direct calculation,

$$\frac{\partial}{\partial x_k}\left(\sum_{j=1}^n \left(\frac{\partial\rho}{\partial x_j}\right)^2\right)^{1/2} = |\nabla\rho|^{-1}\sum_{j=1}^n \frac{\partial\rho}{\partial x_j}\frac{\partial^2\rho}{\partial x_j \partial x_k},$$

$$\begin{aligned} \frac{\partial^2}{\partial x_k^2}|\nabla\rho| &= |\nabla\rho|^{-1}\left(\sum_{j=1}^n \frac{\partial\rho}{\partial x_j}\frac{\partial^3\rho}{\partial x_j \partial x_k^2} + \sum_{j=1}^n \left(\frac{\partial^2\rho}{\partial x_j \partial x_k}\right)^2\right) - |\nabla\rho|^{-3}\left(\sum_{j=1}^n \frac{\partial\rho}{\partial x_j}\frac{\partial^2\rho}{\partial x_j \partial x_k}\right)^2 \\ &\geq |\nabla\rho|^{-1}\sum_{j=1}^n \frac{\partial\rho}{\partial x_j}\frac{\partial^3\rho}{\partial x_j \partial x_k^2}. \end{aligned}$$

Here we used Schwarz's inequality. Summing up the above in $k = 1, 2, \dots, n$, we have the desired inequality. □

LEMMA 11. *Let $E \subset \mathbb{R}^n$ be a domain with C^2 boundary and Γ be a relatively open subset of ∂E. Let $\rho \in C^2(\overline{E})$ be a real-valued function with $\partial\rho/\partial\nu = 0$ on Γ, where ν is the unit outward normal vector on ∂E. Then*

$$\frac{1}{2}\frac{\partial}{\partial\nu}|\nabla\rho|^2 = -h(\operatorname{grad}\rho, \operatorname{grad}\rho) \qquad \text{on} \quad \Gamma, \tag{4.6}$$

where $h(\cdot,\cdot)$ is the second fundamental form of the inclusion $\partial E \hookrightarrow \mathbb{R}^n$ with respect to $-\nu$ (cf. [15, Chap. 7]).

From the Neumann boundary condition of ρ, $\operatorname{grad}\rho_{|\partial\Gamma}$ can be identified with a first-order differential operator on Γ as well as a vector field on Γ.

Proof. First, extend $\nu(x) = (\nu_1(x), \dots, \nu_n(x))$ as a C^2 vector field on some neighborhood of Γ.

$$\frac{1}{2}\frac{\partial}{\partial\nu}|\nabla\rho|^2 = \frac{1}{2}\sum_{k=1}^{n}\nu_k\frac{\partial}{\partial x_k}\sum_{j=1}^{n}\left(\frac{\partial\rho}{\partial x_j}\right)^2 = \sum_{k=1}^{n}\sum_{j=1}^{n}\nu_k\frac{\partial\rho}{\partial x_j}\frac{\partial^2\rho}{\partial x_j\partial x_k}.$$

On the other hand, we have $\sum_{k=1}^{n}\nu_k(x)\partial\rho/\partial x_k = 0$ on Γ. Since $\operatorname{grad}\rho$ is a differential operator on Γ, we can operate $\operatorname{grad}\rho$ on the above equation and we get

$$0 = \operatorname{grad}\rho\left(\sum_{k=1}^{n}\nu_k(x)\partial\rho/\partial x_k\right) = \sum_{k,j}\left(\nu_k\frac{\partial\rho}{\partial x_j}\frac{\partial^2\rho}{\partial x_j\partial x_k} + \frac{\partial\nu_k}{\partial x_j}\frac{\partial\rho}{\partial x_j}\frac{\partial\rho}{\partial x_k}\right).$$

Using this, we have

$$\frac{1}{2}\frac{\partial}{\partial\nu}|\nabla\rho|^2 = -\sum_{j,k}\frac{\partial\nu_k}{\partial x_j}\frac{\partial\rho}{\partial x_j}\frac{\partial\rho}{\partial x_k}.$$

This completes the proof of Lemma 11. □

Proof of Proposition 8. Equation (4.1) is a typical case in the framework given in [1] and [23], so we briefly discuss the existence of a solution. Define two constant functions as follows:

$$Z_{+,\alpha}(r,z) = 1, \quad Z_{-,\alpha}(r,z) = 1 - \frac{d}{\alpha} \quad \text{in} \quad \Sigma.$$

It is easy to check that if $d > 0$ is large enough, $Z_{-,\alpha} \le Z_{+,\alpha}$ are a lower–upper solution pair of (4.1) for $\alpha > d$. Applying [1] or [23], we get a solution Z_α such that $Z_{-,\alpha} \le Z_\alpha \le Z_{+,\alpha}$ in Σ. The uniqueness of the positive solution can be proved in the same manner as in that of Lemma 3.1 in [13], so we omit it. The first estimate in (4.2) directly follows from the inequality

$$1 - \frac{d}{\alpha} \le Z_\alpha(r,z) \le 1 \quad \text{in} \quad \Sigma \tag{4.7}$$

for $\alpha > d$. We prove the remaining estimates of (4.2). Regarding Z_α as a function defined in Ω by $Z(x) = Z(\sqrt{x_1^2 + x_2^2}, x_3)$, we have

$$\Delta_x Z_\alpha - \frac{m^2 Z_\alpha}{x_1^2 + x_2^2} + \alpha Z_\alpha(1 - Z_\alpha^2) = 0 \quad \text{in} \quad \Omega, \tag{4.8}$$

with the Neumann boundary condition on $\partial\Omega$. The nonlinear term in (4.8) is bounded in $C^0(\overline{\Omega})$ for $\alpha > d$ by (4.7), and the Schauder estimate for the elliptic boundary value problem yields that $\{Z_\alpha\}_{\alpha > d}$ is bounded in $C^{1+\gamma}(\overline{\Omega})$, where $\gamma \in [0, 1)$ is an arbitrary constant. This implies that $\{|\nabla Z_\alpha|\}_{\alpha > d}$ is relatively compact in $C^0(\overline{\Omega})$, and hence it follows from (4.7) that $|\nabla Z_\alpha(x)|$ converges to 0 uniformly in $\overline{\Omega}$ as $\alpha \to \infty$. Applying grad Z_α to equation (4.8) with the aid of Lemma 10, we obtain the following differential inequalities:

$$(4.9)\qquad \begin{cases} |\nabla Z_\alpha|\Delta|\nabla Z_\alpha| + g_\alpha(x) \geq 0 \quad \text{in} \quad \Omega\cap G_\alpha, \\ |\nabla Z_\alpha|\dfrac{\partial}{\partial\nu}|\nabla Z_\alpha| = -h(\operatorname{grad} Z_\alpha, \operatorname{grad} Z_\alpha) \leq 0 \quad \text{on} \quad \partial\Omega\cap G_\alpha, \end{cases}$$

where

$$G_\alpha = \{x \in \overline{\Omega} \mid |\nabla Z_\alpha(x)| > 0\},$$

$$g_\alpha(x) = -\frac{m^2|\nabla Z_\alpha|^2}{x_1^2+x_2^2} - \sum_{j=1}^{2}\frac{\partial}{\partial x_j}\left(\frac{m^2}{x_1^2+x_2^2}\right)\frac{\partial Z_\alpha}{\partial x_j}Z_\alpha + \alpha(1-3Z_\alpha^2)|\nabla Z_\alpha|^2.$$

The first inequality in (4.9) is deduced by direct calculation. The second inequality follows from Lemma 11. Recall that grad Z_α is normal to the longitudinal direction of $\partial\Omega$, i.e., $Z_\alpha = Z_\alpha(\sqrt{x_1^2+x_2^2}, x_3)$ is constant in the longitudinal direction. Considering the sign of the second fundamental form and that the cross-section Σ of Ω is convex, we obtain $h(\operatorname{grad} Z_\alpha, \operatorname{grad} Z_\alpha) \geq 0$ on $\partial\Omega$. (4.9) is verified.

We shall prove that $\alpha\|\nabla Z_\alpha\|_{L^\infty}$ is bounded when $\alpha \to \infty$ with the aid of the maximum principle. Z_α is not a constant function and the set F_α defined by

$$F_\alpha = \{x \in \overline{\Omega} \mid 0 < |\nabla Z_\alpha(x)| = \max_{\Omega} |\nabla Z_\alpha|\},$$

is not empty for $\alpha > d$. By virtue of (4.7), there exists $c > 0$ such that $0 \leq \alpha(1-Z_\alpha^2) \leq c$ in Σ for $\alpha > d$. We estimate $|\nabla Z_\alpha|$ in F_α from above. We divide the argument into the following two cases (I and II) of α.

Case I. $F_\alpha \not\subset \partial\Omega$. Take any point $x_0 \in F_\alpha \setminus \partial\Omega$. We have $\Delta|\nabla Z_\alpha| \leq 0$ at $x = x_0$, and we have $g_\alpha(x_0) \geq 0$. By a simple calculation, we get

$$-\frac{m^2|\nabla Z_\alpha|^2}{x_1^2+x_2^2} - \sum_{j=1}^{2}\frac{\partial}{\partial x_j}\left(\frac{m^2}{x_1^2+x_2^2}\right)\frac{\partial Z_\alpha}{\partial x_j}Z_\alpha + \alpha(1-3Z_\alpha^2)|\nabla Z_\alpha|^2 \geq 0 \quad \text{in} \quad F_\alpha,$$

$$-\frac{1}{|\nabla Z_\alpha|}\sum_{j=1}^{2}\frac{\partial}{\partial x_j}\left(\frac{m^2}{x_1^2+x_2^2}\right)\frac{\partial Z_\alpha}{\partial x_j}Z_\alpha + \alpha(1-Z_\alpha^2)|\nabla Z_\alpha| \geq 2\alpha Z_\alpha^2|\nabla Z_\alpha| \quad \text{in} \quad F_\alpha,$$

$$2\alpha Z_\alpha^2|\nabla Z_\alpha| \leq \left(\sum_{j=1}^{2}(\frac{\partial}{\partial x_j}(\frac{m^2}{x_1^2+x_2^2}))^2\right)^{1/2} + 2c|\nabla Z_\alpha|.$$

For $\alpha \geq 2d + 8c$,

$$(4.10)\qquad \alpha|\nabla Z_\alpha| \leq 4\left(\sum_{j=1}^{2}\left(\frac{\partial}{\partial x_j}\left(\frac{m^2}{x_1^2+x_2^2}\right)\right)^2\right)^{1/2} + 2c|\nabla Z_\alpha| \quad \text{in} \quad F_\alpha.$$

Case II. $F_\alpha \subset \partial\Omega$. From Lemma 11,

$$(4.11)\qquad \partial|\nabla Z_\alpha|/\partial\nu \leq 0 \quad \text{on} \quad \partial\Omega\cap G_\alpha.$$

First, we show that $g_\alpha(x) \geq 0$ in F_α. If $g_\alpha(x_0) < 0$ for some $x_0 \in F_\alpha$, we see that $\Delta|\nabla Z_\alpha| > 0$ in some neighborhood of x_0 and we have $\partial|\nabla Z_\alpha|/\partial\nu > 0$ at $x = x_0$ from Hopf's boundary-point lemma (cf. [21]). This is contrary to (4.11). We conclude that g_α is nonnegative in F_α. By this fact, we get a similar estimate to (4.10) in Case I.

This completes the proof of the first two properties in (4.2). From these estimates, $C^1(\Sigma)$ norm of the nonlinear term of (4.1) is bounded when $\alpha \to \infty$. Again applying the Schauder estimate to (4.1), we get that $\|Z_\alpha\|_{C^{2+\gamma}(\overline{\Sigma})}$ is bounded when $\alpha \to \infty$ for any fixed $0 \le \gamma < 1$. This implies that ΔZ_α converges to 0 uniformly in $\overline{\Sigma}$ as $\alpha \to \infty$ and we obtain the last part of (4.2). This completes the proof of Proposition 8. □

Proof of Proposition 9. We remark that for $m = 0$, we can reduce (4.3) to (4.1) by setting $Y \equiv 0$. For $m \neq 0$, if (W, Y) is a solution to (4.3), then $(W, -Y)$ is a solution to (4.3) obtained by replacing m by $-m$. Hence in the construction of solutions, we can assume without loss of generality that m is a positive integer. Since m is positive, (4.3) becomes a so-called cooperation system for functions such that $0 \le W \le 1$, $0 \le Y \le m$. Actually, the nonlinear term in the former equation is nondecreasing in Y and the one in the latter equation is nondecreasing in W provided that W and Y are in the above region. Hence the comparison method (upper–lower solution method) is again applicable. However, in this case, we have the difficulty that D is an unbounded domain and a coefficient of the equation of Y is singular on ∂D (see the operator L_2). We deal with this difficulty by considering an approximation problem by taking a bounded subdomain D_p where the coefficients are bounded and we can thereby get the desired solution in D by taking the limit $p \to \infty$. First, we introduce some auxiliary comparison functions.

$$(4.12) \qquad \begin{cases} W_1(r,z) = 1 - \dfrac{d_1}{\alpha}, \quad W_2(r,z) = 1, \\ Y_1(r,z) = d_2 r^2 e^{-\eta((r-a)^2+(z-b)^2)}, \quad Y_2'(r,z) = \dfrac{r^2}{1+r^2+z^2}, \\ Y_3'(r,z) = \dfrac{r}{(r^2+z^2)^s}, \quad Y_2(r,z) = \min(d_3 Y_2'(r,z), d_3 Y_3'(r,z), m), \end{cases}$$

where $d_1 > 0$, $d_2 > 0$, $d_3 > 0$ and $\eta > 0$ are positive constants and s is a constant such that $4s^2 - 2s - 1 < 0$, $1/2 < s$ (for instance, $s = 3/4$), and (a, b) is an arbitrarily fixed point in Σ. Through easy calculation, we get

(4.13)
$$\begin{cases} L_2 Y_1(r,z) = d_2 \eta r e^{-\eta((r-a)^2+(z-b)^2)} \left(4\eta r(r-a)^2 + 4\eta r(z-b)^2 - 10r + 6a\right), \\ L_2 Y_2'(r,z) = -\dfrac{10r^2 + 2r^2z^2 + 2r^4}{(1+r^2+z^2)^3}, \\ L_2 Y_3'(r,z) = \dfrac{(4s^2-2s-1)r^2 - z^2}{r(r^2+z^2)^{s+1}}. \end{cases}$$

LEMMA 12. *There are constants $d_1 > 0$, $d_2 > 0$, $d_3 > 0$, $\eta > 0$ such that the following inequalities hold for large $\alpha > 0$.*

$$(4.14) \qquad 0 < W_1 \le W_2 \le 1 \quad in \quad \Sigma, \qquad 0 < Y_1 \le Y_2 \le m \quad in \quad D,$$

$$(4.15) \qquad \begin{cases} L_1 W_1 - \dfrac{1}{r^2}(m - Y_1)^2 W_1 + \alpha W_1(1 - W_1^2) \ge 0 \quad in \quad \Sigma, \\ L_2 Y_1 + (m - Y_1) W_1^2 \Lambda_\Sigma \ge 0 \quad in \quad D, \\ \dfrac{\partial W_1}{\partial \mathbf{n}} = 0 \quad on \quad \partial\Sigma, \quad Y_1 = 0 \quad on \quad \partial D, \end{cases}$$

$$(4.16) \qquad \begin{cases} L_1 W_2 - \dfrac{1}{r^2}(m - Y_2)^2 W_2 + \alpha W_2(1 - W_2^2) \le 0 \quad in \quad \Sigma, \\ L_2 Y_2 + (m - Y_2) W_2^2 \Lambda_\Sigma \le 0 \quad in \quad D, \\ \dfrac{\partial W_2}{\partial \mathbf{n}} = 0 \quad on \quad \partial\Sigma, \quad Y_2 = 0 \quad on \quad \partial D, \end{cases}$$

We should remark that Y_2 *does not belong to* $C^2(D)$, *and we have to take the differential inequality of* Y_2 *in* (4.16) *in the weak sense, that is,*

$$\int_D (Y_2 L_2 S + (m - Y_2) W_2^2 \Lambda_\Sigma S) r\, dr dz \le 0 \tag{4.17}$$

for any $S = S(r,z) \in C^\infty(D)$ *satisfying* $S \ge 0$ *in* D *and* $\operatorname{supp} S \Subset D$.

Proof. By taking $\eta > 0$ large, we see that

$$\{(r,z) \in D \mid 4\eta r(r-a)^2 + 4\eta r(z-b)^2 - 10r + 6a < 0\} \subset \Sigma.$$

We take $d_2 > 0$ so small such that $0 < Y_1 \le m$ in D. Then

$$L_1 W_1 - \frac{1}{r^2}(m - Y_1)^2 W_1 + \alpha W_1(1 - W_1^2) \ge W_1 \left(d_1 - \frac{m^2}{r^2}\right) \ge 0$$

for large fixed $d_1 > 0$. We can retake $d_2 > 0$ smaller such that the second inequality of (4.15) is valid. Next, we prove (4.16). The first inequality is trivial. We see that $L_2 Y_2'(r,z)$ and $L_2 Y_3'(r,z)$ are negative in D. Take $d_3 > 0$ large so that $Y_2 = m$ in Σ and $Y_1 \le Y_2$ in D. The first inequality can in the sense of distribution be checked from the definition of Y_2. □

We approximate the domain D. Let $D_p = \{(r,z) \in D \mid 1/p < r,\ r^2 + z^2 < p^2\}$, where $p \in \mathbb{N}$ is a parameter. We consider the following boundary value problem,

$$\begin{cases} L_1 W - \dfrac{1}{r^2}(m - Y)^2 W + \alpha W(1 - W^2) = 0 \quad \text{in} \quad \Sigma, \\ L_2 Y + (m - Y) W^2 \Lambda_\Sigma = 0 \quad \text{in} \quad D_p, \\ \dfrac{\partial W}{\partial \mathbf{n}} = 0 \quad \text{on} \quad \partial\Sigma, \quad Y = Y_1 \quad \text{on} \quad \partial D_p. \end{cases} \tag{4.18}$$

This is a cooperation system in the region $0 < W < 1$, $0 < Y < m$. If we have a lower solution and an upper solution, we can conclude that there exists a solution between them by using the standard theory (cf. [17], [18], [23]). In our case, the situation is a little different from those dealt with in the aforementioned literature because the domains of definition of W and Y are different. However, we can carry out a completely similar argument with using (W_1, Y_1), (W_2, Y_2) as lower–upper solutions and get a solution (W, Y) such that $W_1 \le W \le W_2$ in Σ and $Y_1 \le Y \le Y_2$ in D_p. Thus we have the following approximate sequence of solutions to (4.3).

LEMMA 13. *For large* $p \in \mathbb{N}$, *there exists a solution* $(W^{(p)}, Y^{(p)}) \in C^{2+\gamma}(\overline{\Sigma}) \times C^{1+\gamma}(\overline{D}_p)$ *to* (4.18) *such that*

$$\begin{cases} W_1(r,z) \le W^{(p)}(r,z) \le W_2(r,z) \quad in \quad \Sigma, \\ Y_1(r,z) \le Y^{(p)}(r,z) \le Y_2(r,z) \quad in \quad D_p, \end{cases} \tag{4.19}$$

where $0 \le \gamma < 1$ *is an arbitrarily fixed constant.*

Applying the Schauder estimates with (4.19) to (4.18), we obtain that for any large k, the set of approximate solutions $\{(W^{(p)}, Y^{(p)})\}_{p \ge k+1}$ are relatively compact in $C^2(\overline{\Sigma}) \times C^1(\overline{D}_k)$. Applying the diagonal argument, we get a convergent subsequence and, consequently, a solution $(W_\alpha, Y_\alpha) \in C^2(\overline{\Sigma}) \times C^1(D)$ to equation (4.2) with the same estimate as (4.19), that is,

$$\begin{cases} W_1(r,z) \le W_\alpha(r,z) \le W_2(r,z) \quad \text{in} \quad \Sigma, \\ Y_1(r,z) \le Y_\alpha(r,z) \le Y_2(r,z) \quad \text{in} \quad D, \end{cases} \tag{4.20}$$

for large $\alpha > 0$. The former estimate in (4.20) implies the first property in (4.4). Hence the nonlinear terms of (4.3) are bounded uniformly in α. Using the Schauder estimate, we have a constant $c > 0$ such that

$$|\nabla W_\alpha| + |\nabla Y_\alpha| \le c \quad \text{in} \quad \overline{\Omega} \quad (\text{large } \alpha > 0). \tag{4.21}$$

To deduce the remaining estimates in (4.4), we can carry out a quite similar argument to that in the proof of Proposition 8. Using the orthogonal coordinate (x_1, x_2, x_3), let us set $W_\alpha(x) = W_\alpha((x_1^2 + x_2^2)^{1/2}, x_3)$. The first equation of (4.2) becomes

$$\begin{cases} \Delta W_\alpha - \dfrac{1}{x_1^2 + x_2^2}(m - Y_\alpha)^2 W_\alpha + \alpha W_\alpha(1 - W_\alpha^2) = 0 \quad \text{in} \quad \Omega, \\ \dfrac{\partial W_\alpha}{\partial \nu} = 0 \quad \text{on} \quad \partial\Omega. \end{cases} \tag{4.22}$$

Applying $\operatorname{grad} W_\alpha$ to the above equation and using Lemmas 10 and 11, we get

$$\begin{cases} |\nabla W_\alpha| \Delta |\nabla W_\alpha| + g_\alpha(x) \ge 0 \quad \text{in} \quad \Omega \cap G_\alpha, \\ |\nabla W_\alpha| \dfrac{\partial}{\partial \nu} |\nabla W_\alpha| \le 0 \quad \text{on} \quad \partial\Omega \cap G_\alpha, \end{cases} \tag{4.23}$$

where $G_\alpha = \{x \in \overline{\Omega} \mid |\nabla W_\alpha| > 0\}$ and

$$g_\alpha(x) = -\frac{1}{x_1^2 + x_2^2}(m - Y_\alpha)^2 |\nabla W_\alpha|^2 - \left\langle \nabla W_\alpha \nabla \left(\frac{(m - Y_\alpha)^2}{x_1^2 + x_2^2} \right) \right\rangle W_\alpha$$
$$+ \alpha |\nabla W_\alpha|^2 (1 - 3W_\alpha^2).$$

Consider the set

$$F_\alpha = \left\{ x \in \overline{\Omega} \mid 0 < |\nabla W_\alpha(x)| = \max_{\overline{\Omega}} |\nabla W_\alpha| \right\}$$

and apply the same argument as for Z_α in the proof of Proposition 8; we have $g_\alpha(x) \ge 0$ in F_α for large $\alpha > 0$. Thus we obtain a uniform bound for $\alpha |\nabla W_\alpha|$ in Ω. From this estimate, the $C^1(\overline{\Omega})$ norm of the nonlinear term of (4.22) is bounded when $\alpha \to \infty$. Again from the Schauder estimate, $\{W_\alpha\}_\alpha$ is bounded in $C^{2+\gamma}$ for $0 \le \gamma < 1$. Therefore, ΔW_α uniformly converges to 0 in Ω as $\alpha \to \infty$ and we obtain the desired convergence in the last property of (4.4). □

5. Stability of Φ_α in Theorem 7. In this section, we complete the proof of Theorem 7. We will prove the stability of Φ_α, which was constructed through (4.1) by Proposition 8,

$$\Phi_\alpha(x) = Z_\alpha(r, z) e^{im\theta}. \tag{5.1}$$

To prove the stability of Φ_α in (1.6) for large $\alpha > 0$, we show that $\mathcal{L}_0$ is positive definite in $N_0(\Phi_\alpha)$. For this purpose, we consider the linearized eigenvalue problem of (1.6). For convenience of notation, we discuss the problem in terms of real-valued functions. Let u_α and v_α be the real and the imaginary parts of Φ_α, i.e., $u_\alpha(x) = Z_\alpha(r, z) \cos m\theta$ and $v_\alpha(x) = Z_\alpha(r, z) \sin m\theta$. Thus we consider the following eigenvalue problem:

$$\begin{cases} \Delta \begin{pmatrix} \phi \\ \psi \end{pmatrix} + \alpha \begin{pmatrix} 1 - 3u_\alpha^2 - v_\alpha^2 & -2u_\alpha v_\alpha \\ -2u_\alpha v_\alpha & 1 - u_\alpha^2 - 3v_\alpha^2 \end{pmatrix} \begin{pmatrix} \phi \\ \psi \end{pmatrix} \\ \qquad + \mu \begin{pmatrix} \phi \\ \psi \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \ \text{in } \Omega, \\ \dfrac{\partial \phi}{\partial \nu} = \dfrac{\partial \psi}{\partial \nu} = 0 \quad \text{on} \quad \partial\Omega, \end{cases} \tag{5.2}$$

where

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2} = \frac{1}{r}\frac{\partial}{\partial r}\left(r \frac{\partial}{\partial r} \right) + \frac{\partial^2}{\partial z^2} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}.$$

ϕ and ψ are real-valued functions in Ω. We can show that (5.2) is a self-adjoint eigenvalue problem with real (countable) eigenvalues $\{\mu_\ell(\alpha)\}_{\ell=1}^{\infty}$ which are arranged in increasing order (counting multiplicity). It is easy to see that the set of eigenvalues contains 0 because $(\phi, \psi) = (-v_\alpha, u_\alpha)$ satisfies (5.11) with $\mu = 0$. This is due to the invariance in (1.7). We will prove that $\mu_1(\alpha) = 0$ and $\mu_2(\alpha) > 0$ is bounded away from 0 when $\alpha > 0$ is large. We change the variables as follows:

$$\text{(5.3)} \quad \begin{pmatrix} \widehat{\phi}(r,\theta,z) \\ \widehat{\psi}(r,\theta,z) \end{pmatrix} = R(-m\theta) \begin{pmatrix} \phi(r,\theta,z) \\ \psi(r,\theta,z) \end{pmatrix}, \quad \text{where } R(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

The eigenvalue problem is written in terms of $\widehat{\phi}$ and $\widehat{\psi}$ as follows.

$$\text{(5.4)} \quad \begin{cases} \Delta \begin{pmatrix} \widehat{\phi} \\ \widehat{\psi} \end{pmatrix} - \dfrac{2m}{r^2} \begin{pmatrix} \partial\widehat{\psi}/\partial\theta \\ -\partial\widehat{\phi}/\partial\theta \end{pmatrix} + \left(\alpha(1 - Z_\alpha^2) - \dfrac{m^2}{r^2}\right) \begin{pmatrix} \widehat{\phi} \\ \widehat{\psi} \end{pmatrix} \\ \qquad - 2\alpha Z_\alpha^2 \begin{pmatrix} \widehat{\phi} \\ 0 \end{pmatrix} + \mu \begin{pmatrix} \widehat{\phi} \\ \widehat{\psi} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{in } \Omega, \\ \dfrac{\partial\widehat{\phi}}{\partial\nu} = \dfrac{\partial\widehat{\psi}}{\partial\nu} = 0 \quad \text{on} \quad \partial\Omega, \end{cases}$$

We express $\left(\begin{smallmatrix} \widehat{\phi} \\ \widehat{\psi} \end{smallmatrix}\right)$ in the form of the Fourier expansion

$$\text{(5.5)} \quad \begin{pmatrix} \widehat{\phi}(r,\theta,z) \\ \widehat{\psi}(r,\theta,z) \end{pmatrix} = \frac{1}{\sqrt{2}}\xi_0(r,z) + \sum_{k=1}^{\infty} (\xi_k(r,z)\cos k\theta + \zeta_k(r,z)\sin k\theta),$$

where the vector-valued functions

$$\text{(5.6)} \quad \xi_k(r,z) = \begin{pmatrix} \xi_{k,1}(r,z) \\ \xi_{k,2}(r,z) \end{pmatrix}, \quad \zeta_k(r,z) = \begin{pmatrix} \zeta_{k,1}(r,z) \\ \zeta_{k,2}(r,z) \end{pmatrix}$$

are defined in Σ. Substitute (5.5) into (5.4); we can decompose the eigenvalue problem (5.4) into an infinite series of elliptic eigenvalue problems.

$$\text{(5.7)} \quad \begin{cases} L_1\xi_0 + \left(\alpha(1 - Z_\alpha^2) - \dfrac{m^2}{r^2}\right)\xi_0 - 2\alpha Z_\alpha^2 \begin{pmatrix} \xi_{0,1} \\ 0 \end{pmatrix} + \mu\xi_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ in } \Sigma, \\ \dfrac{\partial}{\partial\mathbf{n}}\xi_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{on} \quad \partial\Sigma. \end{cases}$$

$$\text{(5.8)} \quad \begin{cases} L_1\xi_k - \dfrac{k^2}{r^2}\xi_k - \dfrac{2mk}{r^2}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\zeta_k + \left(\alpha(1 - Z_\alpha^2) - \dfrac{m^2}{r^2}\right)\xi_k \\ \qquad - 2\alpha Z_\alpha^2 \begin{pmatrix} \xi_{k,1} \\ 0 \end{pmatrix} + \mu\xi_k = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ in } \Sigma, \\ L_1\zeta_k - \dfrac{k^2}{r^2}\zeta_k + \dfrac{2mk}{r^2}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\xi_k + \left(\alpha(1 - Z_\alpha^2) - \dfrac{m^2}{r^2}\right)\zeta_k \\ \qquad - 2\alpha Z_\alpha^2 \begin{pmatrix} \zeta_{k,1} \\ 0 \end{pmatrix} + \mu\zeta_k = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{in } \Sigma, \\ \dfrac{\partial}{\partial\mathbf{n}}\xi_k = \dfrac{\partial}{\partial\mathbf{n}}\zeta_k = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{on} \quad \partial\Sigma \quad (k \geq 1). \end{cases}$$

(5.8) is rewritten as follows:

$$\begin{cases} L_1\begin{pmatrix}\xi_{k,1}\\ \zeta_{k,2}\end{pmatrix} - \dfrac{k^2}{r^2}\begin{pmatrix}\xi_{k,1}\\ \zeta_{k,2}\end{pmatrix} - \dfrac{2mk}{r^2}\begin{pmatrix}0&1\\1&0\end{pmatrix}\begin{pmatrix}\xi_{k,1}\\ \zeta_{k,2}\end{pmatrix}\\ \quad + \left(\alpha(1-Z_\alpha^2) - \dfrac{m^2}{r^2}\right)\begin{pmatrix}\xi_{k,1}\\ \zeta_{k,2}\end{pmatrix} - 2\alpha Z_\alpha^2\begin{pmatrix}\xi_{k,1}\\ 0\end{pmatrix} + \mu\begin{pmatrix}\xi_{k,1}\\ \zeta_{k,2}\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix} \text{ in } \Sigma,\\ \dfrac{\partial}{\partial \mathbf{n}}\begin{pmatrix}\xi_{k,1}\\ \zeta_{k,2}\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix} \quad \text{on} \quad \partial\Sigma \qquad (k \geq 1).\end{cases} \tag{5.9}$$

$$\begin{cases} L_1\begin{pmatrix}\zeta_{k,1}\\ -\xi_{k,2}\end{pmatrix} - \dfrac{k^2}{r^2}\begin{pmatrix}\zeta_{k,1}\\ -\xi_{k,2}\end{pmatrix} - \dfrac{2mk}{r^2}\begin{pmatrix}0&1\\1&0\end{pmatrix}\begin{pmatrix}\zeta_{k,1}\\ -\xi_{k,2}\end{pmatrix}\\ \quad + \left(\alpha(1-Z_\alpha^2) - \dfrac{m^2}{r^2}\right)\begin{pmatrix}\zeta_{k,1}\\ -\xi_{k,2}\end{pmatrix} - 2\alpha Z_\alpha^2\begin{pmatrix}\zeta_{k,1}\\ 0\end{pmatrix} + \mu\begin{pmatrix}\zeta_{k,1}\\ -\xi_{k,2}\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix} \text{ in } \Sigma,\\ \dfrac{\partial}{\partial \mathbf{n}}\begin{pmatrix}\zeta_{k,1}\\ -\xi_{k,2}\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix} \quad \text{on} \quad \partial\Sigma \qquad (k \geq 1).\end{cases} \tag{5.10}$$

We see that (5.9) and (5.10) are identical as an eigenvalue problem. Both of them are rewritten as follows:

$$\begin{cases} L_1\begin{pmatrix}\tau\\ \sigma\end{pmatrix} - \dfrac{k^2}{r^2}\begin{pmatrix}\tau\\ \sigma\end{pmatrix} - \dfrac{2mk}{r^2}\begin{pmatrix}0&1\\1&0\end{pmatrix}\begin{pmatrix}\tau\\ \sigma\end{pmatrix}\\ \quad + \left(\alpha(1-Z_\alpha^2) - \dfrac{m^2}{r^2}\right)\begin{pmatrix}\tau\\ \sigma\end{pmatrix} - 2\alpha Z_\alpha^2\begin{pmatrix}\tau\\ 0\end{pmatrix} + \mu\begin{pmatrix}\tau\\ \sigma\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix} \quad \text{in } \Sigma,\\ \dfrac{\partial}{\partial \mathbf{n}}\begin{pmatrix}\tau\\ \sigma\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix} \quad \text{on} \quad \partial\Sigma,\end{cases} \tag{5.11}$$

where σ and τ are real-valued functions in Σ. We can regard (5.7) as the case where $k = 0$ in (5.11). We consider (5.11) for each $k \geq 0$. Let

$$\{\mu_\ell^{(k)}(\alpha)\}_{\ell=1}^\infty \quad \text{and} \quad \left\{\begin{pmatrix}\tau_{\ell,\alpha}^{(k)}\\ \sigma_{\ell,\alpha}^{(k)}\end{pmatrix}\right\}_{\ell=1}^\infty \subset L^2(\Sigma)\times L^2(\Sigma) \tag{5.12}$$

be the set of the eigenvalues arranged in increasing order (counting multiplicity) and the complete system of the corresponding eigenfunctions orthonormalized in $L^2(\Sigma) \times L^2(\Sigma)$. In this and the following sections, $L^2(\Sigma) = L^2(\Sigma; rdrdz)$ is the space of the real-valued square integrable functions with respect to the measure $rdrdz$, and it is the Hilbert space equipped with the inner product with this measure. Thus we obtain

$$\frac{1}{\sqrt{2\pi}}\begin{pmatrix}\tau_{\ell,\alpha}^{(0)}(r,z)\\ \sigma_{\ell,\alpha}^{(0)}(r,z)\end{pmatrix}, \quad \frac{1}{\sqrt{\pi}}\begin{pmatrix}\tau_{\ell,\alpha}^{(k)}(r,z)\cos k\theta\\ \sigma_{\ell,\alpha}^{(k)}(r,z)\sin k\theta\end{pmatrix}, \quad \frac{1}{\sqrt{\pi}}\begin{pmatrix}\tau_{\ell,\alpha}^{(k)}(r,z)\sin k\theta\\ -\sigma_{\ell,\alpha}^{(k)}(r,z)\cos k\theta\end{pmatrix}$$

($k \geq 1$, $\ell \geq 1$), which form a complete orthonormal system of eigenfunctions of (5.4) in $L^2(\Omega) \times L^2(\Omega)$.

We study the asymptotic properties of these eigenvalues and eigenfunctions for $\alpha \to \infty$.

LEMMA 14. *For each nonnegative integer k,*

$$\lim_{\alpha\to\infty} \mu_\ell^{(k)}(\alpha) = \mu_\ell^{(k)} \quad (k \geq 0,\ \ell \geq 1), \tag{5.13}$$

$$\lim_{\alpha\to\infty} \left(\|\nabla\tau_{\ell,\alpha}^{(k)}\|_{L^2(\Sigma)}^2 + \alpha\|\tau_{\ell,\alpha}^{(k)}\|_{L^2(\Sigma)}^2\right) = 0, \tag{5.14}$$

where $\{\mu_\ell^{(k)}\}_{\ell=1}^\infty$ *is the set of the eigenvalues arranged in increasing order (counting multiplicity) of the following eigenvalue problem:*

$$\begin{cases} L_1\sigma - \dfrac{k^2}{r^2}\sigma + \mu\sigma = 0 & in \quad \Sigma, \\ \dfrac{\partial\sigma}{\partial\mathbf{n}} = 0 & on \quad \partial\Sigma. \end{cases} \tag{5.15}$$

Proof. We prove this lemma by the aid of the variational characterization of the eigenvalues. Let $\{\sigma_\ell^{(k)}\}_{\ell=1}^\infty$ be an orthonormal system of eigenfunctions of (5.15) corresponding to $\{\mu_\ell^{(k)}\}_{\ell=1}^\infty$, i.e., $\int_\Sigma \sigma_i^{(k)}\sigma_j^{(k)} rdrdz = \delta_{i,j}$ for $i, j \geq 1$. For each given k, we prove (5.13) and (5.14). For simplicity of notation, we drop the number k and denote $\mu_\ell^{(k)}$, $\mu_\ell^{(k)}(\alpha)$, $\tau_{\ell,\alpha}^{(k)}$, $\sigma_{\ell,\alpha}^{(k)}$, and $\sigma_\ell^{(k)}$ by μ_ℓ, $\mu_\ell(\alpha)$, $\tau_{\ell,\alpha}$, $\sigma_{\ell,\alpha}$, and σ_ℓ, respectively. We prove that for any sequence $\{\alpha_j\}_{j=1}^\infty$ which tends to ∞ as $j \to \infty$, there exists a subsequence on which the limits in (5.13) and (5.14) hold for any ℓ. From (5.11) and the variational characterization of the eigenvalue of the self-adjoint operator (cf. [22]), we have

$$\mu_1(\alpha) = \inf\left\{ J_\alpha(\tau,\sigma) \mid \sigma,\tau \in H^1(\Sigma), \int_\Sigma (\sigma^2+\tau^2) rdrdz = 1 \right\}, \tag{5.16}$$

where

$$J_\alpha(\tau,\sigma) = \int_\Sigma \left(\left(\frac{\partial\sigma}{\partial r}\right)^2 + \left(\frac{\partial\sigma}{\partial z}\right)^2 + \left(\frac{\partial\tau}{\partial r}\right)^2 + \left(\frac{\partial\tau}{\partial z}\right)^2 + \frac{k^2}{r^2}(\sigma^2+\tau^2) + \frac{4mk}{r^2}\sigma\tau \right.$$
$$\left. - \left(\alpha(1-Z_\alpha^2) - \frac{m^2}{r^2}\right)(\sigma^2+\tau^2) + 2\alpha Z_\alpha^2\tau^2 \right) rdrdz.$$

Using the test function $(\tau,\sigma) = (0,\sigma_1)$, we have $\mu_1(\alpha) \leq J_\alpha(0,\sigma_1)$. Using Proposition 8, we obtain

$$\limsup_{\alpha\to\infty} \mu_1(\alpha) \leq \mu_1. \tag{5.17}$$

On the other hand, the eigenfunction $(\tau_{1,\alpha},\sigma_{1,\alpha})$ satisfies

$$\int_\Sigma \left(\left(\frac{\partial\tau_{1,\alpha}}{\partial r}\right)^2 + \left(\frac{\partial\tau_{1,\alpha}}{\partial z}\right)^2 + \frac{k^2}{r^2}\tau_{1,\alpha}^2 + \frac{2mk}{r^2}\sigma_{1,\alpha}\tau_{1,\alpha} \right.$$
$$\left. - \left(\alpha(1-Z_\alpha^2) - \frac{m^2}{r^2}\right)\tau_{1,\alpha}^2 + 2\alpha Z_\alpha^2\tau_{1,\alpha}^2 \right) rdrdz = \mu_1(\alpha)\|\tau_{1,\alpha}\|_{L^2(\Sigma)}^2, \tag{5.18}$$

$$\int_\Sigma \left(\left(\frac{\partial\sigma_{1,\alpha}}{\partial r}\right)^2 + \left(\frac{\partial\sigma_{1,\alpha}}{\partial z}\right)^2 + \frac{k^2}{r^2}\sigma_{1,\alpha}^2 + \frac{2mk}{r^2}\sigma_{1,\alpha}\tau_{1,\alpha} \right.$$
$$\left. - \left(\alpha(1-Z_\alpha^2) - \frac{m^2}{r^2}\right)\sigma_{1,\alpha}^2 \right) rdrdz = \mu_1(\alpha)\|\sigma_{1,\alpha}\|_{L^2(\Sigma)}^2. \tag{5.19}$$

From Proposition 8 and the boundedness of $\mu_1(\alpha)$ (when $\alpha \to \infty$), it follows that

$$\|\tau_{1,\alpha}\|_{L^2(\Sigma)}^2 \sim O(1/\alpha) \qquad (\alpha \to \infty).$$

Considering this fact with regard to (5.18), we obtain

$$\alpha\|\tau_{1,\alpha}\|_{L^2(\Sigma)}^2 \to 0, \quad \|\nabla\tau_{1,\alpha}\|_{L^2(\Sigma)}^2 \to 0 \quad \text{as} \quad \alpha \to \infty,$$

and

$$\lim_{\alpha\to\infty} \|\sigma_{1,\alpha}\|_{L^2(\Sigma)}^2 = 1.$$

In view of (5.19), $\sigma_{1,\alpha}$ is bounded in $H^1(\Sigma)$ and also relatively compact in the weak topology. Thus there exists a subsequence $\{\eta_j\}_j$ of $\{\alpha_j\}_j$ such that σ_{1,η_j} weakly converges to a certain $\widehat{\sigma}_1 \in H^1(\Sigma)$ and $\mu_1(\eta_j)$ converges to a μ' $(\leq \mu_1)$. From the lower semicontinuity of the H^1 norm in the weak convergence, we see

$$\|\widehat{\sigma}_1\|_{L^2(\Sigma)} = 1, \quad \liminf_{j\to\infty} \|\nabla\sigma_{1,\eta_j}\|^2_{L^2(\Sigma)} \geq \|\nabla\widehat{\sigma}_1\|^2_{L^2(\Sigma)}. \tag{5.20}$$

Taking the limit-inf as $j \to \infty$ in (5.19) for $\alpha = \eta_j$, we get

$$\int_\Sigma \left(\left(\frac{\partial\widehat{\sigma}_1}{\partial r}\right)^2 + \left(\frac{\partial\widehat{\sigma}_1}{\partial z}\right)^2 + \frac{k^2}{r^2}\widehat{\sigma}_1^2 \right) rdrdz \leq \mu' \|\widehat{\sigma}_1\|^2_{L^2(\Sigma)}. \tag{5.21}$$

From the variational characterization of μ_1, (5.21) implies that $\mu' \geq \mu_1$, $\mu' = \mu_1$ follows from (5.17), and $\widehat{\sigma}_1$ is a first eigenfunction of (5.15). Therefore, $\lim_{j\to\infty}\mu_1(\eta_j) = \mu_1$ holds. Moreover, we see from (5.19) that

$$\lim_{j\to\infty} \|\nabla\sigma_{1,\eta_j}\|^2_{L^2(\Sigma)} = \|\nabla\widehat{\sigma}_1\|^2_{L^2(\Sigma)}.$$

This concludes the first step of the induction. Next take an element $\widetilde{\sigma} \in \mathrm{L.h.}[\sigma_1,\sigma_2]$ such that $(\widetilde{\sigma}\cdot\widehat{\sigma}_1)_{L^2(\Sigma)} = 0$ and $\|\widetilde{\sigma}\|_{L^2(\Sigma)} = 1$, where $\mathrm{L.h.}[\sigma_1,\sigma_2]$ is the subspace spanned by σ_1, σ_2. Recall that

$$\mu_2(\alpha) = \inf\{J_\alpha(\tau,\sigma) \mid \|\sigma\|^2_{L^2(\Sigma)} + \|\tau\|^2_{L^2(\Sigma)} = 1, \quad (\sigma\cdot\sigma_{1,\alpha})_{L^2(\Sigma)} + (\tau\cdot\tau_{1,\alpha})_{L^2(\Sigma)} = 0\}.$$

By taking the test function $(\tau,\sigma) = (0,\widetilde{\sigma})$, we have

$$\mu_2(\alpha) \leq \frac{J_\alpha(-(\widetilde{\sigma}\cdot\sigma_{1,\alpha})_{L^2(\Sigma)}\tau_{1,\alpha}, \widetilde{\sigma} - (\widetilde{\sigma}\cdot\sigma_{1,\alpha})_{L^2(\Sigma)}\sigma_{1,\alpha})}{\| - (\widetilde{\sigma}\cdot\sigma_{1,\alpha})_{L^2(\Sigma)}\tau_{1,\alpha}\|^2_{L^2(\Sigma)} + \|\widetilde{\sigma} - (\widetilde{\sigma}\cdot\sigma_{1,\alpha})_{L^2(\Sigma)}\sigma_{1,\alpha}\|^2_{L^2(\Sigma)}}. \tag{5.22}$$

By the result obtained in the first step and taking the sequence $\alpha = \eta_j$ $(j = 1,2,3,\ldots)$, we get, by a direct calculation,

$$\limsup_{j\to\infty} \mu_2(\eta_j) \leq \mu_2.$$

From the above inequality and (4.2),

$$\int_\Sigma \Bigg(\left(\frac{\partial\tau_{2,\alpha}}{\partial r}\right)^2 + \left(\frac{\partial\tau_{2,\alpha}}{\partial z}\right)^2 + \frac{k^2}{r^2}\tau_{2,\alpha}^2 + \frac{2mk}{r^2}\sigma_{2,\alpha}\tau_{2,\alpha} - \left(\alpha(1-Z_\alpha^2) - \frac{m^2}{r^2}\right)\tau_{2,\alpha}^2 + 2\alpha Z_\alpha^2\tau_{2,\alpha}^2 \Bigg) rdrdz = \mu_2(\alpha)\|\tau_{2,\alpha}\|^2_{L^2(\Sigma)},$$

$$\int_\Sigma \Bigg(\left(\frac{\partial\sigma_{2,\alpha}}{\partial r}\right)^2 + \left(\frac{\partial\sigma_{2,\alpha}}{\partial z}\right)^2 + \frac{k^2}{r^2}\sigma_{2,\alpha}^2 + \frac{2mk}{r^2}\sigma_{2,\alpha}\tau_{2,\alpha} - \left(\alpha(1-Z_\alpha^2) - \frac{m^2}{r^2}\right)\sigma_{2,\alpha}^2 \Bigg) rdrdz = \mu_2(\alpha)\|\sigma_{2,\alpha}\|^2_{L^2(\Sigma)};$$

we have

$$\eta_j\|\tau_{2,\eta_j}\|^2_{L^2(\Sigma)} \to 0, \qquad \|\nabla\tau_{2,\eta_j}\|^2_{L^2(\Sigma)} \to 0, \qquad \|\sigma_{2,\eta_j}\|^2_{L^2(\Sigma)} \to 1 \quad \text{as } j\to\infty$$

and

$$\limsup_{j\to\infty} \|\sigma_{2,\eta_j}\|_{H^1(\Sigma)} < +\infty.$$

There exist a subsequence $\{\kappa_j\}_{j=1}^\infty \subset \{\eta_j\}_{j=1}^\infty$, μ'' $(\leq \mu_2)$, and $\widehat{\sigma}_2 \in H^1(\Sigma)$ such that

$$\lim_{j\to\infty} \mu_2(\kappa_j) = \mu'', \quad \lim_{j\to\infty} \sigma_{2,\kappa_j} = \widehat{\sigma}_2 \quad \text{weakly in} \quad H^1(\Sigma).$$

It follows that

$$\|\widehat{\sigma}_2\|_{L^2(\Sigma)} = 1, \quad (\widehat{\sigma}_2, \widehat{\sigma}_1)_{L^2(\Sigma)} = 0, \tag{5.23}$$

$$\int_\Sigma \left(\left(\frac{\partial \widehat{\sigma}_2}{\partial r} \right)^2 + \left(\frac{\partial \widehat{\sigma}_2}{\partial z} \right)^2 + \frac{k^2}{r^2} \widehat{\sigma}_2^2 \right) r dr dz \le \mu'' \|\widehat{\sigma}_2\|^2_{L^2(\Sigma)}. \tag{5.24}$$

By this we conclude that $\mu'' \ge \mu_2$. As in the first step, we obtain that $\mu'' = \mu_2$, $\widehat{\sigma}_2$ is a second eigenfunction of (5.15), and, moreover,

$$\lim_{j \to \infty} \|\nabla \sigma_{2,\kappa_j}\|^2_{L^2(\Sigma)} = \|\nabla \widehat{\sigma}_2\|^2_{L^2(\Sigma)}.$$

For higher eigenvalues ($\ell \ge 3$), we can repeat the similar argument inductively. Consequently, for each ℓ, there exists a subsequence $\{\alpha_{\ell,j}\}_{j=1}^\infty \subset \{\alpha_j\}_{j=1}^\infty$ such that

$$\lim_{j \to \infty} \mu_\ell(\alpha_{\ell,j}) = \mu_\ell, \tag{5.25}$$

$$\lim_{j \to \infty} \left(\|\nabla \tau_{\ell,\alpha_{\ell,j}}\|^2_{L^2(\Sigma)} + \alpha_{\ell,j} \|\tau_{\ell,\alpha_{\ell,j}}\|^2_{L^2(\Sigma)} \right) = 0. \tag{5.26}$$

From the arbitrariness of the sequence $\{\alpha_j\}_{j=1}^\infty$, (5.13) and (5.14) hold. □

Now we are in a position to prove the stability of Φ_α.

Proof of Theorem 7. From (5.15), it is easy to see that

$$\mu_\ell^{(k)} \ge \mu_1^{(k)} \ge \frac{k^2}{r_0^2} \quad \text{and} \quad \mu_\ell^{(k)} \ge \mu_2^{(0)} > 0 \quad (k \ge 0,\ \ell \ge 2),$$

where $r_0 = \inf\{r \mid (r,z) \in \Sigma\} > 0$. By Lemma 14, all $\mu_{\ell,\alpha}^{(k)}$'s except for $\mu_{1,\alpha}^{(0)}$ are positive and bounded away from 0 when $\alpha \to \infty$. On the other hand, $\mu_{1,\alpha}^{(0)} = 0$ because we can take $(\tau, \sigma) = (0, Z_\alpha)$ in (5.11) with $\mu = 0$ and $k = 0$. Then there exist constants $\delta_0 > 0$ and $\alpha_* > 0$ such that

$$J_\alpha(\widehat{\phi}, \widehat{\psi}) \ge \delta_0 \left(\|\widehat{\phi}\|^2_{L^2(\Omega)} + \|\widehat{\psi}\|^2_{L^2(\Omega)} \right) \quad \text{for} \quad (\widehat{\phi}, \widehat{\psi}) \in H^1(\Omega) \times H^1(\Omega) \text{ with } \int_\Omega Z_\alpha \widehat{\psi} dx = 0,$$

for $\alpha > \alpha_*$. Translating this inequality into the one in terms of (ϕ, ψ) (see (5.3)), we obtain

$$\mathcal{L}_0(\phi, \psi) \ge \delta_0 \left(\|\phi\|^2_{L^2(\Omega)} + \|\psi\|^2_{L^2(\Omega)} \right)$$
$$\text{for} \quad (\phi, \psi) \in H^1(\Omega) \times H^1(\Omega) \text{ with } \int_\Omega (\phi v_\alpha - \psi u_\alpha) dx = 0$$

for $\alpha > \alpha_*$. This completes the proof of Theorem 7. □

6. Stability of (Φ_α, A_α) in Theorem 6. In this section, we prove the stability of (Φ_α, A_α), which we constructed in §4.

$$\begin{cases} A_\alpha(x) = Y_\alpha(r,z) \left(\dfrac{-\sin\theta}{r}, \dfrac{\cos\theta}{r}, 0 \right), \\ \Phi_\alpha(x) = W_\alpha(r,z) e^{im\theta}. \end{cases} \tag{6.1}$$

As is the case in §5, we express Φ_α in terms of real-valued functions, i.e., we put $u_\alpha(x) = W_\alpha(r,z)\cos m\theta$, $v_\alpha(x) = W_\alpha(r,z)\sin m\theta$. We estimate the second variation $\mathcal{L}(\phi, \psi, B)$ on $\overline{N}(\Phi_\alpha, A_\alpha) = \overline{N}(u_\alpha, v_\alpha, A_\alpha)$ from below. We change the variables ϕ and ψ into $\widehat{\phi}$ and $\widehat{\psi}$ by

$$\begin{pmatrix} \widehat{\phi} \\ \widehat{\psi} \end{pmatrix} = R(-m\theta) \begin{pmatrix} \phi \\ \psi \end{pmatrix} \quad \text{(cf. (5.3))}.$$

Using formula (2.3), we can express $\mathcal{L}(u_\alpha, v_\alpha, A_\alpha, \phi, \psi, B)$ in terms of $\widehat{\phi}$, $\widehat{\psi}$, and B for the solution $(\Phi_\alpha, A_\alpha) = (u_\alpha, v_\alpha, A_\alpha)$. If $\operatorname{div} B = 0$ and $\langle B \cdot \nu \rangle = 0$ on $\partial\Omega$ (this is valid for $(\phi, \psi, B) \in \overline{N}(u_\alpha, v_\alpha, A_\alpha)$), the second variation is written concretely as follows:

$$\mathcal{L}(u_\alpha, v_\alpha, A_\alpha, \phi, \psi, B) = I_1(\widehat{\phi}, \widehat{\psi}) + I_2(B) + I_3(\widehat{\phi}, \widehat{\psi}, B),$$

where

$$\begin{aligned} &I_1(\widehat{\phi}, \widehat{\psi}) \\ &= \int_\Omega \left(\left(\frac{\partial \widehat{\phi}}{\partial r}\right)^2 + \left(\frac{\partial \widehat{\phi}}{\partial z}\right)^2 + \left(\frac{\partial \widehat{\psi}}{\partial r}\right)^2 + \left(\frac{\partial \widehat{\psi}}{\partial z}\right)^2 - \alpha(1 - W_\alpha^2)(\widehat{\phi}^2 + \widehat{\psi}^2) + 2\alpha W_\alpha^2 \widehat{\phi}^2 \right) dx \\ &\quad + \int_\Omega \left(\frac{(Y_\alpha - m)^2}{r^2} (\widehat{\phi}^2 + \widehat{\psi}^2) + \frac{1}{r^2} \left(\left(\frac{\partial \widehat{\phi}}{\partial \theta}\right)^2 + \left(\frac{\partial \widehat{\psi}}{\partial \theta}\right)^2 \right) + \frac{2(Y_\alpha - m)}{r^2} \left(\widehat{\psi} \frac{\partial \widehat{\phi}}{\partial \theta} - \widehat{\phi} \frac{\partial \widehat{\psi}}{\partial \theta} \right) \right) dx, \end{aligned}$$

$$I_2(B) = \int_{\mathbb{R}^3} |\operatorname{rot} B|^2 dx + \int_\Omega W_\alpha^2 B^2 dx,$$

$$I_3(\widehat{\phi}, \widehat{\psi}, B) = 4 \int_\Omega \left(\frac{(Y_\alpha - m) S_1 W_\alpha \widehat{\phi}}{r^2} + S_2 \frac{\partial W}{\partial r} \widehat{\psi} + S_3 \frac{\partial W}{\partial z} \widehat{\psi} \right) dx,$$

where S_1, S_2, and S_3 are defined from B through

$$B = \left(-S_1 \frac{\sin\theta}{r} + S_2 \cos\theta, S_1 \frac{\cos\theta}{r} + S_2 \sin\theta, S_3 \right).$$

To investigate the coerciveness of I_1, we consider the eigenvalue problem

$$\text{(6.2)} \qquad \begin{cases} \Delta \begin{pmatrix} \widehat{\phi} \\ \widehat{\psi} \end{pmatrix} + \dfrac{2}{r^2}(Y_\alpha - m) \begin{pmatrix} \partial\widehat{\psi}/\partial\theta \\ -\partial\widehat{\phi}/\partial\theta \end{pmatrix} - \dfrac{1}{r^2}(Y_\alpha - m)^2 \begin{pmatrix} \widehat{\phi} \\ \widehat{\psi} \end{pmatrix} \\ \qquad + \alpha(1 - W_\alpha^2) \begin{pmatrix} \widehat{\phi} \\ \widehat{\psi} \end{pmatrix} - 2\alpha W_\alpha^2 \begin{pmatrix} \widehat{\phi} \\ 0 \end{pmatrix} + \mu \begin{pmatrix} \widehat{\phi} \\ \widehat{\psi} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ in } \Omega, \\ \dfrac{\partial \widehat{\phi}}{\partial \nu} = \dfrac{\partial \widehat{\psi}}{\partial \nu} = 0 \quad \text{on} \quad \partial\Omega. \end{cases}$$

As in the proof of Theorem 7, we express $\binom{\widehat{\phi}}{\widehat{\psi}}$ in the Fourier expansion as follows:

$$\text{(6.3)} \qquad \begin{pmatrix} \widehat{\phi}(r, \theta, z) \\ \widehat{\psi}(r, \theta, z) \end{pmatrix} = \frac{1}{\sqrt{2}} \xi_0(r, z) + \sum_{k=1}^{\infty} \left(\xi_k(r, z) \cos k\theta + \zeta_k(r, z) \sin k\theta \right),$$

where the real vector functions

$$\text{(6.4)} \qquad \xi_k(r, z) = \begin{pmatrix} \xi_{k,1}(r, z) \\ \xi_{k,2}(r, z) \end{pmatrix} \quad (k \geq 0), \quad \zeta_k(r, z) = \begin{pmatrix} \zeta_{k.1}(r, z) \\ \zeta_{k,2}(r, z) \end{pmatrix} \quad (k \geq 1).$$

The eigenvalue problem (6.2) is decomposed into the following:

$$\text{(6.5)} \qquad \begin{cases} L_1 \begin{pmatrix} \xi_{k,1} \\ \zeta_{k,2} \end{pmatrix} - \dfrac{k^2}{r^2} \begin{pmatrix} \xi_{k,1} \\ \zeta_{k,2} \end{pmatrix} + \left(\alpha(1 - W_\alpha^2) - \dfrac{1}{r^2}(Y_\alpha - m)^2 \right) \begin{pmatrix} \xi_{k,1} \\ \zeta_{k,2} \end{pmatrix} \\ \quad - \dfrac{2k}{r^2}(Y_\alpha - m) F \begin{pmatrix} \xi_{k,1} \\ \zeta_{k,2} \end{pmatrix} - 2\alpha W_\alpha^2 \begin{pmatrix} \xi_{k,1} \\ 0 \end{pmatrix} + \mu \begin{pmatrix} \xi_{k,1} \\ \zeta_{k,2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ in } \Sigma, \\ \dfrac{\partial}{\partial \mathbf{n}} \begin{pmatrix} \xi_{k,1} \\ \zeta_{k,2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{on} \quad \partial\Sigma \end{cases}$$

and

$$\text{(6.6)}\quad \begin{cases} L_1\begin{pmatrix}\zeta_{k,1}\\ -\xi_{k,2}\end{pmatrix} - \dfrac{k^2}{r^2}\begin{pmatrix}\zeta_{k,1}\\ -\xi_{k,2}\end{pmatrix} + \left(\alpha(1-W_\alpha^2) - \dfrac{1}{r^2}(Y_\alpha - m)^2\right)\begin{pmatrix}\zeta_{k,1}\\ -\xi_{k,2}\end{pmatrix} \\ \quad - \dfrac{2k}{r^2}(Y_\alpha - m)F\begin{pmatrix}\zeta_{k,1}\\ -\xi_{k,2}\end{pmatrix} - 2\alpha W_\alpha^2\begin{pmatrix}\zeta_{k,1}\\ 0\end{pmatrix} + \mu\begin{pmatrix}\zeta_{k,1}\\ -\xi_{k,2}\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix} \text{ in }\Sigma, \\ \dfrac{\partial}{\partial \mathbf{n}}\begin{pmatrix}\zeta_{k,1}\\ -\xi_{k,2}\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix} \quad \text{on} \quad \partial\Sigma, \end{cases}$$

respectively, where $F = \left(\begin{smallmatrix}0 & 1\\ 1 & 0\end{smallmatrix}\right)$. It is easy to see that (6.5) and (6.6) are equivalent to each other as eigenvalue problems. Both of them are rewritten as follows:

$$\text{(6.7)}\quad \begin{cases} L_1\begin{pmatrix}\tau\\ \sigma\end{pmatrix} - \dfrac{k^2}{r^2}\begin{pmatrix}\tau\\ \sigma\end{pmatrix} + \left(\alpha(1-W_\alpha^2) - \dfrac{1}{r^2}(Y_\alpha - m)^2\right)\begin{pmatrix}\tau\\ \sigma\end{pmatrix} \\ \qquad - \dfrac{2k}{r^2}(Y_\alpha - m)F\begin{pmatrix}\tau\\ \sigma\end{pmatrix} - 2\alpha W_\alpha^2\begin{pmatrix}\tau\\ 0\end{pmatrix} + \mu\begin{pmatrix}\tau\\ \sigma\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix} \text{ in }\Sigma, \\ \dfrac{\partial}{\partial \mathbf{n}}\begin{pmatrix}\tau\\ \sigma\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix} \quad \text{on} \quad \partial\Sigma. \end{cases}$$

Let

$$\{\mu_\ell^{(k)}(\alpha)\}_{\ell=1}^\infty \quad \text{and} \quad \left\{\begin{pmatrix}\tau_{\ell,\alpha}^{(k)}(r,z)\\ \sigma_{\ell,\alpha}^{(k)}(r,z)\end{pmatrix}\right\}_{\ell=1}^\infty \subset L^2(\Sigma)\times L^2(\Sigma)$$

be the eigenvalues arranged in increasing order (with counting multiplicity) and a complete system of the corresponding orthonormal eigenfunctions of (6.7). We can apply an argument completely similar to that in Lemma 14 and obtain the following asymptotic behaviors of the eigenvalues and eigenfunctions.

LEMMA 15. *For each nonnegative integer k,*

$$\text{(6.8)}\qquad \lim_{\alpha\to\infty}\mu_\ell^{(k)}(\alpha) = \mu_\ell^{(k)} \quad (k\ge 0,\ \ell\ge 1),$$

$$\text{(6.9)}\qquad \lim_{\alpha\to\infty}\left(\|\nabla\tau_{\ell,\alpha}^{(k)}\|_{L^2(\Sigma)}^2 + \alpha\|\tau_{\ell,\alpha}^{(k)}\|_{L^2(\Sigma)}^2\right) = 0 \quad (k\ge 0,\ \ell\ge 1),$$

where $\{\mu_\ell^{(k)}\}_{\ell=1}^\infty$ is the set of the eigenvalues arranged in increasing order (with counting multiplicity) of the following eigenvalue problem:

$$\text{(6.10)}\qquad \begin{cases} L_1\sigma - \dfrac{k^2}{r^2}\sigma + \mu\sigma = 0 \quad \textit{in} \quad \Sigma, \\ \dfrac{\partial\sigma}{\partial\mathbf{n}} = 0 \quad \textit{on} \quad \partial\Sigma. \end{cases}$$

LEMMA 16. *The family of functions*

$$\text{(6.11)}\quad \frac{1}{\sqrt{2\pi}}\begin{pmatrix}\tau_{\ell,\alpha}^{(0)}(r,z)\\ \sigma_{\ell,\alpha}^{(0)}(r,z)\end{pmatrix}, \quad \frac{1}{\sqrt{\pi}}\begin{pmatrix}\tau_{\ell,\alpha}^{(k)}(r,z)\cos k\theta\\ \sigma_{\ell,\alpha}^{(k)}(r,z)\sin k\theta\end{pmatrix}, \quad \frac{1}{\sqrt{\pi}}\begin{pmatrix}\tau_{\ell,\alpha}^{(k)}(r,z)\sin k\theta\\ -\sigma_{\ell,\alpha}^{(k)}(r,z)\cos k\theta\end{pmatrix}$$

($k\ge 1, \ell\ge 1$) form a complete orthonormal basis in $L^2(\Omega)\times L^2(\Omega)$.

Now we can expand any $\left(\begin{smallmatrix}\widehat{\phi}\\ \widehat{\psi}\end{smallmatrix}\right) \in L^2(\Omega)\times L^2(\Omega)$ in terms of the above basis:

$$\text{(6.12)}\qquad \begin{pmatrix}\widehat{\phi}\\ \widehat{\psi}\end{pmatrix} = \frac{1}{\sqrt{\pi}}\sum_{k\ge 1,\ell\ge 1}\left(c_{k,\ell}\begin{pmatrix}\tau_{\ell,\alpha}^{(k)}\cos k\theta\\ \sigma_{\ell,\alpha}^{(k)}\sin k\theta\end{pmatrix} + d_{k,\ell}\begin{pmatrix}\tau_{\ell,\alpha}^{(k)}\sin k\theta\\ -\sigma_{\ell,\alpha}^{(k)}\cos k\theta\end{pmatrix}\right)$$

$$+\frac{1}{\sqrt{2\pi}}\sum_{\ell\geq 1} g_\ell \begin{pmatrix} \tau^{(0)}_{\ell,\alpha} \\ \sigma^{(0)}_{\ell,\alpha} \end{pmatrix}, \quad g_\ell,\ c_{k,\ell},\ d_{k,\ell} \in \mathbb{R}.$$

Here g_ℓ, $c_{k,\ell}$, and $d_{k,\ell}$ are determined from $\widehat{\phi}$ and $\widehat{\psi}$ by

$$(6.13)\qquad \begin{cases} g_\ell = \dfrac{1}{\sqrt{2\pi}}\left((\widehat{\phi}\cdot\tau^{(0)}_{\ell,\alpha})_{L^2(\Omega)} + (\widehat{\psi}\cdot\sigma^{(0)}_{\ell,\alpha})_{L^2(\Omega)}\right), \\ c_{k,\ell} = \dfrac{1}{\sqrt{\pi}}\left((\widehat{\phi}\cdot\tau^{(k)}_{\ell,\alpha}\cos k\theta)_{L^2(\Omega)} + (\widehat{\psi}\cdot\sigma^{(k)}_{\ell,\alpha}\sin k\theta)_{L^2(\Omega)}\right), \\ d_{k,\ell} = \dfrac{1}{\sqrt{\pi}}\left((\widehat{\phi}\cdot\tau^{(k)}_{\ell,\alpha}\sin k\theta)_{L^2(\Omega)} - (\widehat{\psi}\cdot\sigma^{(k)}_{\ell,\alpha}\cos k\theta)_{L^2(\Omega)}\right), \\ \|\widehat{\phi}\|^2_{L^2(\Omega)} + \|\widehat{\psi}\|^2_{L^2(\Omega)} = \displaystyle\sum_{\ell=1}^{\infty} g_\ell^2 + \sum_{k=1}^{\infty}\sum_{\ell=1}^{\infty}(c^2_{k,\ell} + d^2_{k,\ell}). \end{cases}$$

We prepare an auxiliary result concerning a complete orthonormal basis of a product of Hilbert spaces, which we use in the proof of Lemma 19 below.

LEMMA 17. *Let H_1 and H_2 be two real Hilbert spaces with inner products $(\cdot,\cdot)_{H_1}$ and $(\cdot,\cdot)_{H_2}$, respectively, and let H be the product Hilbert space $H_1 \times H_2$ with the following inner product:*

$$\left(\begin{pmatrix}\phi\\ \psi\end{pmatrix}, \begin{pmatrix}\phi'\\ \psi'\end{pmatrix}\right)_H \equiv (\phi,\phi')_{H_1} + (\psi,\psi')_{H_2} \quad \textit{for} \quad \begin{pmatrix}\phi\\ \psi\end{pmatrix}, \begin{pmatrix}\phi'\\ \psi'\end{pmatrix} \in H.$$

If there exists an orthonormal basis $\{\binom{\phi_n}{\psi_n}\}_{n=1}^{\infty} \subset H$, then

$$(6.14)\qquad \begin{cases} \displaystyle\sum_{n=1}^{\infty}(\phi,\phi_n)_{H_1}(\psi,\psi_n)_{H_2} = 0 \quad \textit{for any} \quad \begin{pmatrix}\phi\\ \psi\end{pmatrix} \in H, \\ \|\phi\|^2_{H_1} = \displaystyle\sum_{n=1}^{\infty}(\phi,\phi_n)^2_{H_1}, \quad \|\psi\|^2_{H_2} = \sum_{n=1}^{\infty}(\psi,\psi_n)^2_{H_2}. \end{cases}$$

Proof. Take any $\binom{\phi}{\psi} \in H$ and expand with respect to the given orthonormal basis. Then

$$(6.15)\qquad \begin{pmatrix}\phi\\ \psi\end{pmatrix} = \sum_{n=1}^{\infty}((\phi,\phi_n)_{H_1} + (\psi,\psi_n)_{H_2})\begin{pmatrix}\phi_n\\ \psi_n\end{pmatrix},$$

$$(6.16)\qquad \left(\begin{pmatrix}\phi\\ \psi\end{pmatrix}, \begin{pmatrix}\phi\\ \psi\end{pmatrix}\right)_H = \|\phi\|^2_{H_1} + \|\psi\|^2_{H_2} = \sum_{n=1}^{\infty}((\phi,\phi_n)_{H_1} + (\psi,\psi_n)_{H_2})^2.$$

Taking $\phi = 0 \in H_1$ or $\psi = 0 \in H_2$, we get the second and third equalities in (6.14). The first equality in (6.14) follows immediately from (6.16). □

The following lemma directly follows from the above lemmas.

LEMMA 18. *For any $(\widehat{\phi},\widehat{\psi}) \in L^2(\Omega) \times L^2(\Omega)$, the following equalities hold:*

$$\sum_{\ell=1}^{\infty}(\widehat{\phi}\cdot\tau^{(0)}_{\ell,\alpha})_{L^2(\Omega)}(\widehat{\psi}\cdot\sigma^{(0)}_{\ell,\alpha})_{L^2(\Omega)} + 2\sum_{k,\ell\geq 1}(\widehat{\phi}\cdot\tau^{(k)}_{\ell,\alpha}\cos k\theta)_{L^2(\Omega)}(\widehat{\psi}\cdot\sigma^{(k)}_{\ell,\alpha}\sin k\theta)_{L^2(\Omega)}$$
$$- 2\sum_{k,\ell\geq 1}(\widehat{\phi}\cdot\tau^{(k)}_{\ell,\alpha}\sin k\theta)_{L^2(\Omega)}(\widehat{\psi}\cdot\sigma^{(k)}_{\ell,\alpha}\cos k\theta)_{L^2(\Omega)} = 0,$$

$$\begin{aligned}&\|\widehat{\phi}\|^2_{L^2(\Omega)}\\&=\frac{1}{2\pi}\left(\sum_{\ell=1}^{\infty}(\widehat{\phi}\cdot\tau^{(0)}_{\ell,\alpha})^2_{L^2(\Omega)}+2\sum_{\ell,k\geq 1}(\widehat{\phi}\cdot\tau^{(k)}_{\ell,\alpha}\cos k\theta)^2_{L^2(\Omega)}+2\sum_{\ell,k\geq 1}(\widehat{\phi}\cdot\tau^{(k)}_{\ell,\alpha}\sin k\theta)^2_{L^2(\Omega)}\right),\end{aligned}$$

$$\begin{aligned}&\|\widehat{\psi}\|^2_{L^2(\Omega)}\\&=\frac{1}{2\pi}\left(\sum_{\ell=1}^{\infty}(\widehat{\psi}\cdot\sigma^{(0)}_{\ell,\alpha})^2_{L^2(\Omega)}+2\sum_{\ell,k\geq 1}(\widehat{\psi}\cdot\sigma^{(k)}_{\ell,\alpha}\sin k\theta)^2_{L^2(\Omega)}+2\sum_{\ell,k\geq 1}(\widehat{\psi}\cdot\sigma^{(k)}_{\ell,\alpha}\cos k\theta)^2_{L^2(\Omega)}\right).\end{aligned}$$

Proof. Set $H_1 = H_2 = L^2(\Omega)$, $(\cdot,\cdot)_{H_1} = (\cdot,\cdot)_{H_2} = (\cdot,\cdot)_{L^2(\Omega)}$, and $(\widehat{\phi}\cdot\widehat{\psi})_{L^2(\Omega)} = \int_\Omega \widehat{\phi}(r,\theta,z)\widehat{\psi}(r,\theta,z)rdrdzd\theta$ for $\widehat{\phi},\widehat{\psi}\in L^2(\Omega)$. Combining Lemmas 16 and 17 concludes the proof. □

$I_1(\widehat{\phi},\widehat{\psi})$ is expressed in terms of the Fourier coefficients of $\widehat{\phi}$ and $\widehat{\psi}$:

$$I_1(\widehat{\phi},\widehat{\psi}) = \sum_{\ell=1}^{\infty}\mu^{(0)}_\ell(\alpha)g_\ell^2 + \sum_{k=1}^{\infty}\sum_{\ell=1}^{\infty}\mu^{(k)}_\ell(\alpha)(c^2_{k,\ell}+d^2_{k,\ell}). \tag{6.17}$$

We remark that $\tau^{(0)}_{1,\alpha}(r,z) = 0$, $\sigma^{(0)}_{1,\alpha}(r,z) = e_\alpha W_\alpha(r,z)$, $\mu^{(0)}_1(\alpha) = \mu^{(0)}_1 = 0$, $\mu^{(0)}_2 > 0$, $e_\alpha \neq 0$, is a certain real number which satisfies $\lim_{\alpha\to\infty} e_\alpha^2 = 1/|\Omega|$.

We have the following coercive inequality.

LEMMA 19. *For any $c > 0$ and $\eta > 0$, there exist constants $\alpha_1 > 0$ and $c' > 0$ such that*

$$I_1(\widehat{\phi},\widehat{\psi}) \geq c\|\widehat{\phi}\|^2_{L^2(\Omega)} + \left(\min(\mu^{(0)}_2(\alpha),\mu^{(1)}_1(\alpha)) - \eta\right)\|\widehat{\psi}\|^2_{L^2(\Omega)} - c'(\widehat{\psi}\cdot W_\alpha)^2_{L^2(\Omega)} \tag{6.18}$$

for any $\widehat{\phi},\widehat{\psi}\in H^1(\Omega)$ and $\alpha\geq\alpha_1$.

Proof. In view of the eigenvalues of (6.10) and Lemma 15, for a given $c > 0$, we can take a natural number N so that $\mu^{(k)}_\ell(\alpha) \geq c+1$ for $k+\ell > N$, $k\geq 0$, $\ell\geq 1$, and for any large $\alpha > 0$. Thus we have,

$$\begin{aligned}I_1(\widehat{\phi},\widehat{\psi}) \geq &\sum_{\ell=1}^{N}\mu^{(0)}_\ell(\alpha)g^2_\ell + \sum_{k+\ell\leq N}\mu^{(k)}_\ell(\alpha)(c^2_{k,\ell}+d^2_{k,\ell})\\&+(c+1)\left(\sum_{\ell>N}g^2_\ell + \sum_{k\geq 1,\ell\geq 1,k+\ell>N}(c^2_{k,\ell}+d^2_{k,\ell})\right).\end{aligned}$$

Substituting (6.13), we have

$$\begin{aligned}2\pi I_1 \geq &\sum_{\ell=1}^{N}\mu^{(0)}_\ell(\alpha)(\widehat{\phi}\tau^{(0)}_{\ell,\alpha})^2_{L^2(\Omega)}\\&+2\sum_{k+\ell\leq N}\mu^{(k)}_\ell(\alpha)\left((\widehat{\phi}\tau^{(k)}_{\ell,\alpha}\cos k\theta)^2_{L^2(\Omega)}+(\widehat{\phi}\tau^{(k)}_{\ell,\alpha}\sin k\theta)^2_{L^2(\Omega)}\right)\\&+(c+1)\left\{\sum_{\ell>N}(\widehat{\phi}\tau^{(0)}_{\ell,\alpha})^2_{L^2(\Omega)}+2\sum_{k+\ell>N}\left((\widehat{\phi}\tau^{(k)}_{\ell,\alpha}\cos k\theta)^2_{L^2(\Omega)}+(\widehat{\phi}\tau^{(k)}_{\ell,\alpha}\sin k\theta)^2_{L^2(\Omega)}\right)\right\}\\&+2\sum_{\ell=1}^{N}\mu^{(0)}_\ell(\alpha)(\widehat{\phi}\tau^{(0)}_{\ell,\alpha})_{L^2(\Omega)}(\widehat{\psi}\sigma^{(0)}_{\ell,\alpha})_{L^2(\Omega)}\\&+4\sum_{k+\ell\leq N}\mu^{(k)}_\ell(\alpha)(\widehat{\phi}\tau^{(k)}_{\ell,\alpha}\cos k\theta)_{L^2(\Omega)}(\widehat{\psi}\sigma^{(k)}_{\ell,\alpha}\sin k\theta)_{L^2(\Omega)}\end{aligned}$$

$$
\begin{aligned}
&-4\sum_{k+\ell\le N}\mu_\ell^{(k)}(\alpha)(\widehat{\phi\tau}^{(k)}_{\ell,\alpha}\sin k\theta)_{L^2(\Omega)}(\widehat{\psi\sigma}^{(k)}_{\ell,\alpha}\cos k\theta)_{L^2(\Omega)}\\
&+2(c+1)\sum_{\ell>N}(\widehat{\phi\tau}^{(0)}_{\ell,\alpha})_{L^2(\Omega)}(\widehat{\psi\sigma}^{(0)}_{\ell,\alpha})_{L^2(\Omega)}\\
&+4(c+1)\sum_{k+\ell>N}\Big((\widehat{\phi\tau}^{(k)}_{\ell,\alpha}\cos k\theta)_{L^2(\Omega)}(\widehat{\psi\sigma}^{(k)}_{\ell,\alpha}\sin k\theta)_{L^2(\Omega)}\\
&\qquad\qquad\qquad\qquad-(\widehat{\phi\tau}^{(k)}_{\ell,\alpha}\sin k\theta)_{L^2(\Omega)}(\widehat{\psi\sigma}^{(k)}_{\ell,\alpha}\cos k\theta)_{L^2(\Omega)}\Big)\\
&+\sum_{\ell=1}^{N}\mu_\ell^{(0)}(\alpha)(\widehat{\psi\sigma}^{(0)}_{\ell,\alpha})^2_{L^2(\Omega)}\\
&+2\sum_{k+\ell\le N}\mu_\ell^{(k)}(\alpha)\Big((\widehat{\psi\sigma}^{(k)}_{\ell,\alpha}\sin k\theta)^2_{L^2(\Omega)}+(\widehat{\psi\sigma}^{(k)}_{\ell,\alpha}\cos k\theta)^2_{L^2(\Omega)}\Big)\\
&+(c+1)\Big\{\sum_{\ell>N}(\widehat{\psi\sigma}^{(0)}_{\ell,\alpha})^2_{L^2(\Omega)}+2\sum_{k+\ell>N}\Big((\widehat{\psi\sigma}^{(k)}_{\ell,\alpha}\sin k\theta)^2_{L^2(\Omega)}+(\widehat{\psi\sigma}^{(k)}_{\ell,\alpha}\cos k\theta)^2_{L^2(\Omega)}\Big)\Big\}.
\end{aligned}
$$

From Lemma 18,

$$
\begin{aligned}
(6.19)\qquad I_1(\widehat{\phi},\widehat{\psi})\ge{}&\frac{c+1}{2\pi}\|\widehat{\phi}\|^2_{L^2(\Omega)}+\frac{1}{2\pi}\sum_{\ell=1}^{N}(\mu_\ell^{(0)}(\alpha)-c-1)(\widehat{\phi\tau}^{(0)}_{\ell,\alpha})^2_{L^2(\Omega)}\\
&+\frac{1}{\pi}\sum_{k+\ell\le N}(\mu_\ell^{(k)}(\alpha)-c-1)\Big((\widehat{\phi\tau}^{(k)}_{\ell,\alpha}\cos k\theta)^2_{L^2(\Omega)}+(\widehat{\phi\tau}^{(k)}_{\ell,\alpha}\sin k\theta)^2_{L^2(\Omega)}\Big)\\
&+\frac{1}{2\pi}\sum_{\ell=1}^{N}2(\mu_\ell^{(0)}(\alpha)-c-1)(\widehat{\phi\tau}^{(0)}_{\ell,\alpha})_{L^2(\Omega)}(\widehat{\psi\sigma}^{(0)}_{\ell,\alpha})_{L^2(\Omega)}\\
&+\frac{1}{2\pi}\sum_{k+\ell\le N}4(\mu_\ell^{(k)}(\alpha)-c-1)(\widehat{\phi\tau}^{(k)}_{\ell,\alpha}\cos k\theta)_{L^2(\Omega)}(\widehat{\psi\sigma}^{(k)}_{\ell,\alpha}\sin k\theta)_{L^2(\Omega)}\\
&+\frac{1}{2\pi}\sum_{k+\ell\le N}(-4)(\mu_\ell^{(k)}(\alpha)-c-1)(\widehat{\phi\tau}^{(k)}_{\ell,\alpha}\sin k\theta)_{L^2(\Omega)}(\widehat{\psi\sigma}^{(k)}_{\ell,\alpha}\cos k\theta)_{L^2(\Omega)}\\
&+\min(\mu_2^{(0)}(\alpha),\mu_1^{(1)}(\alpha))\left(\|\widehat{\psi}\|^2_{L^2(\Omega)}-\frac{1}{2\pi}(\widehat{\psi\sigma}^{(0)}_{1,\alpha})^2_{L^2(\Omega)}\right).
\end{aligned}
$$

We used $\mu_1^{(0)}(\alpha)=0$, $\tau^{(0)}_{1,\alpha}=0$, $\sigma^{(0)}_{1,\alpha}=e_\alpha W_\alpha$. From Lemma 15, equation (6.9), to the right-hand side of (6.19), the terms that include $\tau^{(k)}_{\ell,\alpha}$ can be absorbed in those that include $\|\widehat{\phi}\|^2_{L^2(\Omega)}$ and $\|\widehat{\psi}\|^2_{L^2(\Omega)}$ for large $\alpha>0$. We have

$$
I_1(\widehat{\phi},\widehat{\psi})\ge c\|\widehat{\phi}\|^2_{L^2(\Omega)}+(\min(\mu_2^{(0)}(\alpha),\mu_1^{(1)}(\alpha))-\eta)\|\widehat{\psi}\|^2_{L^2(\Omega)}-c'(\widehat{\psi}\cdot W_\alpha)^2_{L^2(\Omega)}
$$

for large $\alpha>0$. We obtain (6.18). □

LEMMA 20. *For $B\in L^2_{\mathrm{loc}}(\mathbb{R}^3;\mathbb{R}^3)$ such that $\nabla B\in L^2(\mathbb{R}^3;\mathbb{R}^{3\times3})$,*

$$
\|\nabla B\|^2_{L^2(\mathbb{R}^3;\mathbb{R}^{3\times3})}=\|\operatorname{div}B\|^2_{L^2(\mathbb{R}^3)}+\|\operatorname{rot}B\|^2_{L^2(\mathbb{R}^3;\mathbb{R}^3)}.
$$

Proof. This equality is proved by the Fourier transform. □

Now we estimate $\mathcal{L}(\widehat{\phi},\widehat{\psi},B)$ from below.

Proof of Theorem 6. Assume that $(\phi,\psi,B)\in\overline{N}(u_\alpha,v_\alpha,A_\alpha)$.

$$
\mathcal{L}(\phi,\psi,B)=I_1(\widehat{\phi},\widehat{\psi})+I_2(B)+I_3(\widehat{\phi},\widehat{\psi},B).
$$

We prove that $|I_3(\widehat{\phi},\widehat{\psi},B)|$ is dominated by I_1 and I_2.

$$\begin{aligned}|I_3(\widehat{\phi},\widehat{\psi},B)| &\le \left|\int_\Omega \frac{(Y_\alpha - m)S_1 W_\alpha \widehat{\phi}}{r^2} dx\right| + \left|\int_\Omega \left(S_2 \frac{\partial W}{\partial r}\widehat{\psi} + S_3 \frac{\partial W}{\partial z}\widehat{\psi}\right) dx\right| \\ &\le \int_\Omega \frac{|mS_1\widehat{\phi}|}{r^2} dx + \sup_\Omega |\nabla W_\alpha| \int_\Omega |\widehat{\psi}|(|S_2| + |S_3|)dx \\ &\le \frac{|m|}{r_0}\left(\epsilon \int_\Omega \frac{S_1^2}{r^2}dx + \frac{1}{4\epsilon}\int_\Omega \widehat{\phi}^2 dx\right) + \sup_\Omega |\nabla W_\alpha| \int_\Omega \left(\widehat{\psi}^2 + \frac{S_2^2 + S_3^2}{2}\right)dx \\ &= \sup_\Omega |\nabla W_\alpha| \cdot \|\widehat{\psi}\|^2_{L^2(\Omega)} + \frac{|m|}{4\epsilon r_0}\|\widehat{\phi}\|^2_{L^2(\Omega)} \\ &\quad + \frac{|m|\epsilon}{r_0}\int_\Omega \frac{S_1^2}{r^2}dx + \sup_\Omega |\nabla W_\alpha| \int_\Omega \frac{S_2^2 + S_3^2}{2} dx.\end{aligned}$$

From $|B|^2 = S_1^2/r^2 + S_2^2 + S_3^2$ and Proposition 9, take $\epsilon > 0$ so that $\epsilon h|m|/r_0 = 1/2$. Next, take c in Lemma 19 such that $c = |m|/(4\epsilon r_0) + 1$. From Lemma 19 and Proposition 9, we can take α_1 large so that the following inequality is true for $\delta = \min(\mu_2^{(0)}/2, \mu_1^{(1)}/2, 1) > 0$:

$$\begin{aligned}(6.20)\qquad \mathcal{L}(\phi,\psi,B) &\ge \delta\left(\|\widehat{\phi}\|^2_{L^2(\Omega)} + \|\widehat{\psi}\|^2_{L^2(\Omega)} + \|B\|^2_{L^2(\Omega;\mathbb{R}^3)} + \|\mathrm{rot}\, B\|^2_{L^2(\mathbb{R}^3;\mathbb{R}^3)}\right) \\ &= \delta\left(\|\phi\|^2_{L^2(\Omega)} + \|\psi\|^2_{L^2(\Omega)} + \|B\|^2_{L^2(\Omega;\mathbb{R}^3)} + \|\mathrm{rot}\, B\|^2_{L^2(\mathbb{R}^3;\mathbb{R}^3)}\right) \quad \text{for} \quad \alpha \ge \alpha_1.\end{aligned}$$

We used $(\widehat{\psi} \cdot W_\alpha)_{L^2(\Omega)} = 0$ for $(\phi,\psi,B) \in \overline{N}(u_\alpha, v_\alpha, A_\alpha)$. We will get a similar inequality on $N(u_\alpha, v_\alpha, A_\alpha)$. First, we recall the following inequality:

$$\int_{\mathbb{R}^3} \frac{\varphi(x)^2}{|x-y|^2} dx \le 4\int_{\mathbb{R}^3} |\nabla\varphi(x)|^2 dx \quad (\forall y \in \mathbb{R}^3, \forall \varphi \in H^1(\mathbb{R}^3)) \quad \text{(cf. [16])}.$$

Since $\Omega \subset \mathbb{R}^3$ is a bounded domain, by fixing y outside of Ω, we see that there exists a constant $R_1 > 0$ such that

$$(6.21)\qquad R_1 \int_\Omega \varphi^2 dx \le \int_{\mathbb{R}^3} |\nabla\varphi|^2 dx \quad (\forall \varphi \in H^1(\mathbb{R}^3)).$$

From Proposition 9, equation (4.4), there exist constants $R_2 > 0$ and $\alpha_2 > 0$ such that for any $\alpha > \alpha_2$,

$$(6.22)\qquad R_2 \int_\Omega \varphi^2 dx \le \int_\Omega |\nabla\varphi|^2 dx \quad \left(\forall \varphi \in H^1(\Omega);\ \int_\Omega (u_\alpha^2 + v_\alpha^2)\,\varphi dx = 0\right).$$

It is also true that there exists a $R_3 > 0$ such that

$$(6.23)\qquad \int_{\partial\Omega} \varphi^2 dS \le R_3 \int_\Omega (|\varphi|^2 + |\nabla\varphi|^2) dx \quad (\forall \varphi \in H^1(\Omega)\,).$$

Take any $(\phi,\psi,B) \in N(u_\alpha, v_\alpha, A_\alpha)$ and we have

$$(\phi,\psi,B) = (-v\xi, u\xi, \nabla\xi) + (\overline{\phi}, \overline{\psi}, \overline{B}) \in T(u,v,A) + \overline{N}(u,v,A),$$

which is, equivalently, $\phi = -v\xi + \overline{\phi}$, $\psi = u\xi + \overline{\psi}$, $B = \nabla\xi + \overline{B}$, and

$$(6.24)\qquad \int_\Omega (u_\alpha^2 + v_\alpha^2)\,\xi(x)dx = 0, \quad \Delta\xi = 0 \text{ in } \Omega, \quad \frac{\partial\xi}{\partial\nu} = \langle B \cdot \nu\rangle \text{ on } \partial\Omega.$$

From these equations, we see that

$$(6.25)\qquad \mathrm{rot}\, B = \mathrm{rot}\, \overline{B} \quad \text{in} \quad \mathbb{R}^3,$$

$$\phi^2+\psi^2=(-v\xi+\overline{\phi})^2+(u\xi+\overline{\psi})^2\leq 2(\overline{\phi}^2+\overline{\psi}^2)+2\xi^2 \quad \text{in} \quad \Omega. \tag{6.26}$$

By (6.22) and (6.24), we obtain

$$R_2\int_\Omega \xi^2 dx \leq \int_\Omega |\nabla\xi|^2 dx. \tag{6.27}$$

On the other hand, (6.24) yields

$$0=\int_\Omega \xi\Delta\xi dx=\int_{\partial\Omega}\xi\frac{\partial\xi}{\partial\nu}dS-\int_\Omega|\nabla\xi|^2dx$$

and, subsequently,

$$\begin{aligned}\int_\Omega|\nabla\xi|^2dx&=\int_{\partial\Omega}\xi\langle B\cdot\nu\rangle dS\leq\frac{\epsilon}{2}\int_{\partial\Omega}\xi^2dS+\frac{1}{2\epsilon}\int_{\partial\Omega}|B|^2dS\\&\leq\frac{\epsilon R_3}{2}\int_\Omega(\xi^2+|\nabla\xi|^2)dx+\frac{R_3}{2\epsilon}\int_\Omega(|B|^2+|\nabla B|^2)dx.\end{aligned} \tag{6.28}$$

Combining (6.27) and (6.28) and taking $\epsilon=R_2/R_3(R_2+1)$, we have

$$\int_\Omega|\nabla\xi|^2dx\leq\frac{R_3^2(1+1/R_2)}{2h^2}\int_\Omega\left(|B|^2+|\nabla B|^2\right)dx.$$

Using (6.26)–(6.29), we conclude that there exists a constant $c>0$, (which is independent of $(\phi,\psi,B)\in N(u_\alpha,v_\alpha,A_\alpha)$) such that

$$\int_\Omega(\phi^2+\psi^2)dx+\int_\Omega|B|^2dx+\int_{\mathbb{R}^3}|\mathrm{rot}B|^2dx\leq c\int_\Omega(\overline{\phi}^2+\overline{\psi}^2)dx+\int_\Omega|\overline{B}|^2dx+\int_{\mathbb{R}^3}|\mathrm{rot}\overline{B}|^2dx$$

On the other hand, from $\operatorname{div}B=0$ in $\mathbb{R}^3$ and Lemma 20, we see that

$$\|\nabla B\|_{L^2(\mathbb{R}^3;\mathbb{R}^{3\times3})}=\|\mathrm{rot}\,B\|_{L^2(\mathbb{R}^3;\mathbb{R}^3)}$$

and $\mathcal{L}(\phi,\psi,B)=\mathcal{L}(\overline{\phi},\overline{\psi},\overline{B})$ (cf. Proposition 3). Hence we obtain the desired inequality (3.4) from (6.20), which completes the proof of Theorem 6. □

REFERENCES

[1] H. Amann, *On the existence of positive solutions of nonlinear elliptic boundary value problems*, Indiana Univ. Math. J., 21 (1971), pp. 125–146.

[2] P. Bauman, N. N. Carlson, and D. Phillips, *On the zeros of solutions to Ginzburg–Landau type systems*, SIAM J. Math. Anal., 24 (1993), pp. 1283–1293.

[3] M. Berger and Y. Chen, *Symmetric vortices for the Ginzburg–Landau equations of superconductivity and the nonlinear desingularization phenomenon*, J. Funct. Anal., 82 (1989), pp. 259–295.

[4] F. Bethuel, H. Brezis, and F. Helein, *Limite singulière pour la minimisation de fonctionnelles du type Ginzburg–Landau*, C. R. Acad. Sci. Paris Sér. I Math., 314 (1992), pp. 891–896.

[5] N. A. Bobylev, *Topological index of extremals of multidimensional variational problems*, Funct. Anal. Appl., 20 (1986), pp. 89–93.

[6] R. W. Carroll and A. J. Glick, *On the Ginzburg–Landau equations*, Arch. Rational Mech. Anal., 16 (1968), pp. 373–384.

[7] Y. Chen, *Nonsymmetric vortices for the Ginzburg–Landau equations on the bounded domain*, J. Math. Phys., 30 (1989), pp. 1942–1950.

[8] Q. Du, M. Gunzberger, and J. Peterson, *Analysis and approximation of the Ginzburg–Landau model of superconductivity*, SIAM Review, 34 (1992), pp. 54–81.

[9] C. Elliot, H. Matano, and Q. Tang, *Zeros of a complex Ginzburg–Landau order parameter with application to superconductivity*, European J. Appl. Math., 5 (1994), pp. 431–448.

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

[11] V. Ginzburg and L. Landau, *On the theory of superconductivity*, Zh. Éksper. Teoret. Fiz., 20 (1950), pp. 1064–1082.

[12] A. Jaffe and C. Taubes, *Vortices and Monopoles*, Birkhäuser, Boston, 1980.

[13] S. Jimbo and Y. Morita, *Stability of non-constant steady state solutions to a Ginzburg–Landau equation in higher space dimensions*, Nonlinear Anal., 22 (1994), pp. 753–770.

[14] V. S. Klimov, *Nontrivial solutions of the Ginzburg–Landau equation*, Theort. Math. and Phys., 50 (1982), pp. 383–389.

[15] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry, Vol.* II, Interscience, New York, 1963.

[16] O. A. Ladyzenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Beach, New York, 1969.

[17] H. Matano and M. Mimura, *Pattern formation in competion diffusion systems in nonconvex domains*, Publ. Res. Inst. Math. Sci., 19 (1983), pp. 1049–1079.

[18] H. Matano, *Existence of nontrivial unstable sets for equilibriums of strongly order preserving systems*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 30 (1984), pp. 645–673.

[19] A. B. Monvel-Berthier, V. Georgescu, and R. Pruce, *A boundary value problem related with the Ginzburg–Landau model*, Comm. Math. Phys., 142 (1991), pp. 1–23.

[20] F. Odeh, *Existence and bifurcation theorems for the Ginzburg–Landau equations*, J. Math. Phys., 8 (1967), pp. 2351–2356.

[21] M. H. Protter and H. F. Weinberger, *Maximum Principle in Differential Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1967.

[22] M. Reed and B. Simon, *Methods of Mathematical Physics* IV, Academic Press, New York, 1978.

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

[24] R. Temam, *Navier–Stokes Equations*, North–Holland, Amsterdam, 1979.

[25] Y. Yang, *Existence, regularity and asymptotic behavior of the solution to the Ginzburg–Landau equations on* $\mathbb{R}^3$, Comm. Math. Phys., 123 (1989), pp. 147–161.

[26] ———, *Boundary value problems of the Ginzburg–Landau equations*, Proc. Roy. Soc. Edinburgh, 114A (1990), pp. 355–365.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1386–1405, September 1996

REGULARITY OF THE GAIN TERM AND STRONG L^1 CONVERGENCE TO EQUILIBRIUM FOR THE RELATIVISTIC BOLTZMANN EQUATION*

HÅKAN ANDRÉASSON†

Abstract. The main purpose of the paper is to show that the gain term of the relativistic collision operator is regularizing. This is a generalization of P. L. Lions' analogous result in the nonrelativistic situation. The regularizing theorem has many applications in kinetic theory, and a few are discussed in this paper. In particular, the asymptotic behaviour of periodic solutions to the relativistic Boltzmann equation is studied. We show that such solutions converge strongly in L^1 to a global Jüttner equilibrium solution (sometimes called a relativistic Maxwellian) provided that the initial data satisfy the physically natural bounds of finite energy and entropy.

Key words. relativistic Boltzmann equation, regularity, Fourier integral operators, stationary phase, strong L^1 convergence, Jüttner solution

AMS subject classifications. 76P05, 83A05, 35S30

1. Introduction.

The relativistic Boltzmann equation models the space–time behaviour of the one-particle distribution function, corresponding to a many-particle system obeying the laws of relativistic mechanics. We refer to [ACB] for applications to different fields such as plasma and nuclear physics, and we mention the books of Synge [Sy], deGroot et al. [GLW], and Stewart [St] for background on the relativistic equation.

In §1, the relativistic Boltzmann equation is defined and the assumptions used throughout the paper are specified. Also, some general facts about the equation are presented. Section 2 is the main part of the paper. Here we show that the gain term of the relativistic collision operator is regularizing. In the classical setting, this result was first obtained by P. L. Lions [L]. Recently, a simpler proof has been given by Wennberg [W] by rewriting the gain term via Carleman's representation and using the fact that it then takes the form of a generalized Radon transform. Lions' approach [L] relies on the method of stationary phase and some facts from the theory of Fourier integral operators, and we will also follow this approach in the relativistic situation. The regularizing property of the gain term is a consequence of the specific nature of the collision geometry. For relativistic interactions, the collision geometry is different from the classical one due to the fact that the collison invariants have a different form. It is well known that the collision geometry for classical interactions is spherical and invariant under translations, i.e., the collision sphere is unchanged as long as the relative velocity remains the same. This is not the case in the relativistic situation, where we get ellipsoids instead of spheres and where the eccentricity of an ellipsoid changes with the energies of the particles which take part in the collision process. Therefore, the translation invariance from the classical situation is not carried over to the relativistic case. The translation invariance is used by Lions to simplify the operator to be studied. A simplification in the relativistic case is even more important since the explicit calculations needed in the proof will become quite involved otherwise. We use the Lorentz invariance of the relativistic particle mechanics to perform this simplification.

In §3, we present two applications of the regularizing theorem. First, we discuss

* Received by the editors September 12, 1994; accepted for publication February 27, 1995.

† Department of Mathematics, Chalmers University of Technology, S-412 96 Göteborg, Sweden.

the problem of the asymptotic behaviour of solutions to the relativistic Boltzmann equation. This problem has recently been studied by Glassey and Strauss in [GlS2] and [GlS3]. We extend one of their results by proving strong L^1 convergence to a global Jüttner equilibrium solution when the initial data, periodic in the space variables, satisfy the natural bounds of finite energy and entropy. The analogous problem in the nonrelativistic situation was first solved by Arkeryd [Ar5] using a nonstandard method. A standard proof was then obtained by Lions [L] by applying the regularizing theorem. Both the nonstandard approach and the standard approach will be discussed.

Next, we apply the regularizing theorem to a functional equation important in relativistic kinetic theory. The result obtained, interesting in its own right, is then used in the appendix to discuss a point of connection between Arkeryd's and Lions' approaches to the asymptotic problem discussed above.

We end this introduction with a presentation of the specific assumptions used in this paper for the relativistic Boltzmann equation. Let the speed of light be normalized to $c = 1$ and the particle rest mass to $m = 1$, and let $(+ - - -)$ be the signature. The relativistic Boltzmann equation models the space–time behaviour of the one-particle distribution function, $f = f(x,p,t)$. The equation has the form

$$\left(\partial_t + \frac{p}{p_0}\cdot\nabla_x\right) f = Q(f,f), \tag{1.1}$$

where the collision operator is defined by

$$Q(f,g)(p) = \frac{1}{p_0}\int_{\mathbb{R}^3}\int_{\mathbb{S}^2}(f(p')g(q') - f(p)g(q))B(g,\theta)d\Omega\frac{dq}{q_0}. \tag{1.2}$$

Here $d\Omega$ is the element of surface area on $\mathbb{S}^2$, $p^\mu = (p_0, p)$ is the four-momentum ($p_0 \in \mathbb{R}$, $p \in \mathbb{R}^3$, $\mu = 0,1,2,3$), and $p_0 = \sqrt{1+p^2}$ is the particle energy. The total energy and relative momentum in the center-of-mass system are $s^{1/2} = |q^\mu + p^\mu|$ and $2g = |q^\mu - p^\mu|$, respectively. Primed momenta denote the associated momenta in the scattering process; hence $p^\mu + q^\mu = p^{\mu'} + q^{\mu'}$. The scattering angle θ in the center-of-mass system satisfies

$$\cos\theta = 1 - 2\frac{(p^\mu - q^\mu)(p'_\mu - q'_\mu)}{|p^\mu - q^\mu|^2}. \tag{1.3}$$

The kernel $B(g,\theta)$ and the scattering cross-section $\sigma(g,\theta)$ are related by

$$B(g,\theta) = \frac{gs^{1/2}}{2}\sigma(g,\theta). \tag{1.4}$$

Remark. Introducing the Møller velocity

$$v_M = \frac{gs^{1/2}}{p_0q_0} = \frac{2g\sqrt{1+g^2}}{p_0q_0} \tag{1.5}$$

with the identity $4g^2 = s - 4$, we see that the equation is quite similar to the classical Boltzmann equation. In particular, if we consider the classical limit where $|p|+|q| \ll 1$, we get $2g \approx |p - q|$. Thus the Møller velocity tends to the relative velocity, $v_M \approx 2g \approx |p - q|$, and we have recovered the classical equation.

In order to discuss the trend to equilibrium in §3, we adopt the assumptions on the scattering cross-section used by Dudyński and Ekiel-Jeżewska [DEJ], thereby proving the global existence of equation (1.1), i.e.,

$$B > 0 \text{ a.e.}, \qquad B \in L^1_{\text{loc}}(\mathbb{R}^3 \times \mathbb{S}^2), \tag{1.6}$$

$$\frac{1}{p_0}\int_{\{|q|\le R\}}\int_{\mathbb{S}^2} B(g,\theta)d\Omega\frac{dq}{q_0} \to 0 \quad \text{as } |p| \to \infty \quad \forall R < \infty. \tag{1.7}$$

These assumptions are of the same type as those used by DiPerna and Lions [DPL1] in the nonrelativistic situation but modified in a natural way to the relativistic case.

The collision operator (1.2) is defined in the center of mass system. We could equivalently use a different representation (see Appendix II in [GlS2] for a derivation):

$$\begin{aligned} Q(f,g) = &\int_{\mathbb{R}^3}\int_{\mathbb{S}^2} k(p,q,\omega) \\ &\times[f(p+a(p,q,\omega)\omega)g(q-a(p,q,\omega)\omega) - f(p)g(q)]dpd\omega, \end{aligned} \tag{1.8}$$

where

$$k(p,q,\omega) = 4s\sigma(p_0+q_0)^2\frac{|\omega\cdot(\hat{q}-\hat{p})|}{(e^2-(\omega\cdot(p+q))^2)^2}, \tag{1.9}$$

$$a(p,q,\omega) = \frac{2ep_0q_0(\omega\cdot(\hat{q}-\hat{p}))}{e^2-(\omega\cdot(p+q))^2}. \tag{1.10}$$

Here $e := p_0 + q_0$ is the total energy and $\hat{x} := x/x_0$, so $\hat{q} - \hat{p}$ is the relative velocity. The function a is the distance from p to p' (and from q to q'). The quantities s and $\sigma(g,\theta)$ are defined above. The explicit form of the kernel k is not essential for our purpose, but we want to give a proper definition of the gain term from a kinetic point of view. Furthermore, the transformation property of k under a change of coordinates $(p,q) \to (p',q')$ is important in the study of the relativistic Boltzmann equation, so a short presentation on this topic will be given. For classical interactions, it is well known that the Jacobian of the corresponding transformation is unity and that the cross-section (kernel) is invariant. The behaviour is slightly different in the relativistic situation, but, as is physically necessary, the relation

$$k(p,q,\omega)dpdq = k(p',q',\omega)dp'dq' \tag{1.11}$$

holds, meaning that there is local reversibility. In fact, the Jacobian is given by

$$\frac{\partial(p',q')}{\partial(p,q)} = \frac{p_0'q_0'}{p_0q_0} \tag{1.12}$$

(see [GlS1]), and this in turn implies

$$k(p,q,\omega) = k(p',q',\omega)\frac{\partial(p',q')}{\partial(p,q)} \tag{1.13}$$

(see [GlS2]). Finally, the collision operator can be written in an obvious way as

$$Q(f,g) = Q^+(f,g) - Q^-(f,g),$$

where Q^+ and Q^- are referred to as the gain and loss terms, respectively.

2. A regularity theorem for the gain term. This section generalizes P. L. Lions' result [L] concerning Sobolev estimates of the gain term to the relativistic setting. Some familiarity with [L] will simplify the reading. The estimates obtained in [L] rely on the specific nature of the collision geometry. For relativistic interactions, the collision geometry changes according to the relativistic conservation laws

$$p_0 + q_0 = p_0' + q_0', \tag{2.1}$$

$$p + q = p' + q'. \tag{2.2}$$

For fixed p and q, the possible values of p' and q' will all belong to an ellipsoid instead of a sphere, as is the case for classical interactions. Thus the integration in the gain term is performed over ellipsoids instead of spheres. However, the eccentricity of an ellipsoid depends on the energies of the ingoing particles, whereas the eccentricity is constant ($= 0$) in the classical situation, or, in other words, the translation-invariance is not carried over to the relativistic case. Lions [L] uses this translation invariant property to simplify the operator to be studied. A simplification of the operator is even more important in the relativistic situation since the explicit computations in the proof become quite involved without any reductions. Below we will reduce the complexity of the operator by making use of the Lorentz invariance.

Let us take the dimension arbitrary ($N \geq 2$). The gain term then takes the form

$$Q^+(f,g) = \int_{\mathbb{R}^N} \int_{\mathbb{S}^{N-1}} b(p,q,\omega) f(p + a(p,q,\omega)\omega) g(q - a(p,q,\omega)\omega) dp d\omega. \tag{2.3}$$

Here the notation of the kernel is changed to b instead of k. The kernel b is equipped with specific regularity properties, as will be clear from the formulation of the following main theorem.

THEOREM 1. *Let $b \in C^\infty(\mathbb{R}^N \times \mathbb{R}^N \times \mathbb{S}^{N-1})$, $0 \leq g \in L^1_{\text{loc}}(\mathbb{R}^N)$, $0 \leq f \in L^2_{\text{loc}}(\mathbb{R}^N)$. Assume that b vanishes if $|p|$ is large, if $|q|$ is large, or if $|q-p|$ is small. Also assume that b vanishes if $|(\hat{q} - \hat{p}) \cdot \omega|$ is near 0 or if $|(q-p)\cdot\omega|$ is near $|q-p|$ uniformly in ω. Then*

$$\|Q^+(f,g)\|_{H^{\frac{N-1}{2}}(\mathbb{R}^N)} \leq C\|f\|_{L^2(K_q)}\|g\|_{L^1(K_p)} \tag{2.4}$$

for some $C > 0$ depending only on b. Here K_q and K_p are compact sets in $\mathbb{R}^N$ such that supp $b \subset K_q \times K_p \times \mathbb{S}^{N-1}$.

Remark. The hypotheses in Theorem 1 differ slightly from what one could argue should be the natural extension of [L] to the relativistic case. In [L], b vanishes if the relative velocity is small and large, respectively. It seems natural that a translation of this condition to the relativistic case, i.e., a vanishing of b for small and large relative momentum, should be sufficient. However, uniform control of the eccentricity of the ellipsoids is necessary for carrying out the proof, and that control is not quite achieved with only this hypothesis on b. Indeed, an ellipsoid has its principal axis directed along $p + q$, so the remaining $N - 1$ axes have equal lengths. If we denote by α the length of the principal axis and by β the length of one of the other axes, we have

$$\frac{\beta^2}{\alpha^2} = \frac{e^2 - (p+q)^2}{e^2}. \tag{2.5}$$

Hence the numerator could be fixed as the denominator approaches infinity, so control of the relative momentum does not imply control of the eccentricity.

Let us now consider g as fixed in L^1, and let the linear operator Q_g be given by

$$Q_g(f) = Q^+(f,g). \tag{2.6}$$

We wish to compute the transpose of Q_g in the $L^2 \times L^2$ duality (of real functions). Therefore, for all $\varphi \in C_0^\infty(\mathbb{R}^N)$, consider $\int_{\mathbb{R}^N} Q_g(f)\varphi dp$ or, explicitly,

$$\int_{\mathbb{R}^N} Q_g(f)\varphi dp$$

$$= \iint_{\mathbb{R}^{2N}} dpdq \int_{\mathbb{S}^{N-1}} d\omega b(p,q,\omega) f(q + a(p,q,\omega)\omega) g(p - a(p,q,\omega)\omega)\varphi(p). \tag{2.7}$$

Let us change variables $(p,q) \to (p',q')$ in (2.7), where this map and its inverse (see [GlS2]) are given by

$$\begin{aligned} p' &= p - a(p,q,\omega)\omega, & q' &= q + a(p,q,\omega)\omega, \\ p &= p' + a(p',q',\omega)\omega, & q &= q' - a(p',q',\omega)\omega. \end{aligned} \tag{2.8}$$

We wish to show that the essential properties of the kernel are preserved under this transformation, i.e., $b(p,q,\omega) \to \tilde{b}(p',q',\omega)$ such that $\tilde{b}$ also satisfies the conditions of Theorem 1. We saw above (1.12) that the Jacobian of the map $(p,q) \to (p',q')$ is harmless. Next, we observe from equation (2.8) that a large value of $|p'|$ (or $|q'|$) requires that $|p|$ or $|q|$ is large, i.e., b vanishes, so this property is preserved. And if $|q' - p'|$ is small, then necessarily $|q - p|$ is small and b vanishes, so this property is also preserved. Finally, since we are working in a compact domain, there is uniform control of the eccentricity. This implies that if $|(\widehat{q' - p'}) \cdot \omega|$ is sufficiently small, then $|(\hat{q} - \hat{p}) \cdot \omega|$ is also small, and if $|(q' - p') \cdot \omega|$ is close to $|q' - p'|$, then $|(q - p) \cdot \omega|$ is close to $|q - p|$. Thus all the properties of b are preserved. Since b and $\tilde{b}$ have the same behaviour, we shall use the same notation for the kernel and write b for $\tilde{b}$ below.

Formally, then,

$$\int_{\mathbb{R}^N} Q_g(f)\varphi dp = \int_{\mathbb{R}^N} dq' f(q') \times \left\{ \int_{\mathbb{R}^N} dp' g(p') \int_{\mathbb{S}^{N-1}} d\omega b(p',q',\omega)\varphi(p' + a(p',q',\omega)\omega) \right\}. \tag{2.9}$$

Due to the compact support of the kernel b in the momentum variables, formula (2.9) is trivially justified.

We wish to rewrite the term in brackets in (2.9), considered as an operator in φ. In order to do so, we introduce the Lorentz transformation $\Lambda(p)$ which carries a particle with momentum p to rest. The transformation $\Lambda(p)$, of course, acts on four-vectors (if $N = 3\ldots$), and by the notation $\Lambda(p)x$, where x is a three-vector (if $N = 3\ldots$), i.e., $x \in \mathbb{R}^N$ and $X = (x_0, x) \in \mathbb{R}^{N+1}$, we will intend the projection of $\Lambda(p)X$ onto the N-dimensional momentum space. The notation $x_p := \Lambda(p)x$ will also appear below. Further, if ψ is a function with a momentum variable as an argument, then $(\Lambda(p)\psi)(x) := \psi(\Lambda(p)x)$. The distinction between primed and unprimed variables has no relevance in what follows, so we will drop the primes from here on. The term in brackets in (2.9) can now be reformulated as follows:

$$\int_{\mathbb{R}^N} dp\, g(p) \int_{\mathbb{S}^{N-1}} d\omega b(p,q,\omega)\varphi(p + a(p,q,\omega)\omega)$$

$$= \int_{\mathbb{R}^N} g(p) \left([\Lambda(p) \circ T_p \circ \Lambda(-p)]\, \varphi \right)(q) dp, \tag{2.10}$$

where

$$(2.11) \quad T_p\varphi(q) = \int_{\mathbb{S}^{N-1}} b(p, q_{-p}, \Gamma \circ A(p, q_{-p}, \tilde{\omega}))J(p, q_{-p}, \tilde{\omega})\varphi(a_0(q, \tilde{\omega})\tilde{\omega})d\tilde{\omega}.$$

Before discussing the validity and purpose of this formulation, we define the quantities in (2.11). First, for fixed p and q ($p \neq q$), the map $\Gamma \circ A : \mathbb{S}^{N-1} \to \mathbb{S}^{N-1}$ is defined by

$$(2.12) \qquad A(p, q, \tilde{\omega}) = \Lambda(-p)(a_0(q_p, \tilde{\omega})\tilde{\omega}) - p \quad \text{if } q \cdot \tilde{\omega} > 0.$$

If $q \cdot \tilde{\omega} < 0$, then define $A(p, q, \tilde{\omega}) := -A(p, q, \tilde{\omega})$, where $a_0(q, \omega) := a(0, q, \omega)$. Also, $\Gamma : \mathbb{R}^N \to \mathbb{S}^{N-1}$ is the projection

$$(2.13) \qquad \Gamma(x) = \frac{x}{|x|} \quad \text{if } x \neq 0.$$

$\Gamma \circ A$ is then continuously extended to all of $\mathbb{S}^{N-1}$. Finally, J is the Jacobian of the map $\tilde{\omega} \to \Gamma \circ A(p, q, \tilde{\omega}) = \omega$ for fixed p and q. The map $\Gamma \circ A$ takes care of the aberration occurring under a change of coordinate frames; $\tilde{\omega}$ is the angle of a collision observed from the rest frame of p if the corresponding angle is given by ω in the laboratory frame. Formula (2.11) is now readily verified if one observes that $p + a(p, q, \omega)\omega = \Lambda(-p)(a_0(q_p, \tilde{\omega})\tilde{\omega})$ with $\tilde{\omega} = \Gamma \circ \Lambda(p)(p + a(p, q, \omega)\omega)$.

T_p is the operator used by an observer in the rest frame of a particle with momentum p. The reason for extracting this operator is that the collision geometry becomes less complicated when one of the particles is at rest. The geometry of the interaction process is crucial for the main part of the proof, which relies on the theory of Fourier integral operators and the method of stationary phase. With no reductions of the geometry, the explicit computations for justifying the hypotheses needed in order to realize the proof become quite involved. However, the explicit form of the kernel has become more complicated due to the change of frames. But the vital features of the kernel are maintained, so this will not affect the proof. The mapping $\omega \to \tilde{\omega}$ is, of course, a diffeomorphism since it is just a result of deforming one ellipsoid into another. The Jacobian J of this map only expresses the change of the eccentricity of the collision ellipsoids under changes of coordinate frames. Again, since b has compact support in p and q, there is uniform control of the eccentricity. Thus the Jacobian is bounded in the support of b, and we may still just denote the kernel bJ by b. Hence the operator T_p takes the form (recall that $\omega = \Gamma \circ A(p, q_{-p}, \tilde{\omega})$)

$$(2.14) \qquad T_p\varphi(q) = \int_{\mathbb{S}^{N-1}} b(p, q_{-p}, \omega)\varphi(a_0(q, \tilde{\omega})\tilde{\omega})d\tilde{\omega},$$

where b satisfies the conditions of Theorem 1.

If we can show that T_p is bounded from $H^{-(N-1)/2}(\mathbb{R}^N)$ into $L^2(\mathbb{R}^N)$ (the bound can certainly be taken independently of p since we are working in a compact domain), then we may conclude the proof of Theorem 1. To see this, first apply Hölder's inequality to equation (2.9) and then Jensen's inequality (recall that $g \geq 0$) to equation (2.10). From here, Theorem 1 follows if we observe that the Jacobian of the mapping $q \to \Lambda(p)q$ is harmless because b vanishes for p or q large. Hence we only have to prove the following.

THEOREM 2. *The operator T_p is bounded from $H^{-(N-1)/2}(\mathbb{R}^N)$ into $L^2(\mathbb{R}^N)$ and, more generally, from $H^s(\mathbb{R}^N)$ into $H^{s+(N-1)/2}(\mathbb{R}^N)$, $s \in \mathbb{R}$.*

We fix p in the support of b. An inverse Fourier transform yields, for all $\varphi \in \mathcal{S}(\mathbb{R}^N)$,

$$T_p\varphi(q) = \int_{\mathbb{R}^N}\int_{\mathbb{S}^{N-1}} (2\pi)^{-N} b(p, q_{-p}, \omega) e^{ia_0(q,\tilde{\omega})(\xi\cdot\tilde{\omega})}\hat{\varphi}(\xi)d\tilde{\omega}d\xi. \tag{2.15}$$

Let

$$s(p,q,\xi) = \int_{\mathbb{S}^{N-1}} b(p, q_{-p}, \omega)e^{i[a_0(q,\tilde{\omega})(\xi\cdot\tilde{\omega})-(q\cdot\xi)]}d\tilde{\omega}. \tag{2.16}$$

Clearly, $s(p,\cdot,\cdot) \in C^\infty(\mathbb{R}^{2N})$, and

$$T_p\varphi(q) = \int_{\mathbb{R}^N} (2\pi)^{-N} e^{i(q\cdot\xi)} s(p,q,\xi)\hat{\varphi}(\xi)d\xi. \tag{2.17}$$

We claim that s satisfies, for some $0 < \delta < 1/4$ depending on p,

$$s(p,q,\xi) = 0 \quad \text{if } |q| < \delta \text{ or if } |q_{-p}| > \delta^{-1}, \tag{2.18}$$

$$s(p,q,\xi) = c(p,q,\xi) \quad \text{if } |q\cdot\xi| \geq (1-2\delta)|q||\xi| \text{ and } |\xi| > 1/2, \tag{2.19}$$

and

$$\begin{aligned} s(p,q,\xi) &= e^{i[\phi_+(p,q,\xi)-(q\cdot\xi)]}a_+(p,q,\xi) + e^{i[\phi_-(p,q,\xi)-(q\cdot\xi)]}a_-(p,q,\xi) \\ &\qquad \text{if } |q\cdot\xi| \leq (1-\delta)|q||\xi| \text{ and } |\xi| > 1/2, \end{aligned} \tag{2.20}$$

where

$$\phi_\pm = \frac{(q\cdot\xi)}{2} \pm \frac{|q|\sqrt{\xi^2(e^2-q^2)+(q\cdot\xi)^2}}{e^2-q^2}. \tag{2.21}$$

Here and below, $e = 1 + q_0$ and

$$c(p,\cdot,\cdot) \in S^{-\infty}(\mathbb{R}^{2N}), \tag{2.22}$$

$$a_+(p,\cdot,\cdot), a_-(p,\cdot,\cdot) \in S^{-\frac{N-1}{2}}(\{|q| > \delta\} \times \{|\xi| > 1/2\}), \tag{2.23}$$

where $S^{-\infty} = \cap_{m\in\mathbb{R}} S^m$ and S^m is the usual class of symbols; see, for instance, [H3]. These claims will follow from the stationary-phase method, as will be clear below. First, we show that they are enough to prove Theorem 2. Therefore, assume that (2.18)–(2.20) are valid. Then we can write

$$\begin{aligned} s(p,q,\xi) &= (1-\zeta)(\xi)s(p,q,\xi) \\ &\quad + (1-\theta)\left(\frac{|q\cdot\xi|}{|q||\xi|}\right)\psi(q)\zeta(\xi)s(p,q,\xi) \\ &\quad + \theta\left(\frac{|q\cdot\xi|}{|q||\xi|}\right)\psi(q)\zeta(\xi)\left[e^{i[\phi_+-(q\cdot\xi)]}a_+ + e^{i[\phi_--(q\cdot\xi)]}a_-\right], \end{aligned} \tag{2.24}$$

where $\psi \in C_0^\infty(\mathbb{R}^N)$ with $\psi \equiv 0$ if $|q| \leq \delta/2$, $\psi \equiv 0$ if $|q_{-p}| \geq (\delta/2)^{-1}$, $\psi \equiv 1$ if $|q| \geq \delta$, and $\psi \equiv 1$ if $|q_{-p}| \leq (\delta)^{-1}$. We have also taken $\theta \in C_0^\infty(\mathbb{R})$ with $\theta(t) \equiv 0$ if $|t| \geq 1-\delta$, $\theta(t) \equiv 1$ if $|t| \leq 1-2\delta$, and $\zeta \in C^\infty(\mathbb{R}^N)$ with $\zeta(\xi) \equiv 0$ if $|\xi| \leq 1/2$, $\zeta(\xi) \equiv 1$ if $|\xi| \geq 1$.

The first term obviously belongs to $S^{-\infty}$, and this is also true for the second term because of (2.19) and (2.22). Thus, in view of (2.17), these two terms define bounded operators from $H^s(\mathbb{R}^N)$ into $H^{s+m}(\mathbb{R}^N)$ for all $m \in \mathbb{R}$ (see, for instance, S. Alinhac and P. Gérard [AG]). For the last term, we note that $\theta\psi\zeta a_\pm \in S^{-(N-1)/2}(\mathbb{R}^{2N})$ due to (2.23) and that the phase functions $\phi_\pm$ are positively homogeneous of degree one in ξ. If $\phi_\pm$ satisfy the crucial nondegeneracy condition on the support of $\theta\psi\zeta$, i.e.,

$$\det\left(\frac{\partial^2\phi_\pm}{\partial q_j \partial\xi_k}\right) \neq 0 \quad \text{on supp } \theta\psi\zeta, \tag{2.25}$$

then we could apply the theory of Fourier integral operators (see, e.g., L. Hörmander [H1], [H4] or A. Cordoba and C. Fefferman [CF]) to conclude that the symbols $\theta\psi\zeta e^{i[\phi_\pm-(q-\xi)]}a_\pm$ define bounded operators from H^s into $H^{s+(N-1)/2}$. Obviously, the claims in (2.18)–(2.20) and Lemma 1 below are enough to ascertain Theorem 2. Let us finish this argument by proving the following.

LEMMA 1. *$\phi_\pm$ are nondegenerate on the support of $\theta\psi\zeta$.*

Proof. The proof for ϕ_+ and ϕ_- are analogous, so we present it only for ϕ_+. We want to prove that (2.25) holds on the support of $\theta\psi\zeta$. The calculations become less tedious with a change of coordinates. Introduce the map $q \to t$,

$$t = \frac{q}{\sqrt{e^2 - q^2}}, \quad q \in \mathbb{R}^N, \tag{2.26}$$

which is invertible:

$$q = 2t\sqrt{1+t^2} =: 2tt_0. \tag{2.27}$$

For ϕ_+, we then obtain

$$\phi_+ = (t\cdot\xi)t_0 + |t|\sqrt{\xi^2 + (t\cdot\xi)^2}. \tag{2.28}$$

The Jacobian of the map $q \to t$ is certainly nonzero, so condition (2.25) is equivalent to

$$\det\left(\frac{\partial^2\phi_+}{\partial t_j \partial\xi_k}\right) \neq 0. \tag{2.29}$$

An explicit calculation yields (recall $|t|, |\xi| > 0$ on the support of $\theta\psi\zeta$)

$$\begin{aligned} \frac{\partial^2\phi_+}{\partial t_j\partial\xi_k} &= \delta_{jk}t_0 + \delta_{jk}|t|\frac{(t\cdot\xi)}{E} + t_jt_k\left(\frac{1}{t_0} + \frac{(t\cdot\xi)}{|t|E}\right) \\ &\quad + t_j\xi_k\frac{1}{|t|E} + \xi_jt_k\left(\frac{|t|}{E} - |t|\frac{(t\cdot\xi)^2}{E^3}\right) - \xi_j\xi_k|t|\frac{(t\cdot\xi)}{E^3}, \end{aligned} \tag{2.30}$$

where $E = \sqrt{\xi^2 + (t\cdot\xi)^2}$. With no loss of generality (the determinant is invariant under a change of basis), we may assume that $t_j = \xi_j = 0$ for $j > 2$. We find

$$\det\left(\frac{\partial^2\phi_+}{\partial t_j\partial\xi_k}\right) = \left(t_0 + |t|\frac{(t\cdot\xi)}{E}\right)^{N-2}\cdot Y, \tag{2.31}$$

where

$$\begin{aligned} Y &= t_0^2 + \left(3t_0|t| + \frac{t_0}{|t|}\right)\frac{(t\cdot\xi)}{E} \\ &\quad + (2t_0^2 + t^2)\frac{(t\cdot\xi)^2}{E^2} + t_0|t|\frac{(t\cdot\xi)^3}{E^3}. \end{aligned} \tag{2.32}$$

The first factor in (2.31) is obviously strictly positive, so it is enough to prove that Y is nonzero on the support of $\theta\psi\zeta$. To see this, we observe that Y factorizes

$$Y = t_0|t|\left(\frac{(t\cdot\xi)}{E}+\frac{|t|}{t_0}\right)\left(\frac{(t\cdot\xi)}{E}+\frac{t_0}{|t|}\right)^2. \tag{2.33}$$

A simple calculation now shows that $((t\cdot\xi)/E+|t|/t_0)=0$ if and only if $(t\cdot\xi)=-|t||\xi|$, which means that $(q\cdot\xi)=-|q||\xi|$. Accordingly, $((t\cdot\xi)/E+|t|/t_0)$ is strictly positive on the support of $\theta\psi\zeta$. This holds trivially for the last factor in (2.33), and we are done. □

Remark. If we had extracted the appropriate operator in the center-of-mass system, where the collision geometry is spherical, then the associated phase function would be degenerate. It is obvious that the operator in the center-of-mass system cannot be regularizing due to the fact that the hypersurfaces we integrate over do not move with q. In Lions [L], there is a discussion of the geometrical condition of the hypersurfaces for obtaining regularizing operators.

Let us now complete the proof of Theorem 2 by proving assertions (2.18)–(2.20). Statement (2.18) follows immediately from the nature of the kernel b. The two others will follow from the stationary-phase method. In order to apply this method, we keep the direction $\xi/|\xi|$ fixed and let the modulus $|\xi|$ go to infinity. We want to find the critical points over $\mathbb{S}^{N-1}$ of the C^∞ function

$$\phi = a_0(q,\tilde\omega)(\xi\cdot\tilde\omega). \tag{2.34}$$

Recall that we only have to consider $(q,\xi)\in\mathbb{R}^{2N}$ such that $|q|\geq\delta$, $|q_{-p}|\leq\delta^{-1}$, and $|\xi|>1/2$. If $\tilde\omega_c$ (here and below, $\tilde\omega_c$ denotes an arbitrary critical point) is a critical point of ϕ over $\mathbb{S}^{N-1}$, then the gradient of ϕ at $\tilde\omega_c$ will point in the direction of $\tilde\omega_c$. Hence the critical points are solutions of

$$\frac{\partial\phi}{\partial\tilde\omega}=\left(\frac{\partial\phi}{\partial\tilde\omega}\cdot\tilde\omega\right)\tilde\omega \tag{2.35}$$

or, explicitly,

$$(\xi\cdot\tilde\omega)(e^2+(q\cdot\tilde\omega)^2)q+(q\cdot\tilde\omega)(e^2-(q\cdot\tilde\omega)^2)\xi=2e^2(q\cdot\tilde\omega)(\xi\cdot\tilde\omega)\tilde\omega. \tag{2.36}$$

Next, observe that if $(q\cdot\tilde\omega)/|q|$ is sufficiently small (depending on p), then $((\hat q_{-p}-\hat p)\cdot\omega)/|\hat q_{-p}-\hat p|$ is also small (a collision sufficiently grazing in the rest frame of p is also grazing in the laboratory frame), so $b(p,q_{-p},\omega)$ vanishes. Therefore, we only have to look for critical points such that $(q\cdot\tilde\omega)\neq 0$. Assume first that q and ξ are not collinear. Then the ansatz

$$\tilde\omega_c=\frac{Aq+B\xi}{|Aq+B\xi|} \tag{2.37}$$

gives exactly four critical points $\pm\tilde\omega_+$ and $\pm\tilde\omega_-$. Here

$$\tilde\omega_\pm=\Gamma\left(\left(|q|(q\cdot\xi)\pm e\sqrt{\xi^2(e^2-q^2)+(q\cdot\xi)^2}\right)q\ +\ |q|(e^2-q^2)\xi\right) \tag{2.38}$$

and Γ is the same as above. As the directions of q and ξ become increasingly coincident (we have the parallel case in mind, the antiparallel case being similar), then $\tilde\omega_+$ approaches $q/|q|$, whereas $\tilde\omega_-$ approaches $\tilde\omega^\perp$, which is a unit vector in the plane generated by q and ξ orthogonal to q. Therefore, as $|q\cdot\xi|/|q||\xi|$ gets close to 1, then $|q\cdot\tilde{\omega}_+|/|q|$ and $|q\cdot\tilde{\omega}_-|/|q|$ approach 1 and 0, respectively. As we saw above, b vanishes if $|q\cdot\tilde{\omega}|/|q|$ is sufficiently small. In fact, this also happens if $|q\cdot\tilde{\omega}|$ is sufficiently close to 1 since then $|(q_{-p}-p)\cdot\omega|/|q_{-p}-p|$ is close to 1 and b vanishes according to the hypotheses made on b. Thus, if $|q\cdot\xi|/|q||\xi|$ is close to 1 no critical points are contained in the support of b, and an application of the stationary-phase lemma without critical points (see, for instance, [H2] and [GuS]) yields (2.19). Finally, we prove that our last claim (2.20) holds. Certainly, we only have to work in the domain

$$D=\left\{(q,\xi):|q|\geq\delta,\ |q_{-p}|\leq\delta^{-1},\ |\xi|>1/2,\ |q\cdot\xi|<(1-\delta)|q||\xi|\right\}. \tag{2.39}$$

We will consider only the two critical points $\tilde{\omega}_+$ and $\tilde{\omega}_-$ since the contributions from $-\tilde{\omega}_\pm$ are similar. (If we also assume that b is even in ω which is physically natural, then, since ϕ is also even in ω, the contributions are, in fact, identical.) We begin by computing the functions $\phi_\pm(q,\xi)$ defined by

$$\phi_\pm(q,\xi)=\phi(q,\xi,\tilde{\omega}_\pm). \tag{2.40}$$

For this, we make use of equation (2.36). With $\tilde{\omega}_c=(Aq+B\xi)/|Aq+B\xi|$, where A and B are defined in (2.38), we obtain (recall that q and ξ are not collinear in D)

$$\phi(q,\xi,\tilde{\omega}_c):=\frac{2e(q\cdot\tilde{\omega}_c)(\xi\cdot\tilde{\omega}_c)}{e^2-(q\cdot\tilde{\omega}_c)^2}=\frac{q\cdot(Aq+B\xi)}{eB}. \tag{2.41}$$

Inserting the expressions for A and B, we obtain

$$\phi_\pm(q,\xi)=\frac{e(q\cdot\xi)\pm|q|\sqrt{\xi^2(e^2-q^2)+(q\cdot\xi)^2}}{e^2-q^2}, \tag{2.42}$$

and (2.21) follows if we observe that $e^2-q^2=2(1+q_0)=2e$. Now, in order to apply the stationary-phase method (the case admitting critical points; see, e.g., [H2] or [GuS]), we have to compute the Hessian of ϕ over $\mathbb{S}^{N-1}$ at the critical points. We remark that, in general, the Hessian of a function f at a point p is well defined (independent of the particular local coordinate system chosen) if p is a critical point. Now let q and ξ be fixed. We introduce the functions $\psi_\pm: M_\epsilon\to\mathbb{R}$ defined by

$$\psi_\pm(\omega)=\phi(q,\xi,\omega)-\frac{e^2}{e^2-(q\cdot\tilde{\omega}_\pm)^2}\phi_\pm(q,\xi)|\omega|^2,\qquad\omega\in M_\epsilon, \tag{2.43}$$

where $M_\epsilon=\{\omega\in\mathbb{R}^N:|\omega|<1+\epsilon\}$ and ϵ is taken so small that ϕ is C^∞ on M_ϵ. (Of course, an ϵ that works for all q is possible to choose since $q\in D$.) The Hessians of $\psi_\pm$ and ϕ over $\mathbb{S}^{N-1}$ coincide, of course, since $|\omega|$ is constant on $\mathbb{S}^{N-1}$. Moreover, the factor in front of $|\omega|^2$ is taken in such a way that the gradient of $\psi_\pm$ with respect to $\mathbb{R}^N$ at the points $\tilde{\omega}_\pm$ vanishes. Thus, in view of the remark made above, we can compute the Hessian of $\psi_\pm$ over $\mathbb{S}^{N-1}$ at the critical points by restricting the Hessian of $\psi_\pm$ over $\mathbb{R}^N$ to the tangent plane at the critical points. To do this, choose coordinates (x,x'), where x is a coordinate in the direction of $\tilde{\omega}_+$ (respectively, $\tilde{\omega}_-$) and x' are coordinates in the tangent plane at $\tilde{\omega}_+$ (respectively, $\tilde{\omega}_-$). The derivatives are then taken with respect to x'. To simplify the computations, set

$$\eta=\frac{\sqrt{2e}\,\omega}{\sqrt{e^2-(q\cdot\omega)^2}},\qquad\omega\in M_\epsilon, \tag{2.44}$$

with inverse

$$\omega = \frac{e\eta}{\sqrt{2e + (q\cdot\eta)^2}}. \tag{2.45}$$

Expressed in these coordinates, $\psi_\pm$ take the form

$$\psi_\pm(\eta) = (q\cdot\eta)(\xi\cdot\eta) - H_\pm \frac{|\eta|^2}{2e + (q\cdot\eta)^2}, \qquad \eta \in \mathrm{Im}(M_\epsilon), \tag{2.46}$$

where

$$\eta_\pm = \frac{\sqrt{2e}\,\tilde{\omega}_\pm}{\sqrt{e^2 - (q\cdot\tilde{\omega}_\pm)^2}} \tag{2.47}$$

and

$$H_\pm = \frac{e^3|\eta_\pm|^2\phi_\pm(q,\xi)}{2} = \frac{e^3|\eta_\pm|^2(q\cdot\eta_\pm)(\xi\cdot\eta_\pm)}{2}. \tag{2.48}$$

As usual, to facilitate the reading, we present only the computations for the "+" case since the "−" case is analogous. Let us write $\eta = \eta_+ + \lambda\eta_+^\perp + \nu^\perp$, where $\eta_+^\perp$ is orthogonal to η_+ in the plane generated by q and ξ with $|\eta_+^\perp| = |\eta_+|$ and $\nu^\perp$ is orthogonal to the plane generated by q and ξ. Also, we introduce the notations $a = (q\cdot\eta_+)$, $\tilde{a} = (\xi\cdot\eta_+)$, $b = (q\cdot\eta_+^\perp)$, and $\tilde{b} = (\xi\cdot\eta_+^\perp)$. We then obtain

$$\psi_+ = a\tilde{a} + \lambda(a\tilde{b} + \tilde{a}b) + \lambda^2 b\tilde{b} - H_+\left(\frac{|\eta_+|^2(1+\lambda^2) + |\nu^\perp|^2}{2e + (a+\lambda b)^2}\right). \tag{2.49}$$

Accordingly, we wish to compute

$$\frac{\partial^2\psi_+}{\partial x_i\partial x_j}|_{(0,\ldots,0)} \quad \text{with } x = (\lambda, \nu_1^\perp, \ldots, \nu_{N-2}^\perp). \tag{2.50}$$

We readily obtain

$$\frac{\partial^2\psi_+}{\partial x_i\partial x_j}|_{(0,\ldots,0)} = 0 \quad \text{if } i \neq j. \tag{2.51}$$

Carrying out the calculations, we obtain

$$\psi_{\lambda\lambda} := \frac{\partial^2\psi_+}{\partial\lambda^2}|_{(0,\ldots,0)} = 2b\tilde{b} - \frac{2H_+|\eta_+|^2}{2e + a^2}\left[1 - \frac{b^2}{2e+a^2} + \frac{4a^2b^2}{(2e+a^2)^2}\right], \tag{2.52}$$

$$\frac{\partial^2\psi_+}{\partial\nu_j^{\perp 2}}|_{(0,\ldots,0)} = \frac{-2H_+}{2e + a^2}. \tag{2.53}$$

In order to apply the stationary-phase method, we have to compute the determinant of the Hessian. Certainly, the determinant depends on the local coordinate system chosen. However, we are only interested in the analytical behaviour of the determinant, so we will not be concerned about the square of the Jacobian which appears. This is necessarily positive and uniformly bounded from below and above since q belongs to a compact set. However, it is null homogeneous in ξ because $\tilde{\omega}_+ = \tilde{\omega}_+(q,\xi)$

is null homogeneous in ξ. Furthermore, observing that $\psi_{\lambda\lambda}$ is positively homogeneous of degree 1, we can write

$$\det\left(\frac{\partial^2\psi_+}{\partial x_i \partial x_j}|_{(0,\dots,0)}\right) = h_N\left(q, \frac{\xi}{|\xi|}\right)|\xi|\left(\phi_+(q,\xi)\right)^{N-2}. \tag{2.54}$$

Here we have inserted the expression for H_+—(2.48)—and $h_N \in C^\infty(\mathbb{R}^N \times \mathbb{S}^{N-1})$. Next, we will show that h_N and ϕ_+ are nonzero for all $q \in D$. Recall that

$$\phi_+(q,\xi) = \frac{e(q\cdot\xi) + |q|\sqrt{\xi^2(e^2-q^2)+(q\cdot\xi)^2}}{e^2-q^2}. \tag{2.55}$$

Since $q^2(\xi^2(e^2-q^2)+(q\cdot\xi)^2) - e^2(q\cdot\xi)^2 = (e^2-q^2)(q^2\xi^2-(q\cdot\xi)^2)$, we immediately conclude that $\phi_+ > \epsilon$ for some $\epsilon > 0$ on the domain D. To see that h_N is nonzero, we now only have to show that $\psi_{\lambda\lambda}$ is nonzero or, in fact, that $\psi_{\lambda\lambda} < 0$. A simple observation gives $a \geq 0$, $\tilde{a} \geq 0$, and $b\tilde{b} \leq 0$ for all q and ξ. Indeed, $a = (q\cdot\eta_+) \geq 0$ if and only if $(q\cdot\tilde{\omega}_+) \geq 0$ if and only if

$$q^2\left(|q|(q\cdot\xi) + e\sqrt{\xi^2(e^2-q^2)+(q\cdot\xi)^2}\right) + |q|(e^2-q^2)(q\cdot\xi) \geq 0. \tag{2.56}$$

But the second term is greater than the sum of the other two terms since

$$\begin{aligned} &e^2q^4(\xi^2(e^2-q^2)+(q\cdot\xi)^2) - q^2e^4(q\cdot\xi)^2 \\ &\quad = e^2q^2(e^2-q^2)(q^2\xi^2-(q\cdot\xi)^2). \end{aligned} \tag{2.57}$$

Similarly, $\tilde{a} = (\xi\cdot\eta_+) \geq 0$ if and only if

$$(q\cdot\xi)\left(|q|(q\cdot\xi) + e\sqrt{\xi^2(e^2-q^2)+(q\cdot\xi)^2}\right) + |q|(e^2-q^2)\xi^2 \geq 0. \tag{2.58}$$

However,

$$\begin{aligned} &\left(|q|(\xi^2(e^2-q^2)+(q\cdot\xi)^2)\right)^2 - e^2(q\cdot\xi)^2\left(\xi^2(e^2-q^2)+(q\cdot\xi)^2\right) \\ &\quad = (e^2-q^2)\left(\xi^2(e^2-q^2)+(q\cdot\xi)^2\right)\left(q^2\xi^2-(q\cdot\xi)^2\right), \end{aligned} \tag{2.59}$$

so a and $\tilde{a}$ are both positive and uniformly bounded a way from zero on D. Next, since $\eta_+ = Aq + B\xi$ with A and B both positive, the fact that $b\tilde{b} < 0$ on D is a purely geometrical consequence of the positivity of a and $\tilde{a}$. In view of (2.48), (2.52), and the discussion above, it is enough to show that

$$\frac{1}{(2e+a^2)^2}[(2e+a^2)^2 - b^2(2e+a^2) + 4a^2b^2] > 0 \tag{2.60}$$

in order to prove $\psi_{\lambda\lambda} < 0$. Evidently, the relation $a^2 + b^2 = q^2\eta_+^2$ holds. Using this relation and the expression for η_+ (2.47), we obtain for the numerator in (2.60)

$$\frac{4e^2}{(e^2-q^2s^2)^2}\left(e^4 - e^2q^2 + s^2e^2q^2 + s^2(4q^2-4s^2q^2)\right), \tag{2.61}$$

where $s = ((q/|q|)\cdot\tilde{\omega}_+)$. Clearly, $0 \leq s \leq 1$ so the positivity of (2.60) holds.

The analogous computations for the "$-$" case give

$$|\det|_\pm = \left|h_N\left(q, \frac{\xi}{|\xi|}\right)\right| |\phi_\pm|^{N-2}|\xi|. \tag{2.62}$$

In view of equation (2.52) and (2.53), we easily conclude that the signature of the Hessian is $-(N-1)$ for the "+" case, and we have, in addition (the "$-$"case),

$$\operatorname{sign}_{\pm} = \mp(N-1). \tag{2.63}$$

We may now apply the stationary-phase lemma to conclude that $a_+ \in S^{-(N-1)/2}$. We could immediately refer to [L] for this task but prefer to give the arguments below. In order to see that $a_+ \in S^{-(N-1)/2}$, we introduce a partition of unity on $\mathbb{S}^{N-1}$; $\chi_1, \dots, \chi_5$ with $\chi_i \in C^\infty(\mathbb{S}^{N-1})$, $0 \le \chi_i \le 1$, $\sum_{i=1}^{5} \chi_i = 1$ on $\mathbb{S}^{N-1}$. Here $\chi_i \equiv 1$, $i = 1, \dots, 4$, in a neighbourhood of the four critical points $\tilde{\omega}_+, \tilde{\omega}_-, -\tilde{\omega}_+$, and $-\tilde{\omega}_-$, respectively. Apparently, no critical points are contained in the support of χ_5, so we readily see that the symbol $\theta\psi\zeta \int_{\mathbb{S}^{N-1}} e^{i[\phi-(q\cdot\xi)]} b\chi_5 d\tilde{\omega} \in S^{-\infty}$. Hence we only have to show that

$$\begin{gathered} s_1 := \theta\left(\frac{|q\cdot\xi|}{|q||\xi|}\right)\psi(q)\zeta(\xi)\int_{\mathbb{S}^{N-1}}(2\pi)^{-N}e^{i[\phi-(q-\xi)]}b\chi_1 d\tilde{\omega} = e^{i[\phi_+-(q\cdot\xi)]}\,a_1, \\ \text{where } a_1 \in S^{-\frac{N-1}{2}}. \end{gathered} \tag{2.64}$$

A consequence of the stationary-phase lemma (see Hörmander [H2, p. 222] or Guillemin and Sternberg [GuS, p. 6]) is that there is an asymptotic expansion of s_1 such that (recall that $|\det|_+ = |\xi|^{N-1}|h_N(q,\xi/|\xi|)||\phi_+(q,\xi/|\xi|)|^{N-2}$)

$$\begin{aligned} &\Bigg| s_1 - e^{i[\phi_+-(q\cdot\xi)]}\left\{\theta\left(\frac{|q\cdot\xi|}{|q||\xi|}\right)\zeta(\xi)|\xi|^{-\frac{N-1}{2}}\right\} \\ &\quad\times\left[a^0\left(q,\frac{\xi}{|\xi|}\right)+|\xi|^{-1}a^1\left(q,\frac{\xi}{|\xi|}\right)+\dots+|\xi|^{-N}a^N\left(q,\frac{\xi}{|\xi|}\right)\right]\Bigg| \\ &\quad\le \frac{C}{(1+|\xi|)^{N+1}} \quad \text{on } \mathbb{R}^{2N}. \end{aligned} \tag{2.65}$$

Here

$$\theta\left(\frac{|q\cdot\xi|}{|q||\xi|}\right)\zeta(\xi)|\xi|^{-(\frac{N-1}{2}+j)}\,a^j\left(q,\frac{\xi}{|\xi|}\right) \in S^{-(\frac{N-1}{2}+j)} \quad \text{for all } j \ge 0. \tag{2.66}$$

It is also clear from (2.64) that for each pair of multiindices α and β, the estimate

$$|\partial_q^\beta\partial_\xi^\alpha a_1(q,\xi)| \le C(1+|\xi|)^{|\beta|}, \qquad (q,\xi) \in (\{|q| > \delta\} \times \{|\xi| > 1/2\}), \tag{2.67}$$

holds for some constant C depending on α and β. The estimates (2.65) and (2.67) are, in fact, enough (see Hörmander [H3, p. 67]) to conclude that $a_1 \in S^{-(N-1)/2}(\{|q| > \delta\} \times \{|\xi| > 1/2\})$, and the proof is complete. □

3. Applications of the regularizing theorem. The trend to a global equilibrium solution for the relativistic Boltzmann equation was recently studied by Glassey and Strauss in [GlS2] and [GlS3]. The second of their papers deals with the question of convergence to equilibrium in full physical space for small perturbations of the Jüttner equilibrium solution. We will not treat the full-space situation but concentrate on the periodic case, which was studied in [GlS2]. There the authors prove convergence in a variety of function spaces for initial data periodic in the space variables and *near* equilibrium. Theorem 3 below extends this result by proving strong L^1 convergence to a global Jüttner equilibrium solution for *arbitrary* initial data, periodic in the space variables, and satisfying the natural bounds of finite energy and entropy.

For the classical Boltzmann equation, the analogous result was first shown by Arkeryd [Ar5]. The essential technique in Arkeryd's work is a nonstandard measure-theoretical analysis of the entropy dissipation term. A standard proof of this result was then obtained by Lions [L] using a method based on the regularizing theorem and on approximations. In the relativistic situation, one may apply Arkeryd's or Lions' approach to obtain strong L^1 convergence to a local Jüttner equilibrium solution. Below we will see that the periodicity in the space variables implies that every local Jüttner solution is, in fact, a global one. This will require some different arguments in comparison to those given by Arkeryd [Ar4] in the nonrelativistic case.

We now specify the assumptions and the asymptotic result in detail. Consider the relativistic Boltzmann equation in a periodic box Λ, which after rescaling can be taken to be $\mathbb{R}^3/\mathbb{Z}^3$. The kernel satisfies the conditions given in §1. Further, we assume that the initial density distribution $f_0 = f(0)$ has finite energy and entropy

$$\int_\Lambda \int_{\mathbb{R}^3} f_0(p_0 + |\log f_0|) dp dx < \infty. \tag{3.1}$$

Solutions to (1.1) will be in the sense of renormalized solutions or any equivalent form (iterated-integral, exponential-multiplier, or mild-solution form). For a discussion of these matters, see [DPL1], [Ar2], and [Ar5].

We assume that the solutions of the initial value problem (1.1) together with (3.1) satisfy

$$\int_\Lambda \int_{\mathbb{R}^3} f(t)(p_0 + |\log f(t)|) dp dx \leq C, \quad t \geq 0, \tag{3.2}$$

and

$$0 \leq \int_0^\infty ds \int_E (f'f'_* - ff_*) \log \frac{f'f'_*}{ff_*} B d\mu \leq C. \tag{3.3}$$

Here $E = \Lambda \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{S}^2$ and

$$d\mu = \frac{dp}{p_0}\frac{dq}{q_0} dx d\Omega,$$

and we write $f = f(t,x,p)$, $f' = f(t,x,p')$, $f_* = f(t,x,q)$, and $f'_* = f(t,x,q')$. The solutions in [DEJ], adapted to the periodic case, satisfy (3.2). Condition (3.3) is a conseqence of the relativistic form of the entropy inequality

$$\begin{aligned} &\int_\Lambda \int_{\mathbb{R}^3} f \log f dp dx - \int_\Lambda \int_{\mathbb{R}^3} f_0 \log f_0 dp dx \\ &\qquad + \frac{1}{4} \int_0^t ds \int_E (f'f'_* - ff_*) \log \frac{f'f'_*}{ff_*} B d\mu \leq 0, \quad t \geq 0. \end{aligned} \tag{3.4}$$

This inequality is easily derived by reworking the proof of the corresponding inequality in the nonrelativistic case [DPL2] using the results in [DEJ]. To see that (3.3) can be derived from (3.4), we only have to show that the first term in (3.4) is a bounded function of t. In fact, since the integrand of the third term is positive, only boundedness from below remains to be shown. For this, we note that for all nonnegative functions g on $\Lambda \times \mathbb{R}^3$ and all $R < \infty$

$$\begin{aligned} \int_\Lambda \int_{B_R} g|\log g| dp dx &= \int_\Lambda \int_{B_R} g \log g dp dx - 2 \iint_{g \leq 1} g \log g dp dx \\ &\leq \int_\Lambda \int_{B_R} (g \log g + 2 g p_0) dp dx + 2 \int_\Lambda \int_{B_R} g \left(\log \frac{1}{g}\right) 1_{(g \leq \exp(-p_0))} dp dx. \end{aligned}$$

Applying the elementary inequality $t\log(1/t) \leq C\sqrt{t}$ for $0 < t < 1$ and some $C \geq 0$ and letting $R \to \infty$, we get

$$\int_\Lambda \int_{\mathbb{R}^3} g|\log g| dpdx \leq \int_\Lambda \int_{\mathbb{R}^3} (g\log g + 2gp_0)dpdx + C' \tag{3.5}$$

for some positive constant C' independent of g. Our claim follows from this general inequality and the fact that the energy and entropy of the solutions are bounded (see (3.2)). The driving force to equilibrium is, of course, the entropy inequality (3.3).

Let us now state the theorem concerning the asymptotic behaviour of solutions to the relativistic Boltzmann equation.

THEOREM 3. *Given a sequence $(t_k)_{k\in N}$, $t_k \nearrow \infty$, there is a subsequence $(t_{k'})$ and a global Jüttner equilibrium solution*

$$J(p) = \exp(\alpha - \beta_\mu p^\mu), \qquad \alpha \in \mathbb{R}, \quad \beta_\mu = (\beta_0, \beta) \in \mathbb{R}^4 \quad \textit{with } \beta_0 > |\beta|,$$

such that for $T > 0$, $f(\cdot + t_{k'}) \to J$ strongly in $L^1(\Lambda \times \mathbb{R}^3 \times [0,T])$, and for $t > 0$, $f(\cdot, t + t_{k'}) \to J$ strongly in $L^1(\Lambda \times \mathbb{R}^3)$.

Remarks. (1) The condition $\beta_0 > |\beta|$ (i.e., β_μ timelike) is necessary and sufficient for $J \in L^1(\mathbb{R}^3)$. (2) As in the classical case, we can not exclude that there are different limits J for different sequences $(t_k)_{k\in N}$. If the open question of energy conservation would find an affirmative resolution, then the uniqueness would follow.

Sketch of proof. Lions' method to prove the nonrelativistic version of Theorem 3 is based on the regularizing theorem and on approximations. The regularization, which crucially depends on the collision geometry, is in our case taken care of by Theorem 1. The approximations are easily adapted to the relativistic setting since they do not explicitly depend on the collision geometry. We remark that the minor difference in the hypothesis of the kernel b in the regularizing theorem, between the classical [L] and relativistic settings (Theorem 1), is irrelevant for the reworking of Lions' proof. Indeed, Lions only has to consider functions with compact support, so the fact that b has compact support according to the hypotheses of Theorem 1 does not affect the proof.

The author has applied Arkeryd's method to the relativistic case in [An]. In order to realize that proof, some specific results concerning the relativistic Boltzmann equation are needed. References of these results and details of the nonstandard approach are found in [An].

In conclusion, whether we rework Arkeryd's nonstandard proof or Lions' standard proof, we will obtain convergence to a local Jüttner solution J in the two senses given in Theorem 3. The local Jüttner solution, periodic in the space variables, satisfies the equation

$$\left(\partial_t + \frac{p}{p_0} \cdot \nabla_x\right) J = 0 \tag{3.6}$$

in the distribution sense and has the form

$$J(x,t,p) = \exp(\alpha(x,t) - \beta_\mu(x,t)p^\mu). \tag{3.7}$$

Here α and β are Lebesgue measurable functions with $\alpha(x,t)$, $\beta(x,t) \in \mathbb{R}$ for a.e. $(x,t) \in \Lambda \times \mathbb{R}_+$, and β_μ is timelike. Next, we show that the periodicity in the space variables implies that α and β are independent of x and t, i.e., J is a global Jüttner solution.

Now, not only J but also $\log J$ satisfies (3.6) in the distribution sense. This follows from a simple generalization of the arguments given by L. Desvillettes [De] in the analogous nonrelativistic situation. Thus

$$\left(\partial_t + \frac{p}{p_0} \cdot \nabla_x\right) [\alpha(x,t) - \beta_\mu(x,t)p^\mu] = 0 \tag{3.8}$$

holds in the distribution sense. Evaluating the derivatives and identifying the coefficients in front of the different functions of p by zero, we obtain

$$\partial_\mu \alpha = 0, \tag{3.9}$$

$$\partial_\mu \beta_\nu + \partial_\nu \beta_\mu = 0, \tag{3.10}$$

where $\partial_\mu f = \frac{\partial f}{\partial x_\mu}$ and $x_\mu = g_{\mu\nu} x^\nu = (t, -x)$, $g_{\mu\nu} = (+ - --)$.

Obviously, the function α is space–time independent, i.e., constant. The system (3.10) consists of ten equations in four unknowns, β_μ, and is called Killing's equation. Now, it is well known that the general solution of (3.10) is

$$\beta_\rho = \gamma_\rho + \omega_{\rho\nu} x^\nu, \tag{3.11}$$

where γ_ρ and $\omega_{\rho\nu}$ are constants and $\omega_{\rho\nu} = -\omega_{\nu\rho}$. Since $\omega_{\rho\nu}$ is antisymmetric, we have $\omega_{00} = 0$, so there is no time dependence in β_0 and the condition $\beta_0 > |\beta|$ implies $\omega_{\rho 0} = 0 = -\omega_{\rho 0}$. The periodicity condition forces the remaining coefficients of $\omega_{\rho\nu}$ to vanish, so $\omega_{\rho\nu} \equiv 0$. Hence J is a global Jüttner solution. □

The next application of Theorem 1 concerns the functional equation

$$f(p)f(q) = f(p')f(q') \quad \text{a.e. on } \mathbb{R}^{2N}_{p,q} \times \mathbb{S}^{N-1}, \tag{3.12}$$

where, as usual, $p + q = p' + q'$ and $p_0 + q_0 = p'_0 + q'_0$. This equation has been extensively studied under different assumptions on the regularity of f. Irrespective of the assumptions, the solutions turn out to be the Jüttner equilibrium solutions. This fact has been shown, for instance, by Chernikov [C], Bichteler [B], Marle [M], and Dijkstra [Di] under the assumption that solutions are differentiable of order 1 [B], continuous [C], [M] or measurable [Di]. However, a beautiful trick of Lions based on the regularizing theorem shows that solutions to the nonrelativistic analogue of (3.12) are necessarily smooth. In view of Theorem 1, the result is also available in the relativistic case. The generalization is straightforward, but we will present the proof and not only refer to Lions. The structure of the proof is essential for the discussion in the appendix, indicating a point of connection between Arkeryd's and Lions' approaches to the asymptotic problem (Theorem 3). Hence let us prove the following lemma.

LEMMA 2. *Solutions* $0 \leq f \in L^1_{\text{loc}}(\mathbb{R}^N)$ *of the functional equation* (3.12) *are smooth.*

Proof. If $0 \leq f \in L^1_{\text{loc}}(\mathbb{R}^N)$ is a solution of (3.12), then $g := \sqrt{f} \in L^2_{\text{loc}}(\mathbb{R}^N)$ also satisfies (3.12). If $g \equiv 0$, we are done. Otherwise, we introduce

$$b_\epsilon(p,q,\omega) = \zeta_R(|q|)\, \varphi^{(1)}_\epsilon(|p|)\, \varphi^{(2)}_\epsilon(|p-q|) \\ \times \varphi^{(3)}_\epsilon\left(\frac{|(\hat{p}-\hat{q})\cdot\omega|}{|\hat{p}-\hat{q}|}\right) \varphi^{(4)}_\epsilon\left(\frac{|(p-q)\cdot\omega|}{|p-q|}\right),$$

where $0 \leq \zeta_R, \varphi^{(j)}_\epsilon \leq 1$, $\zeta_R, \varphi^{(1)}_\epsilon \in C^\infty_0(\mathbb{R})$, $\zeta_R \equiv 1$ for $|q| \leq R$, and $\varphi^{(1)}_\epsilon \equiv 1$ for $|p| \leq \epsilon^{-1}$. Further, $\varphi^{(2,3)}_\epsilon \in C^\infty(\mathbb{R})$, supp $\varphi^{(2,3)}_\epsilon \subset (\epsilon/2, \infty)$, and $\varphi^{(2,3)}_\epsilon(t) = 1$ for

$t \geq \epsilon$. Finally, $\varphi_\epsilon^{(4)} \in C^\infty([0,1))$, supp $\varphi_\epsilon^{(4)} \subset [0, 1-\epsilon/2)$, and $\varphi_\epsilon^{(4)}(t) = 1$ for $t \leq 1 - \epsilon$.

Hence, for each $R > 0$, we have

$$l_\epsilon(q) := \int_{\mathbb{R}^N} \int_{\mathbb{S}^{N-1}} b_\epsilon(p,q,\omega) g(p) d\omega\, dp > 0 \tag{3.13}$$

on $\{|q| \leq R\}$ for ϵ small enough, and $l_\epsilon \in C_0^\infty(\mathbb{R}^N)$. In view of (3.12), we find that $g = Q^+(g,g)/l_\epsilon$ on $\{|q| \leq R\}$, and by Theorem 1, we obtain $g \in H_{\text{loc}}^{(N-1)/2}(\mathbb{R}^N)$. Next, we observe from the proof of this theorem that if $g \in L^1(\mathbb{R}^N)$ and $f \in H^s(\mathbb{R}^N)$, $s \in \mathbb{R}$, we have

$$\|Q^+(f,g)\|_{H^{s+\frac{N-1}{2}}} \leq C\|f\|_{H^s}\|g\|_{L^1}.$$

Therefore, by iterating the argument above, we deduce that $g \in H_{\text{loc}}^{k(N-1)/2}(\mathbb{R}^N)$ for all $k \geq 1$. Accordingly, $g \in C^\infty(\mathbb{R}^N)$. □

Appendix. A connection between the standard and the nonstandard approach. We will present a different and simpler proof of a lemma by Arkeryd [Ar1], [Ar3]. The lemma is crucial in his nonstandard (ns) approach to the asymptotic problem (i.e., the nonrelativistic analogue of Theorem 3). Our proof is based on the proof of Lemma 2 and hence on the regularizing theorem, which is the crucial part in Lions' approach to the asymptotic problem. A link between their methods of proofs is therefore indicated. In what follows, some nonstandard techniques and notations are used, and the reader not familiar with this matter may consult [HL] or [AFHL].

LEMMA 3 (see [Ar1], [Ar3]). *Let $f \in {}^\star L_+^1(\mathbb{R}^N)$ ($N \geq 2$ and finite) be given with*

$$\int_{{}^\star\mathbb{R}^N} p_0 f(p)\, {}^\star dp \tag{A.1}$$

finite and with

$$f(p)f(q) \approx f(p')f(q') \tag{A.2}$$

for Loeb a.e. $(p,q,\omega) \in \text{ns}\, {}^\star(\mathbb{R}^N \times \mathbb{R}^N \times \mathbb{S}^{N-1})$. Then either $f(p) \approx 0$ for Loeb a.e. $p \in \text{ns}\, {}^\star\mathbb{R}^N$ or ${}^\circ f(p) > 0$ for Loeb a.e. $p \in \text{ns}\, {}^\star\mathbb{R}^N$.

Remarks. (1) Condition (A.1) is a slight modification of the relativistic analogue of Arkeryd's original formulation. (2) Condition (A.1) together with $\int f(p) \log f(p)\, {}^\star dp$ finite implies that f is S-integrable, which is essential for the nonstandard approach to the asymptotic problem. For details we refer to [An].

Proof. For simplicity, we will consider only the two-dimensional case. It will be clear from the proof that the result in higher dimensions follows by a simple iteration of our arguments. Now, we rework the proof of Lemma 2 in the nonstandard context. Assume that ${}^\circ \int f\, {}^\star dp > 0$; if that is not the case, we have $f \approx 0$ Loeb a.e. in ${}^\star\mathbb{R}^2$. Observe that condition (A.1) implies that f is essentially concentrated in ns ${}^\star\mathbb{R}^2$; hence for some noninfinitesimal ϵ, we obtain $l_\epsilon > 0$. Accordingly, relation (A.2) implies that $f \approx h$ Loeb a.e., in ns ${}^\star\mathbb{R}^2$, where $\|\varphi h\|_{{}^\star H^1}$ is finite and φ is a localized standard function in $C_0^\infty(\mathbb{R}^2)$. Actually, $\|\varphi h\|_{{}^\star H^m}$ is finite for m finite (so, in addition, $h \in {}^\star C^k$ for k finite). Now, there is a $\star$-measurable set A of finite diameter and positive Loeb measure in ns ${}^\star\mathbb{R}^2$, where $h > \epsilon$ for some $\epsilon > 0$, $\epsilon \in \mathbb{R}$. Otherwise, $h \approx 0$ Loeb a.e. in ns ${}^\star\mathbb{R}^2$, implying $f \approx 0$ Loeb a.e. in ns ${}^\star\mathbb{R}^2$. We will show that $h > \epsilon/2$ on a ball with positive Loeb measure in ${}^\star\mathbb{R}^2$, and this will, in fact, be enough, in view of well-known arguments [Ar1], to conclude that ${}^\circ h > 0$ Loeb a.e. in ns ${}^\star\mathbb{R}^2$. Let $p \in A$, $\hat{d}$ be a unit vector in ${}^\star\mathbb{R}^2$, and $L(p, \delta, \hat{d})$ be the line segment between p and $p + \delta\hat{d}$. We then have the relation $|h(p+\delta\hat{d}) - h(p)| = |\int_{L(p,\delta,\hat{d})}(\nabla h\cdot\hat{d})\,{}^\star dp|$. If we could find a point $p_0 \in A$ and a positive number $\delta_0 \in \mathbb{R}$ such that

$$\left|\int_{L(p_0,\delta,\hat{d})} (\nabla h\cdot\hat{d})\,{}^\star dp\right| < \frac{\epsilon}{2} \quad \forall\hat{d} \in {}^\star\mathbb{S}^1 \text{ and } \forall\delta,\ 0 \le \delta \le \delta_0, \tag{A.3}$$

then ${}^\circ h > \epsilon/2$ on the ball $B(p_0, \delta_0)$. The proof is by contradiction. Assume that there is no point $p_0 \in A$ with such a property. We will show that this assumption implies that φh can not be in ${}^\star H^2$, where $\varphi \in C_0^\infty$ with $\varphi \equiv 1$ on a ball containing A (in N dimensions, we need ${}^\star H^N$). Let each point $p \in A$ be a center of a γ^{-1}-cube $C(p, \gamma^{-1})$, where $\gamma \in \mathbb{R}$ is large. From Wiener's covering lemma, there are at least $\eta := [\gamma^2 a]$ disjoint cubes (in three dimensions, $\eta \sim \gamma^3$), where a is a constant depending only on the dimension and the Loeb measure of A, and $[x]$ denotes the integer part of x. Let us denote the centers of the corresponding cubes by p_j, $j = 1, \ldots, \eta$. The assumption above implies that for each p_j, there is a unit vector $\hat{d}_j$ in ${}^\star\mathbb{R}^2$, such that

$$\int_{L(p_j,\hat{d}_j,\gamma^{-1}/2)} |(\nabla h\cdot\hat{d}_j)|\,{}^\star dp \ge \frac{\epsilon}{2}. \tag{A.4}$$

Without loss of generality, we assume that all of the unit vectors $\hat{d}_j$ are directed along p_1. Indeed, if this is not the case, it is easy to see that the estimates below will be changed by a factor depending only on the dimension. Let us fix j and denote by $L(\alpha)$ the line segment

$$L(\alpha) := \{p : p = p_j + \alpha n_2 + s n_1,\ s \in (0, \gamma^{-1}/2)\}, \tag{A.5}$$

where $n_{1,2}$ are unit vectors directed along the $p_{1,2}$-axes, respectively. Set $I_\alpha := \int_{L(\alpha)} |\frac{\partial h}{\partial p_1}|\,{}^\star dp_1$ and note that $I_0 \ge \epsilon/2$. Now, either $I_\alpha \ge \epsilon/4$ for $\alpha \in [0, \gamma^{-1}/2]$ or there exist some $\alpha_0 \in (0, \gamma^{-1}/2)$ with $I_{\alpha_0} < \epsilon/4$. In the first case,

$$\int_0^{\gamma^{-1}/2}\int_{L(\alpha)} |\partial_1 h|\,{}^\star dp_1 dp_2 \ge \gamma^{-1}\frac{\epsilon}{8}, \tag{A.6}$$

and in the second case, we obtain from Stoke's theorem (recall $|\nabla|f|| = |\nabla f|$ a.e. when $f \in H^1$; see, e.g., [Au, p. 82])

$$\int_0^{\alpha_0}\int_{L(\alpha)} |\partial_2\partial_1 h|\,{}^\star dp_1 dp_2 \ge \left|\int_0^{\alpha_0}\int_{L(\alpha)} \partial_2|\partial_1 h|\,{}^\star dp_1 dp_2\right| \tag{A.7}$$

$$= \left|\int_{L(0)} n_2|\partial_1 h|\,{}^\star dp_1 + \int_{L(\alpha_0)} n_2|\partial_1 h|\,{}^\star dp_1\right| \ge \frac{\epsilon}{2} - \frac{\epsilon}{4} = \frac{\epsilon}{4}. \tag{A.8}$$

There are η cubes, and one of the cases above occurs at least $\eta/2$ times. If the first case occurs $\eta/2$ times, then

$$\sum_{j=1}^{\eta}\int_{C(p_j,\gamma^{-1})} |\partial_1 h|^\star dp_1 dp_2 \ge \eta\epsilon\gamma^{-1}/16 \sim \gamma\epsilon \to \infty \quad \text{as } \gamma \to \infty. \tag{A.9}$$

In the second case,

$$\text{(A.10)} \qquad \sum_{j=1}^{\eta} \int_{C(p_j,\gamma^{-1})} |\partial_1 \partial_2 h|^\star dp_1 dp_2 \geq \eta\epsilon/8 \sim \gamma^2 \epsilon \to \infty \quad \text{as } \gamma \to \infty.$$

Thus φh cannot be in $^\star H^2$, and this completes the proof. □

Acknowledgments. I would like to thank Leif Arkeryd for suggesting the fascinating topic of relativistic kinetics and for the constant interest he has shown regarding my work.

I would also like to thank Bo Berndtsson for instructive and stimulating discussions concerning Fourier integral operators and Bernt Wennberg for valuable discussions and comments on the manuscript.

REFERENCES

[An] H. Andréasson, *A regularity property and strong L^1 convergence to equilibrium for the relativistic Boltzmann equation,* Technical Report 21, Department of Mathematics, Chalmers University of Technology, Göteborg, Sweden, 1994.

[Ar1] L. Arkeryd, *On the Boltzmann equation in unbounded space far from equilibrium, and the limit of zero mean free path,* Comm. Math. Phys., 105 (1986), pp. 205–219.

[Ar2] ———, *On the Enskog equation in two space variables,* Transport. Theory Stat. Phys., 15 (1986), pp. 673–691.

[Ar3] ———, *The nonlinear Boltzmann equation far from equilibrium,* in Nonstandard Analysis and Its Applications, Cambridge University Press, Cambridge, UK, 1988.

[Ar4] ———, *On the long time behaviour of the Boltzmann equation in a periodic box,* Technical Report 23, Department of Mathematics, Göteborg University, Göteborg, Sweden, 1988.

[Ar5] ———, *On the strong L^1 trend to equilibrium for the Boltzmann equation,* Stud. Appl. Math., 87 (1992), pp. 283–288.

[ACB] A. Anile and Y. Choquet-Bruhat, eds., *Relativistic Fluid Dynamics*, Lecture Notes in Math., Springer-Verlag, Berlin, 1989.

[AFHL] S. Albeverio, J. E. Fenstad, R. Høegh-Krohn, and T. Lindstrøm, *Nonstandard Methods in Stochastic Analysis and Mathematical Physics*, Academic Press, New York, 1986.

[AG] S. Alinhac and P. Gérard, *Opérateurs Pseudo-Différentiels et Théorème de Nash-Moser*, Inter-Editions et Editions du CNRS, Paris, 1991.

[Au] T. Aubin, *Nonlinear Analysis on Manifolds: Monge–Ampère Equations,* Springer-Verlag, New York, 1982.

[B] K. Bichteler, *Bemerkungen über relativistische Stossinvarianten,* Z. Physik, 182 (1965), pp. 521–523.

[C] N. A. Chernikov, *Equilibrium distribution of the relativistic gas,* Acta Phys. Polon., 27 (1964), pp. 1069–1092.

[CF] A. Cordoba and C. Fefferman, *Wave packets and Fourier integral operators,* Comm. Partial Differential Equations, 3 (1978), pp. 979–1005.

[De] L. Desvillettes, *Convergence to equilibrium in large time for Boltzmann and B.G.K. equations,* Rapport Scientifique pour obtenir l'habilitation à diriger des recherches en mathématiques, École Normale de l'Enseignement Technique, Cachan, France, 1994.

[DEJ] M. Dudyński and M. Ekiel-Jeżewska, *Global existence proof for the relativistic Boltzmann equation,* J. Stat. Phys., 66 (1992), pp. 991–1001.

[Di] J. J. Dijkstra, *The mathematical aspects of relativistic kinetic theory* II: *The summational invariants* (*A*), Proc. Kon. Nederl. Akad. Wetensch., 81B (1978), pp. 265–275.

[DPL1] R. DiPerna and P. L. Lions, *On the Cauchy Problem for Boltzmann equations: Global existence and weak stability,* Ann. Math., 130 (1989), pp. 321–366.

[DPL2] ———, *Global solutions of Boltzmann's equation and the entropy inequality,* Arch. Rational Mech. Anal., 114 (1991), pp. 47–55.

[GLW] S. R. de Groot, W. A. van Leeuwen, and C. G. van Weert, *Relativistic Kinetic Theory,* North–Holland, Amsterdam, 1980.

[GlS1] R. Glassey and W. Strauss, *On the derivatives of the collision map of relativistic particles,* Transport Theory Stat. Phys., 20 (1991), pp. 55–68.

[GlS2] ———, *Asymptotic stability of the relativistic equilibrium,* Publ. Res. Inst. Math. Sci., 29 (1992),, pp. 301–347.

[GlS3] ———, *Asymptotic stability of the relativistic Maxwellian via fourteen moments,* Transport Theory Stat. Phys., 24 (1995), pp. 657–678.

[GuS] V. Guillemin and S. Sternberg, *Geometrical Asymptotics,* Surveys 14, American Mathematical Society, Providence, RI, 1977.

[H1] L. Hörmander, *Fourier integral operators* I, Acta Math., 127 (1971), pp. 79–183.

[H2] ———, *The Analysis of Linear Partial Differential Operators* I, 2nd ed., Springer-Verlag, Berlin, 1990.

[H3] ———, *The Analysis of Linear Partial Differential Operators* III, Springer-Verlag, Berlin, 1985.

[H4] ———, *The Analysis of Linear Partial Differential Operators* IV, Springer-Verlag, Berlin, 1985.

[HL] A. Hurd and P. Loeb, *An Introduction to Nonstandard Real Analysis,* Academic Press, New York, 1985.

[L] P. L. Lions, *Compactness in Boltzmann's equation via Fourier integral operators and applications* I, J. Math. Kyoto Univ., 34 (1994), pp. 391–427.

[M] C. Marle, *Sur l'établissement des équations de l'hydrodynamique des fluides relativistes dissipatifs* I: *L'équation de Boltzmann relativiste,* Ann. Inst. Henri Poincaré, 10 (1969), pp. 67–126.

[St] J. M. Stewart, *Non-Equilibrium Relativistic Kinetic Theory,* Lecture Notes in Phys. 10, Springer-Verlag, Berlin 1971.

[Sy] J. L. Synge, *The Relativistic Gas*, North–Holland, Amsterdam, 1957.

[W] B. Wennberg, *Regularity estimates for the Boltzmann equation,* Comm. Partial Differential Equations, 19 (1994), pp. 2057–2074.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1406–1423, September 1996

012

A TRANSMISSION PROBLEM IN THE SCATTERING OF ELECTROMAGNETIC WAVES BY A PENETRABLE OBJECT*

RODOLFO H. TORRES†

Abstract. Layer-potential techniques are used to study a transmission problem arising in the scattering of electromagnetic waves by a penetrable object. The method proposed does not involve the use of the calculus of pseudodifferential operators and hence it can be applied in domains with very little regularity. The solutions are represented as a combination of a curl and a double curl of a single layer-potential operator. The work relies on the important harmonic-analysis tools developed in recent years to study boundary-value problems in domains with minimal regularity assumptions.

Key words. Maxwell equations, reduced wave equation, layer-potential methods, transmission problems, scattering theory, nonsmooth domains

AMS subject classifications. 35J05, 35Q60, 45P05, 58G20, 31A25

1. Introduction.

A classical problem arising in electromagnetism is that of determining the field scattered by a penetrable object from the knowledge of the tangential component on the surface of the object of an incoming field. See, e.g., [12] and [15]. The mathematical formulation of this problem leads to a transmission problem for the Maxwell equations on a bounded domain (see §2 below for the precise statement). The problem has been studied using several approaches based on layer-potentials techniques. In particular, we want to mention works by Wilde [20] and Costabel and Stephan [5]. Reference to related works can be found therein.

For time-harmonic electromagnetic waves, the solution of Maxwell equations are divergence-free solutions of the vector Helmholtz equation. In [20], very general transmission problems for the vector Helmholtz equation are considered. The solutions of the problems are obtained as a combination of several single- and double-layer potentials after solving, in appropriate Hölder spaces, a 4×4 system of of integral equations of the second kind on the boundary of the domain. This classical method requires the domain to be at least of class C^2 and, as a consequence, the solutions have continuous partial derivatives up to the boundary of the domain. On the other hand, in [5], the so-called direct method is used. This is a general method applicable to strongly elliptic boundary-value problems and relies on the coercivity on certain Sobolev spaces (the energy spaces) of a bilinear form related to the boundary data. In [5], the electromagnetic problem is transformed into a particular transmission problem for the vector Helmholtz equation which is solved, again, by inverting a matrix of operators on the boundary of the domain. In this work, the calculus of pseudodifferential operators is used and hence the domain is assumed to be C^∞. In addition, the boundary values of the solutions are prescribed in the distributional sense and not pointwise. The purpose of this paper is to develop an alternative approach to study the electromagnetic transmission problem in domains which are less regular than the one considered in the works just mentioned, allowing less regular boundary data, but still obtaining solutions whose boundary values are prescribed pointwise (nontangentially).

As is well known, the study of boundary-value problems using layer-potential techniques in domains which are C^1 or Lipschitz is very delicate. One of the main reasons for this is that some of the resulting integral operators on the boundary of the

* Received by the editors May 6, 1994; accepted for publication February 16, 1995. This research was supported in part by NSF grant DMS-9303363.

† Department of Mathematics, University of Michigan, Ann Arbor, MI 48109. Current address: Department of Mathematics, University of Kansas, Lawrence, KS 66045-2142.

domain have to be interpreted as principal-value singular integrals. In particular, to consider L^p data and solutions with boundary values obtained pointwise, deep results from harmonic analysis are necessary. Dirichlet and conormal derivative problems for several equations and system of equations in nonsmooth domains have already been studied using harmonic-analysis techniques. A few examples are [9], [11], [17], [18], [7]. Using similar techniques, transmission problems have been considered in [8] and [16]. This last paper deals with the case of the scalar Helmholtz equation in Lipschitz domains. See also [19] and [6], where an approach to transmission problems related to [5] is used.

The study of the potential operators associated with Maxwell equations in C^1 and Lipschitz domains has been recently carried out in [13] and [14]. In particular, the so-called Maxwell, electric, and magnetic boundary-value problems for a perfect conducting object were solved with optimal estimates in the case of C^1 domains. This work depends heavily on the results in [2] and [3] about the Cauchy integral operator on Lipschitz curves as well as the developments in [9]. For the previously known results about these problems in the case of smoother domains, we refer to [4].

In this paper, we will combine the results of [13] with some of the ideas in [16] to study the electromagnetic transmission problem in domains which are only C^1 or Lipschitz. Unlike the approaches in [20] and [5], we propose as a solution for the electromagnetic transmission problem a combination of the curl and the curlcurl of the single-layer potential. After taking traces, this ansatz leads to a 2×2 system of integral operators on the boundary of the domain. The trace operator associated with the double curl of the single-layer potential is hypersingular (even on smooth domains). Nevertheless, in the case of the electromagnetic transmission problem, this operator appears in a regularized way. This allows us to consider it on an appropriate space of functions: the space $L_T^{2,\mathrm{Div}}$ consisting of tangential vector fields with surface divergence in L^2. It was shown in [14] that $L_T^{2,\mathrm{Div}}$ is the right space of boundary data to work with in domains with little regularity. As in [16], the solution of the system of integrals operators on the boundary relies on the knowledge of the spectrum of a singular integral operator. In our present situation, the singular-integral operator is the one obtained as the tangential component of the trace of the curl of the single-layer potential.

The paper is organized as follows. In §2, we recall some basic facts about nonsmooth domains and state the transmission problem with boundary data in $L_T^{2,\mathrm{Div}}$. In §3, we show for appropriate values of the electromagnetic characteristics of the object and surrounding media the uniqueness of solution to the problem in the case of Lipschitz domains. In §4, we collect several results from [13] about the layer-potential operators associated with Maxwell equations and include some new results regarding the double curl of the single-layer potential. In §5, we show some existence results.

2. The electromagnetic transmission problem.

The notation that we use is standard for the subject. In particular, we will follow very closely that of [13] which is our main reference. For the purposes of this paper, a Lipschitz, respectively, C^1, domain will always be an open, simply connected domain D of R^3, whose boundary, ∂D, is given locally by the graph of a Lipschitz, respectively, C^1, function. Let N be the exterior unit normal to ∂D and let $d\sigma$ denote surface measure on the boundary. The spaces $L^2(\partial D)$ of functions or vector fields and the space $L_T^2(\partial D)$ of tangential vector fields are defined with respect to $d\sigma$. The space $L^{2,1}(\partial D)$ is, as usual, the space of L^2 functions with tangential derivatives also in L^2. A vector field $A \in L_T^2(\partial D)$ is said to have a surface divergence if there exists a function $b \in L^2(\partial D)$ such that

$$\int_{\partial D} \langle \nabla_T \psi, A \rangle \, d\sigma = - \int_{\partial D} \psi \, b \, d\sigma$$

for all functions ψ which are Lipschitz in a neighborhood of ∂D. Here ∇_T denotes the tangential gradient and $\langle \cdot, \cdot \rangle$ denotes the inner product in R^3. The function b is denoted by $\operatorname{Div} A$ and the space of all such vector fields (see, e.g., [14]) is denoted by $L_T^{2,\mathrm{Div}}(\partial D)$. The space is equipped with the norm

$$\|A\|_{L_T^{2,\mathrm{Div}}(\partial D)} = \|A\|_{L^2(\partial D)} + \|\operatorname{Div} A\|_{L^2(\partial D)}.$$

At every point Q in the boundary of the domain, we consider an open, right-circular, doubly truncated cone $\Gamma(Q)$, with vertex at Q and two convex components, $\Gamma_i(Q)$ in D and $\Gamma_e(Q)$ in $\mathrm{R}^3 \backslash \overline{D}$, so that the resulting family of cones is a regular family in the sense of [17]. For a function u defined in D, the nontangential maximal function of u is defined by

$$u^*(P) = \sup_{X \in \Gamma_i(P)} |u(X)|.$$

The boundary values of functions defined inside D are assumed to be taken in nontangential fashion and almost everywhere with respect to $d\sigma$. That is, $u|_{\partial D}$ is to be interpreted as

$$u(P) = \lim_{\substack{X \longrightarrow P \\ X \in \Gamma_i(P)}} u(X),$$

whenever such a limit exists for almost every point in ∂D. Similar definitions apply for derivatives of a function and for each component of a vector-valued function. For example, if × denotes the exterior product in R^3 and A is a vector field defined inside D, then $N \times \operatorname{curl} A|_{\partial D}$ is given by

$$N \times \operatorname{curl} A(P) = \lim_{\substack{X \longrightarrow P \\ X \in \Gamma_i(P)}} N(P) \times \operatorname{curl} A(X).$$

For functions defined in the exterior of D, the nontangential maximal function and the boundary values are defined in the same way but using $\Gamma_e(P)$.

We can now state the electromagnetic transmission problem that we want to study. We follow the classical description in [15]. Let D represent an object made of an homogeneous material, and assume that the object is immersed in an homogeneous medium represented by the exterior of D. In all space, we consider a time-harmonic electromagnetic wave with frequency ω, described by the electric and magnetic vector fields E and H. These fields satisfy the Maxwell equations

$$\operatorname{curl} E = i\omega\mu_i H$$

$$\operatorname{curl} H = -i\omega\epsilon_i E \quad \text{in } D,$$

and

$$\operatorname{curl} E = i\omega\mu_e H,$$

$$\operatorname{curl} H = -i\omega\epsilon_e E \quad \text{in } \mathrm{R}^3\backslash\overline{D}.$$

The electromagnetic parameters of the object and the surrounding medium in the above equations are, respectively,

$$\epsilon_i = \epsilon_{0i} + \frac{i\sigma_i}{\omega}, \qquad \mu_i = \mu_{0i} + \frac{i\hat{\sigma}_i}{\omega},$$

$$\epsilon_e = \epsilon_{0e} + \frac{i\sigma_e}{\omega}, \qquad \mu_e = \mu_{0e} + \frac{i\hat{\sigma}_e}{\omega},$$

where ϵ_{0i} and ϵ_{0e} are the dielectric constants, μ_{0i} and μ_{0e} are the permeability, σ_i and σ_e are the electric conductivity, and $\hat{\sigma}_i$ and $\hat{\sigma}_e$ are the magnetic conductivity of each medium. The usual restrictions on the values of these parameters are

$$(1) \qquad 0 \leq \arg\omega < \pi,$$

$$(2) \qquad \epsilon_{0i},\, \epsilon_{0e} > 0 \quad \text{and} \quad \mu_{0i},\, \mu_{0e} > 0,$$

$$(3) \qquad \sigma_i,\, \sigma_e \geq 0 \quad \text{and} \quad \hat{\sigma}_i,\, \hat{\sigma}_e \geq 0$$

(see [15]). We will assume $\epsilon_i \neq \epsilon_e$ and $\mu_i \neq \mu_e$. The wave numbers in the interior and exterior of the obstacle are defined by

$$k_i^2 = \omega^2\epsilon_i\mu_i \quad \text{and} \quad k_e^2 = \omega^2\epsilon_e\mu_e,$$

where we assume

$$(4) \qquad 0 \leq \arg k_i, \qquad \arg k_e < \pi.$$

In the exterior of D, the vector fields are decomposed as the sum of a known incoming field and an unknown scattered field,

$$E = E_{\text{in}} + E_{\text{sc}},$$

$$H = H_{\text{in}} + H_{\text{sc}}.$$

Both the incoming and scattered fields satisfy Maxwell's equations in the exterior of D. We also assume that the scattered fields satisfy the radiation conditions

$$(5) \qquad \omega\mu_e\frac{X}{|X|} \times H_{\text{sc}} + k_e E_{\text{sc}} = o(|X|^{-1}) \quad \text{and} \quad E_{\text{sc}} = O(|X|^{-1})$$

as $|X| \to \infty$. The tangential components of the total vector fields must extend continuously across the boundary, so on ∂D we must have

$$N \times E - N \times E_{\text{sc}} = N \times E_{\text{in}},$$

$$N \times H - N \times H_{\text{sc}} = N \times H_{\text{in}},$$

where the values of $N \times E$ and $N \times H$ are taken from inside D. It follows that, in order to obtain the total electric and magnetic fields from the knowledge of the incoming fields, we can consider a transmission boundary-value problem with the tangential components of the incoming fields as datum. Because of the results in [13] and [14], we will assume that these tangential components are in $L_T^{2,\mathrm{Div}}(\partial D)$, and we will require the solutions to have nontangential maximal functions bounded on $L^2(\partial D)$. In fact, it was shown in [14] that if E and H solve Maxwell equations in a C^1 or Lipschitz domain and E has pointwise (nontangentially) boundary values in $L^2(\partial D)$ with bounded nontangential maximal function, then the companion field H also have pointwise boundary values if and only if the tangential component of E is in $L_T^{2,\mathrm{Div}}(\partial D)$. Since the roles of E and H can be interchanged, we have to require the same kind of boundary data for the tangential component of H. Thus we are lead to consider the following problem. Given two tangential vector fields A and B in $L_T^{2,\mathrm{Div}}(\partial D)$, find two vector fields in D, E_i and H_i, and two vector fields in $\mathrm{R}^3\backslash\overline{D}$, E_e and H_e, satisfying the radiation condition (5) and such that

$$(T)\begin{cases} \operatorname{curl} E_i = i\omega\mu_i H_i & \text{in } D,\\ \operatorname{curl} H_i = -i\omega\epsilon_i E_i & \text{in } D,\\ \|E_i^*\|_{L^2(\partial D)} + \|H_i^*\|_{L^2(\partial D)} < \infty, & \\ & \\ \operatorname{curl} E_e = i\omega\mu_e H_e & \text{in } \mathrm{R}^3\backslash\overline{D},\\ \operatorname{curl} H_e = -i\omega\epsilon_e E_e & \text{in } \mathrm{R}^3\backslash\overline{D},\\ \|E_e^*\|_{L^2(\partial D)} + \|H_e^*\|_{L^2(\partial D)} < \infty, & \\ & \\ N\times E_e - N\times E_i = A & \text{on } \partial D,\\ N\times H_e - N\times H_i = B & \text{on } \partial D. \end{cases}$$

3. Uniqueness of solution. The uniqueness of solution of problem (T) is given in [12] and [15] for smooth domains and functions continuous up to the boundary. We will consider here the case of Lipschitz domains, boundary data in $L_T^{2,\mathrm{Div}}(\partial D)$, and boundary values obtained nontangentially. We will always assume that the electromagnetic parameters satisfy the constrains in (1)–(4). Additional limitations in their values will be imposed, if necessary, in the statments of the results to be proved.

Usually, the proof of uniqueness results for boundary-value problems involves integral-representation formulas and some application of the divergence theorem. The standard technique to adapt these formulas to the case of nonsmooth domains is an approximation procedure. The main tool is the following lemma from [17].

LEMMA 3.1. *Let D be a bounded Lipschitz domain. Then it is possible to construct a sequence of C^∞ domains $\Omega_j \subset D$ (or $\Omega_j \supset D$) satisfying the following properties:*

(i) *There is a sequence of Lipschitz diffeomorphisms $\Lambda_j : \partial D \to \partial\Omega_j$ such that the Lipschitz constants of Λ_j and its inverse are uniformly bounded in j. Furthermore, $\Lambda_j(Q) \in \Gamma_i(Q)$ (or $\Gamma_e(Q)$) for all j and all $Q \in \partial D$ and $\sup_{Q\in\partial D} |Q - \Lambda_j(Q)| \le C/j$;*

(ii) *There are positive functions $\rho_j : \partial D \to R_+$ bounded away from zero and infinity uniformly in j such that for any measurable set $F \subset \partial D$, $\int_F \rho_j d\sigma = \int_{\Lambda_j(F)} d\sigma_j$ and such that $\rho_j \to 1$ a.e. and in every $L^p(\partial D)$, $1 \le p < \infty$;*

(iii) *The sequence of normal vectors to Ω_j, $N_j(\Lambda_j(\cdot))$ converges a.e. and in every $L^p(\partial D)$, $1 \le p < \infty$, to N.* □

Let k be a complex number with $\operatorname{Im} k \ge 0$ and consider the fundamental solution of the Helmholtz operator $\triangle + k^2$ in R^3,

$$\Phi(X) = -\frac{e^{ik|X|}}{4\pi|X|}.$$

We will need to use the following formulae regarding solutions of the vector Helmholtz equation.

LEMMA 3.2. *Let D be a Lipschitz domain and let E be a smooth vector field in D or $R^3\backslash\overline{D}$. Assume that E,* $\operatorname{curl} E$, *and* $\operatorname{div} E$ *have nontangentially boundary values on ∂D from the inside or the outside accordingly to where E is defined. Assume also that*

$$\|E^*\|_{L^2(\partial D)} + \|(\operatorname{div} E)^*\|_{L^2(\partial D)} + \|(\operatorname{curl} E)^*\|_{L^2(\partial D)} < \infty. \tag{6}$$

The following formulas hold:

(i) *The tangential vector field $N \times E$ has a surface divergence in $L^2(\partial D)$ and*

$$\operatorname{Div}(N \times E) = -\langle N, \operatorname{curl} E\rangle . \tag{7}$$

(ii) *If E is a solution of the vector Helmholtz equation $\triangle E + k^2 E = 0$ in D, then for all $X \in D$,*

$$\begin{aligned} E(X) = &-\int_{\partial D} \operatorname{curl}_X(\Phi(X-Q)N(Q) \times E(Q))d\sigma \\ &+\int_{\partial D} \nabla_X\Phi(X-Q)\,\langle N(Q), E(Q)\rangle\, d\sigma \\ &-\int_{\partial D} \Phi(X-Q)(N(Q) \times \operatorname{curl} E(Q) - \operatorname{div} E(Q)N(Q))d\sigma \end{aligned}$$

and

$$\begin{aligned} &\int_{\partial D} \langle N(Q) \times \overline{E}(Q), \operatorname{curl} E(Q)\rangle + \operatorname{div} E(Q)\,\langle N(Q), \overline{E}(Q)\rangle\, d\sigma \\ &= \int_D |\operatorname{curl} E(X)|^2 + |\operatorname{div} E(X)|^2 - k^2|E(X)|^2\, dX. \end{aligned}$$

(iii) *If E is a solution of the vector Helmholtz equation $\triangle E + k^2 E = 0$ in $R^3\backslash\overline{D}$ that satisfies at infinity the radiation condition*

$$\operatorname{curl} E \times \frac{X}{|X|} + \operatorname{div} E\frac{X}{|X|} - ikE = o(|X|^{-1}), \qquad E = O(|X|^{-1}), \tag{8}$$

then for all $X \in R^3\backslash\overline{D}$,

$$\begin{aligned} E(X) = &\int_{\partial D} \operatorname{curl}_X\{\Phi(X-Q)N(Q) \times E(Q)\}d\sigma \\ &-\int_{\partial D} \nabla_X\Phi(X-Q)\,\langle N(Q), E(Q)\rangle\, d\sigma \\ &+\int_{\partial D} \Phi(X-Q)(N(Q) \times \operatorname{curl} E(Q) - \operatorname{div} E(Q)N(Q))d\sigma \end{aligned}$$

and

$$\lim_{r\to\infty} \left(-\int_{|X|=r} |k|^2|E(X)|^2 + |\operatorname{curl} E(X) \times N(X) + \operatorname{div} E(X)N(X)|^2\, ds_r \right.$$

$$- 2\mathrm{Im}(k) \int_{D_r} |\mathrm{curl}\, E(X)|^2 + |\mathrm{div}\, E(X)|^2 + |k|^2 |E(X)|^2 \, dX \Big)$$

$$= 2\mathrm{Im} \left(k \int_{\partial D} \left(\langle N(Q), E(Q) \times \mathrm{curl}\, \overline{E}(Q) \rangle + \mathrm{div}\, \overline{E}(Q) \, \langle N(Q), E(Q) \rangle \right) d\sigma \right),$$

where ds_r is the surface measure on the ball of radius r, $B_r(0)$, and where $D_r = R^3 \setminus \overline{D} \cap B_r(0)$.

Proof. The above formulas are well known for smooth domains. The validity of them in the case of Lipschitz domains was justified in [14] using Lemma 3.1 and a limiting argument. We shall not repeat the details here (cf. the proof of Theorem 3.4 below). □

A simple consequence of the above lemma is the following result.

LEMMA 3.3. *Let D be a Lipschitz domain. Let E be a solution of the vector Helmholtz equation in $R^3 \setminus \overline{D}$ satisfying (6), the radiation condition (8), and the inequality*

$$\mathrm{Im} \left(k \int_{\partial D} \left(\langle N(Q), E(Q) \times \mathrm{curl}\, \overline{E}(Q) \rangle + \mathrm{div}\, \overline{E}(Q) \, \langle N(Q), E(Q) \rangle \right) d\sigma \right) \geq 0.$$

If $\mathrm{Im}\, k > 0$, *then $E = 0$ in $R^3 \setminus \overline{D}$.*

Proof. If $\mathrm{Im}\, k > 0$, then from the last part of Lemma 3.2,

$$\int_{D_r} |E(X)|^2 \, dX \to 0,$$

which implies that $E = 0$. □

We can now prove a uniqueness results for solutions of the transmission problem (T). Recall that solutions of the Maxwell equations

$$\mathrm{curl}\, E = i\omega\mu H,$$

$$\mathrm{curl}\, H = -i\omega\epsilon E$$

are divergence-free solutions of the vector Helmholtz equation with wave number $k^2 = \omega^2 \epsilon\mu$. Notice also the equivalence between the radiation conditions (5) and (8).

THEOREM 3.4. *Let D be a Lipschitz domain. Assume that* $\mathrm{Im}\, k_i > 0$ *and* $\mathrm{Im}\, k_e > 0$, *and let E_i, H_i, E_e,* and H_e *be solutions of (T) with boundary data $A = B = 0$. Then $E_i = H_i = 0$ in D and $E_e = H_e = 0$ in $R^3 \setminus \overline{D}$.*

Proof. We will use a limiting argument to adapt the proof in [15, p. 282], for the case of smooth domains to the present situation. Let Ω_j be a family of domains approximating D from inside as in Lemma 3.1. Since solutions of Maxwell equations are analytic inside D we can apply the divergence theorem on each domain Ω_j. We obtain

$$\int_{\Omega_j} (i\omega\epsilon_i |E_i|^2 - i\overline{\omega\mu_i} |H_i|^2) dX = \int_{\Omega_j} \mathrm{div}\, (\overline{E}_i \times H_i) dX$$

$$= \int_{\partial\Omega_j} \langle N_j, \overline{E}_i \times H_i \rangle \, d\sigma_j,$$

and using the change of coordinates Λ_j,

$$\int_{\Omega_j} (i\omega\epsilon_i|E_i|^2 - i\overline{\omega\mu_i}|H_i|^2)dX$$

$$= \int_{\partial D} \langle N_j(\Lambda_j(Q)), \overline{E}_i(\Lambda_j(Q)) \times H_i(\Lambda_j(Q))\rangle \, \rho_j d\sigma.$$

The integrals on the left of the above equality are uniformly bounded by

$$C \int_{\partial D} |E_i^*(Q)||H_i^*(Q)|d\sigma.$$

Since we are assuming that $\|E_i^*\|_{L^2(\partial D)} + \|H_i^*\|_{L^2(\partial D)} < \infty$, we can use the properties of the approximating domains together with the dominated-convergence theorem to get

$$\int_D (i\omega\epsilon_i|E_i|^2 - i\overline{\omega\mu_i}|H_i|^2)dX = \int_{\partial D} \langle N, \overline{E}_i \times H_i\rangle \, d\sigma.$$

A similar argument in the exterior of D shows that

$$\int_{D_r} (i\omega\epsilon_i|E_e|^2 - i\overline{\omega\mu_i}|H_e|^2)dX$$

$$= \int_{X=r} \left\langle \frac{X}{r}, \overline{E}_e \times H_e \right\rangle ds - \int_{\partial D} \langle N, \overline{E}_e \times H_e\rangle \, d\sigma.$$

Adding the formulas for the interior and exterior, using the transmission conditions with $A = B = 0$ and the radiation condition at infinity, we get

$$\int_D (i\omega\epsilon_i|E_i|^2 - i\overline{\omega\mu_i}|H_i|^2)dX + \int_{D_r} (i\omega\epsilon_i|E_e|^2 - i\overline{\omega\mu_i}|H_e|^2)dX$$

$$= \int_{X=r} \left\langle \frac{X}{r}, \overline{E}_e \times H_e \right\rangle ds$$

$$= \frac{\overline{k}_e}{\overline{\omega\mu_e}} \int_{X=r} |E_e|^2 ds + o(1).$$

Now, by the constraints on the electromagnetic parameters,

$$\mathrm{Re}\,(i\omega\epsilon) \leq 0 \;\text{ and }\; \mathrm{Re}\,(\overline{i\omega\mu)} \geq 0,$$

where ϵ denotes either ϵ_i or ϵ_e and μ denotes either μ_i or μ_e. In addition,

$$\mathrm{Re}\left(\frac{k_e}{\omega\mu_e}\right) \geq 0.$$

It follows that we must have

$$\mathrm{Re}\left(\int_D (i\omega\epsilon_i|E_i|^2 - i\overline{\omega\mu_i}|H_i|^2)dX\right) = 0, \tag{9}$$

$$\lim_{r\to\infty} \operatorname{Re}\left(\int_{D_r} (i\omega\epsilon_i|E_e|^2 - i\overline{\omega\mu_i}|H_e|^2)dX\right) = 0, \tag{10}$$

and

$$\lim_{r\to\infty} \operatorname{Re}\left(\frac{k_e}{\omega\mu_e}\right)\int_{X=r} |E_e|^2 ds = 0. \tag{11}$$

Moreover, since we are assuming $\operatorname{Im} k_i > 0$, one of the parameters ω, ϵ_i, and μ_i is not a real number. Then either $\operatorname{Re}(i\omega\epsilon_i) < 0$ or $\operatorname{Re}(-i\overline{\omega\mu_i}) < 0$. From (9), one of the fields vanishes in D and so both E_i and H_i must be identically zero in D. Finally, from the transmission conditions, the tangential components of E_e and H_e on the boundary have to be zero and, since $\operatorname{Im} k_e > 0$, Lemma 3.3 implies that E_e and H_e must be identically zero in the exterior of D. □

Remark. The conditions $\operatorname{Im} k_i > 0$ and $\operatorname{Im} k_e > 0$ in the above theorem are removed in [15] for the case of smooth domains by a more elaborated argument. Nevertheless, we will still need those conditions to prove existence of solutions.

We conclude this section with another uniqueness result. As we will see in the proof of existence of solutions, the transmission problem in the next theorem can be used, in a general sense, as adjoint problem for problem (T) (cf. [6]).

THEOREM 3.5. *Let D be a Lipschitz domain in R^3. Assume that* $\operatorname{Im} k_i > 0$ *and* $\operatorname{Im} k_e > 0$. *Assume also that either*

$$\operatorname{Im}\left(k_i\overline{k}_e^{-2}\frac{\mu_e}{\mu_i}\right) \leq 0, \tag{12}$$

$$\operatorname{Im}\left(k_i\frac{\mu_e}{\mu_i}\right) \geq 0 \tag{13}$$

or

$$\frac{\epsilon_e\mu_i}{\mu_e} = \epsilon'_{0e} + \frac{i\sigma'_e}{\omega} \quad \text{with } \epsilon'_{0e} > 0 \text{ and } \sigma'_e \geq 0, \tag{14}$$

$$\frac{\mu_e\epsilon_i}{\epsilon_e} = \mu'_{0e} + \frac{i\hat{\sigma}'_e}{\omega} \quad \text{with } \mu'_{0e} > 0 \text{ and } \hat{\sigma}'_e \geq 0. \tag{15}$$

Then the homogeneous transmission problem for the vector Helmholtz equation,

$$(T')\begin{cases} \triangle E_i + k_e^2 E_i = 0 & \text{in } D,\\ \operatorname{div} E_i = 0 & \text{in } D,\\ \|E_i^*\|_{L^2(\partial D)} + \|(\operatorname{curl} E_i)^*\|_{L^2(\partial D)} < \infty, & \\ \\ \triangle E_e + k_i^2 E_e = 0 & \text{in } R^3\backslash\overline{D},\\ \operatorname{div} E_e = 0 & \text{in } R^3\backslash\overline{D},\\ \|E_e^*\|_{L^2(\partial D)} + \|(\operatorname{curl} E_e)^*\|_{L^2(\partial D)} < \infty, & \\ \\ N\times E_e - N\times E_i = 0 & \text{on } \partial D,\\ \frac{\mu_i}{k_i^2}N\times \operatorname{curl}E_e - \frac{\mu_e}{k_e^2}N\times\operatorname{curl}E_i = 0 & \text{on } \partial D, \end{cases}$$

where E_e satisfies the radiation condition (8) with $k = k_i$, has the unique solution $E_i = 0$ in D and $E_e = 0$ in $R^3\backslash\overline{D}$.

Proof. Assume that conditions (12) and (13) are satisfied. Let E_i and E_e be solutions of problem (T'). Using the transmission conditions and the divergence theorem (whose used can be justified again via Lemma 3.1 and the boundedness of nontangential maximal functions), we get

$$\begin{aligned}\int_{\partial D} \langle N, E_e \times \operatorname{curl} \overline{E}_e \rangle \, d\sigma &= \int_{\partial D} \left\langle N, E_i \times \overline{\left(\frac{k_i^2 \mu_e}{k_e^2 \mu_i}\right)} \operatorname{curl} \overline{E}_i \right\rangle d\sigma \\ &= \int_D \left(\overline{\left(\frac{k_i^2 \mu_e}{k_e^2 \mu_i}\right)} |\operatorname{curl} E_i|^2 - \overline{\left(\frac{k_i^2 \mu_e}{\mu_i}\right)} |E_i|^2 \right) dX.\end{aligned}$$

Now using the constraints on the electromagnetic parameters, we see that

$$\operatorname{Im} \left(k_i \int_{\partial D} \langle N, E_e \times \operatorname{curl} \overline{E}_e \rangle \, d\sigma \right) \geq 0.$$

Since $\operatorname{div} E_e = 0$, Lemma 3.3 implies that $E_e = 0$. Again using the transmission conditions and the representation formula in Lemma 3.2 for the interior of D, we also obtain that $E_i = 0$.

Assume now that conditions (14) and (15) are satisfied. Let E_i and E_e again be solutions of the problem (T'). Then it follows that E_i, $H_i = 1/i\omega\mu_i' \operatorname{curl} E_i$ and E_e, $H_e = 1/i\omega\mu_e' \operatorname{curl} E_e$ are solutions of the homogeneous version of problem (T) with electromagnetic parameters $\epsilon_i' = \epsilon_e$ and $\mu_i' = \mu_e$ in the interior and $\epsilon_e' = \epsilon_e \mu_i / \mu_e$ and $\mu_e' = \mu_e \epsilon_i / \epsilon_e$ in the exterior. By Theorem 3.4, E_i and E_e must be zero. □

Remark. The conditions on the electromagnetic parameters in the above theorem look very technical because we have stated the result in great generality. If some of the parameters are real valued, these conditions become much simpler. See Theorem 5.2 below.

4. Boundary integral operators. We recall some properties about the layer-potential operators associated with the Helmholtz and Maxwell equations. The results are well known for smooth domains; see, e.g., [4]. For nonsmooth domains, we refer for proofs and details to [1] and [16] for the case of the scalar Helmholtz equation and to [14] for the vector-valued case.

Let D be a Lipschitz domain and let f be a function in $L^2(\partial D)$. The single and double acoustic layer potentials are given by

$$\mathcal{S}f(X) = \int_{\partial D} \Phi(X - Q) f(Q) d\sigma(Q), \qquad X \in \mathrm{R}^3,$$

and

$$\mathcal{D}f(X) = \int_{\partial D} \partial_{N_Q} \Phi(X - Q) f(Q) d\sigma(Q), \qquad X \in \mathrm{R}^3 \setminus \partial D.$$

Both $\mathcal{S}f$ and $\mathcal{D}f$ solve the Helmholtz equation in $\mathrm{R}^3 \setminus \partial D$ and, as a consequence of the results in [3], they satisfy

$$\|(\mathcal{S}f)^*\|_{L^2(\partial D)} + \|(\nabla \mathcal{S}f)^*\|_{L^2(\partial D)} + \|(\mathcal{D}f)^*\|_{L^p(\partial D)} \leq C \|f\|_{L^2(\partial D)}.$$

The trace values of $\mathcal{S}f$ are given by

$$\lim_{\substack{X \longrightarrow P \\ X \in \Gamma_i(P)}} \mathcal{S}f(X) = \lim_{\substack{X \longrightarrow P \\ X \in \Gamma_e(P)}} \mathcal{S}f(X) = \mathcal{S}f(P), \qquad P \in \partial D,$$

where

$$\mathcal{S}f(P) = -\frac{1}{4\pi}\int_{\partial D}\frac{e^{ik|Q-P|}}{|Q-P|}f(Q)\,d\sigma(Q), \qquad P \in \partial D.$$

The function $\mathcal{D}f$ has a jump discontinuity given by

$$\lim_{\substack{X\longrightarrow P\\ X\in\Gamma_i(P)}} \mathcal{D}f(X) = \left(\frac{1}{2}I + K\right)f(P), \qquad P \in \partial D,$$

$$\lim_{\substack{X\longrightarrow P\\ X\in\Gamma_e(P)}} \mathcal{D}f(X) = \left(-\frac{1}{2}I + K\right)f(P), \qquad P \in \partial D.$$

where

$$Kf(P) = \frac{1}{4\pi}\text{p.v.}\int_{\partial D}\frac{\langle N(Q), Q-P\rangle}{|Q-P|^3}e^{ik|Q-P|}(1-ik|Q-P|)f(Q)\,d\sigma(Q).$$

The normal derivative of the single-layer potential satisfies

$$\lim_{\substack{X\longrightarrow P\\ X\in\Gamma_i(P)}} \langle N(P), \nabla\mathcal{S}f(X)\rangle = \left(-\frac{1}{2}I + K^*\right)f(P)$$

and

$$\lim_{\substack{X\longrightarrow P\\ X\in\Gamma_e(P)}} \langle N(P), \nabla\mathcal{S}f(X)\rangle = \left(\frac{1}{2}I + K^*\right)f(P),$$

where K^* is the transpose operator of K. On the other hand, the tangential component of $\nabla\mathcal{S}f$ does not jump.

For the rest of the section, we will assume that the imaginary part of the wave number k is positive. This condition guarantees the invertibility results in the next lemma (see [1], [16]).

LEMMA 4.1. *Let D be a Lipschitz domain in R^3. Then the following hold:*

(i) *$S : L^2(\partial D) \longrightarrow L^2(\partial D)$ is compact.*

(ii) *$S : L^2(\partial D) \longrightarrow L^{2,1}(\partial D)$ is invertible.*

(iii) *$\pm\frac{1}{2}I + K : L^2(\partial D) \longrightarrow L^2(\partial D)$ are invertible.*

(iv) *$\pm\frac{1}{2}I + K : L^{2,1}(\partial D) \longrightarrow L^{2,1}(\partial D)$ are invertible.*

(v) *If ∂D is actually of class C^1, then the operator K is compact in $L^2(\partial D)$.* □

The action of the single and double layer-potential operators on vector fields is defined componentwise. In addition, the traces of the divergence and curl of the single-layer potential of a vector field A define bounded operators in $L^2(\partial D)$, and their values are given by

$$\lim_{\substack{X\to P\\ X\in\Gamma_i(P)}} \operatorname{div}\mathcal{S}A(X) = -\frac{1}{2}\langle N, A\rangle(P) + \text{p.v.}\int_{\partial D}\operatorname{div}_P(\Phi(P-Q)A(Q))d\sigma(Q),$$

$$\lim_{\substack{X\to P\\ X\in\Gamma_e(P)}} \operatorname{div}\mathcal{S}A(X) = \frac{1}{2}\langle N, A\rangle(P) + \text{p.v.}\int_{\partial D}\operatorname{div}_P(\Phi(P-Q)A(Q))d\sigma(Q),$$

and

$$\lim_{\substack{X\to P\\ X\in\Gamma_i(P)}} \operatorname{curl}\mathcal{S}A(X) = -\frac{1}{2}(N\times A)(P) + \text{p.v.}\int_{\partial D} \operatorname{curl}_P\left(\Phi(P-Q)A(Q)\right)d\sigma(Q),$$

$$\lim_{\substack{X\to P\\ X\in\Gamma_e(P)}} \operatorname{curl}\mathcal{S}A(X) = \frac{1}{2}(N\times A)(P) + \text{p.v.}\int_{\partial D} \operatorname{curl}_P\left(\Phi(P-Q)A(Q)\right)d\sigma(Q).$$

We also have

$$\|(\operatorname{div}\mathcal{S}A)^*\|_{L^2(\partial D)} + \|(\operatorname{curl}\mathcal{S}A)^*\|_{L^2(\partial D)} \le C\|A\|_{L^2(\partial D)}.$$

The function $\operatorname{curl}\mathcal{S}A$ satisfies the vector Helmholtz equation outside ∂D as well as the radiation condition (8) at infinity. In addition, the tangential component of the trace of the curl of the single-layer potential is given almost everywhere in ∂D by

$$\lim_{\substack{X\longrightarrow P\\ X\in\Gamma_i(P)}} N(P)\times\operatorname{curl}\mathcal{S}A(X) = \left(\frac{1}{2}I + M\right)A(P)$$

and

$$\lim_{\substack{X\longrightarrow P\\ X\in\Gamma_e(P)}} N(P)\times\operatorname{curl}\mathcal{S}A(X) = \left(-\frac{1}{2}I + M\right)A(P),$$

where MA is the tangential vector field defined by

$$MA(P) = \text{p.v.}\int_{\partial D} N(P)\times\operatorname{curl}_P\left(\Phi(P-Q)A(Q)\right)d\sigma(Q).$$

We recall from [13] the following result.

LEMMA 4.2. *Let D be a Lipschitz domain in R^3. Then the operator M maps $L^2_T(\partial D)$ into itself and $L^{2,\mathrm{Div}}_T(\partial D)$ into itself. Moreover, if D is actually C^1, then M is compact on both spaces.* □

In order to study the double curl of the single-layer potential, we need another important result obtained in [14].

LEMMA 4.3. *Let D be Lipschitz domain. A vector field A in $L^2_T(\partial D)$ has a surface divergence in $L^2(\partial D)$ if and only if $\|(\nabla(\operatorname{div}\mathcal{S}A))^*\|_{L^2(\partial D)} < +\infty$. In such a case, $\operatorname{div}\mathcal{S}A = \mathcal{S}(\operatorname{Div}A)$.* □

As a consequence of this last result, we can now prove the following.

LEMMA 4.4. *Let D be a Lipschitz domain. Let A be a vector field in $L^{2,\mathrm{Div}}_T(\partial D)$. Then $\|(\operatorname{curl}\operatorname{curl}\mathcal{S}A)^*\|_{L^2(\partial D)} < +\infty$ and*

$$\lim_{X\to P} N(P)\times\operatorname{curl}\operatorname{curl}\mathcal{S}A(X) = \left(N\times(k^2\mathcal{S}A + \nabla\mathcal{S}(\operatorname{Div}A))\right)(P)$$

nontangentially, both from the inside and outside of D. Moreover, if we define on ∂D the operator

$$LA(P) = \left(N\times(k^2\mathcal{S}A + \nabla\mathcal{S}(\operatorname{Div}A))\right)(P),$$

then L maps $L^{2,\mathrm{Div}}_T(\partial D)$ into itself.

Proof. Let A be a vector field in $L_T^{2,\mathrm{Div}}(\partial D)$. Using the identity

$$\mathrm{curl}\,\mathrm{curl} = -\triangle + \nabla\mathrm{div}$$

and Lemma 4.3, we see that

$$\mathrm{curl}\,\mathrm{curl}\,\mathcal{S}A = k^2\mathcal{S}A + \nabla\mathcal{S}(\mathrm{Div}\,A),$$

which implies the boundedness of the nontangential maximal function and the claimed boundary values (notice that $\mathcal{S}$ and the tangential component of $\nabla\mathcal{S}$ do not have jumps). Clearly, the resulting boundary operator L maps $L_T^{2,\mathrm{Div}}(\partial D)$ into $L_T^2(\partial D)$. If we now apply (7) to the vector field

$$E(X) = \mathrm{curl}\,\mathrm{curl}\,\mathcal{S}A(X),$$

we obtain that

$$\mathrm{Div}\,(LA) = \mathrm{Div}\,(N\times E) = -\langle N, \mathrm{curl}\,\mathrm{curl}\,\mathrm{curl}\,\mathcal{S}A\rangle = -\left\langle N, k^2\mathrm{curl}\,\mathcal{S}A\right\rangle,$$

and, therefore,

$$\begin{aligned}\|LA\|_{L_T^{2,\mathrm{Div}}(\partial D)} &= \|LA\|_{L_T^2(\partial D)} + \|\mathrm{Div}\,(LA)\|_{L^2(\partial D)}\\ &\le \|N\times(k^2\mathcal{S}A+\nabla\mathcal{S}(\mathrm{Div}\,A))\|_{L_T^2(\partial D)} + \|\langle N, k^2\mathrm{curl}\,\mathcal{S}A\rangle\|_{L^2(\partial D)}\\ &\le C(\|\mathcal{S}A\|_{L^2(\partial D)} + \|\nabla\mathcal{S}(\mathrm{Div}\,A)\|_{L^2(\partial D)} + \|\mathrm{curl}\,\mathcal{S}A\|_{L^2(\partial D)})\\ &\le C(\|A\|_{L_T^2(\partial D)} + \|\mathrm{Div}\,A)\|_{L^2(\partial D)}),\end{aligned}$$

which concludes the proof. □

We need to consider the potential-theoretic versions of some of the layer-potential operators already described. Let S_0, M_0, and L_0 be defined using the fundamental solution of the Laplace operator $\triangle$ in R^3,

$$\Phi_0(X) = -\frac{1}{4\pi|X|}.$$

The boundedness properties of the operators S_0, M_0, and L_0 are the same as those of S, M, and L. Moreover, we have the following.

LEMMA 4.5. *Let D be a Lipschitz domain in R^3. Then,*

(i) $M - M_0 : L_T^2(\partial D) \longrightarrow L_T^2(\partial D)$ *is compact.*

(ii) $M - M_0 : L_T^2(\partial D) \longrightarrow L_T^{2,\mathrm{Div}}(\partial D)$ *is bounded.*

(iii) $L - L_0 : L_T^2(\partial D) \longrightarrow L_T^2(\partial D)$ *is compact.*

(iv) $L - L_0 : L_T^2(\partial D) \longrightarrow L_T^{2,\mathrm{Div}}(\partial D)$ *is bounded.*

Proof. A straightforward computation shows that the differences of partial derivatives

$$\partial_i\Phi(P-Q) - \partial_i\Phi_0(P-Q)$$

and

$$\partial_i\partial_j\Phi(P-Q) - \partial_i\partial_j\Phi_0(P-Q)$$

have locally integrable singularities on ∂D. From this easily follows that $M - M_0$ and $L - L_0$ are compact operators in $L_T^2(\partial D)$ (cf. [16]). To show that these operators map $L_T^2(\partial D)$ into $L_T^{2,\mathrm{Div}}(\partial D)$, we notice that

$$\begin{aligned}\mathrm{Div}\,(M-M_0)A &= \mathrm{Div}\,(N\times\mathrm{curl}\,(\mathcal{S}-\mathcal{S}_0)A)\\ &= -\langle N, \mathrm{curl}\,\mathrm{curl}\,(\mathcal{S}-\mathcal{S}_0)A\rangle\\ &= -\left\langle N, k^2\mathrm{curl}\,\mathcal{S}A + \nabla\,\mathrm{div}((\mathcal{S}-\mathcal{S}_0)A)\right\rangle,\end{aligned}$$

which defines an operator with a kernel with a locally integrable singularity and bounded in $L^2(\partial D)$. Similarly,

$$\begin{aligned}\operatorname{Div}(L-L_0)A &= \operatorname{Div}(N\times \operatorname{curl}\operatorname{curl}(\mathcal{S}-\mathcal{S}_0)A)\\ &= -\langle N, \operatorname{curl}\operatorname{curl}\operatorname{curl}(\mathcal{S}-\mathcal{S}_0)A\rangle\\ &= -\left\langle N, k^2 \operatorname{curl}\mathcal{S}A\right\rangle,\end{aligned}$$

again producing a bounded operator in $L^2(\partial D)$. □

We conclude this section with a simple result about the spectrum of M_0 in C^1 domains.

LEMMA 4.6. *Let D be a C^1 domain in R^3. Then, for any complex number λ outside the interval $[-\frac{1}{2},\frac{1}{2}]$, the operator $\lambda I + M_0$ is invertible in $L^2_T(\partial D)$ and in $L_T^{2,\mathrm{Div}}(\partial D)$.*

Proof. Since by the results in [14] the operator M_0 still is L^2-compact in C^1 domains, it is enough to prove that $\lambda I + M_0$ is injective. This is done in [4, pp. 155–157] in the case of C^2 domains. Given the boundedness and invertibility properties of the layer-potential operators discussed in this section, the same proof extends without modification to the case of C^1 domains. Finally, observe that since M_0 is also compact in $L_T^{2,\mathrm{Div}}(\partial D)$, its spectrum in $L^2_T(\partial D)$ and in $L_T^{2,\mathrm{Div}}(\partial D)$ is the same. □

Remark. In the case of Lipschitz domains, the operators M and M_0 may not be compact in $L^2_T(\partial D)$. The invertibility of $\lambda I + M$ or $\lambda I + M_0$ can no longer be handled via Fredholm theory. The usual substitute technique to prove invertibility results in this kind of situation involves the use of Rellich-type identities (see, e.g., [17] and [7]). Such techniques were used in [8] and [10] to study the spectrum in $L^2(\partial D)$ of the double-layer potential for the Laplacian $\lambda I + K_0$. The spectral properties of K_0 in $L^{2,1}(\partial D)$ were studied in [16]. The spectral properties of M_0 in $L^2_T(\partial D)$ in the case of Lipschitz domains remain unknown, but from the results in [13], it follows that if $\lambda I + M_0$ is in vertible in $L^2_T(\partial D)$ for some λ, then it is also invertible in $L_T^{2,\mathrm{Div}}(\partial D)$. This missing information about the spectrum of M_0 in $L^2_T(\partial D)$ is the only additional result that would be necessary to extend Theorem 5.1 in the next section to the case of Lipschitz domains.

5. Existence of solutions.

We now present the existence of solutions to problem (T) using a particular boundary integral representation.

THEOREM 5.1. *Let D be a C^1 domain in R^3. Assume that the electromagnetic parameters satisfy the conditions in Theorem 3.5. Assume also that $\frac{\mu_e+\mu_i}{\mu_e-\mu_i}$ and $\frac{\epsilon_e+\epsilon_i}{\epsilon_e-\epsilon_i}$ are not real numbers in the interval $[-1,1]$. Then the transmission problem (T) has a unique solution for any A and B in $L_T^{2,\mathrm{Div}}(\partial D)$.*

Proof. In view of Theorem 3.4, we only need to show existence of solution. Let U and V be vector fields in $L_T^{2,\mathrm{Div}}(\partial D)$ and consider the ansatz

$$E_e(X) = \mu_e \operatorname{curl}\mathcal{S}_e U(X) + \operatorname{curl}\operatorname{curl}\mathcal{S}_e V(X) \quad \text{in } \mathrm{R}^3\backslash\overline{D},$$

$$E_i(X) = \mu_i \operatorname{curl}\mathcal{S}_i U(X) + \operatorname{curl}\operatorname{curl}\mathcal{S}_i V(X) \quad \text{in } D,$$

and

$$H_e(X) = \frac{1}{i\mu_e\omega}\operatorname{curl} E_e(X),$$

$$H_i(X) = \frac{1}{i\mu_e\omega}\operatorname{curl} E_i(X),$$

where $\mathcal{S}_e$ and $\mathcal{S}_i$ denote the single layer-potential operators defined using the wave numbers k_e and k_i. By the results of the previous section, for any U and V in $L_T^{2,\mathrm{Div}}(\partial D)$, these vector fields solve the Maxwell equations and satisfy the radiation condition at infinity. Then, to solve problem (T), it is enough to show that given A and B in $L_T^{2,\mathrm{Div}}(\partial D)$, we can find U and V such that the above electric fields satisfy on ∂D

$$N \times E_e - N \times E_i = A,$$

$$\frac{1}{\mu_e} N \times \operatorname{curl} E_e - \frac{1}{\mu_i} N \times \operatorname{curl} E_i = B.$$

That is, we need to solve the system

$$\mu_e \left(-\frac{1}{2} I + M_e \right) U + L_e V - \mu_i \left(\frac{1}{2} I + M_i \right) U - L_i V = A,$$

$$L_e U + \frac{k_e^2}{\mu_e} \left(-\frac{1}{2} I + M_e \right) V - L_i U - \frac{k_i^2}{\mu_i} \left(\frac{1}{2} I + M_i \right) V = B.$$

We rewrite this as

$$\begin{pmatrix} -\frac{\mu_e+\mu_i}{2} I + \mu_e M_e - \mu_i M_i & L_e - L_i \\ L_e - L_i & -\left(\frac{k_e^2}{2\mu_e} + \frac{k_i^2}{2\mu_i} \right) I + \frac{k_e^2}{\mu_e} M_e - \frac{k_i^2}{\mu_i} M_i \end{pmatrix} \cdot \begin{pmatrix} U \\ V \end{pmatrix}$$

$$= \begin{pmatrix} A \\ B \end{pmatrix}.$$

Notice that the above system, originally defined in the the product space $L_T^{2,\mathrm{Div}}(\partial D) \times L_T^{2,\mathrm{Div}}(\partial D)$, makes sense in the space $L_T^2(\partial D) \times L_T^2(\partial D)$. Now, we observe that if M_0 is the potential-theoretic version of M, the above matrix of operators can be decompose as the sum of two matrices, $\mathrm{W}_1 + \mathrm{W}_2$, where

$$\mathrm{W}_1 = \begin{pmatrix} -\frac{\mu_e+\mu_i}{2} I + (\mu_e - \mu_i) M_0 & 0 \\ 0 & -\left(\frac{k_e^2}{2\mu_e} + \frac{k_i^2}{2\mu_i} \right) I + \left(\frac{k_e^2}{\mu_e} - \frac{k_i^2}{\mu_i} \right) M_0 \end{pmatrix}$$

and

$$\mathrm{W}_2 = \begin{pmatrix} \frac{\mu_e}{2}(M_e - M_0) + \frac{\mu_i}{2}(M_0 - M_i) & L_e - L_i \\ L_e - L_i & \frac{k_i^2}{\mu_e}(M_e - M_0) + \frac{k_i^2}{\mu_i}(M_0 - M_i) \end{pmatrix}.$$

Since we are assuming that

$$\frac{\mu_e + \mu_i}{\mu_e - \mu_i} \quad \text{and} \quad \frac{\frac{k_e^2}{\mu_e} + \frac{k_i^2}{\mu_i}}{\frac{k_e^2}{\mu_e} - \frac{k_i^2}{\mu_i}} = \frac{\epsilon_e + \epsilon_i}{\epsilon_e - \epsilon_i}$$

are not in the interval $[-1, 1]$, Lemma 4.6 implies that the matrix W_1 is an invertible operator in both $L_T^2(\partial D) \times L_T^2(\partial D)$ and $L_T^{2,\mathrm{Div}}(\partial D) \times L_T^{2,\mathrm{Div}}(\partial D)$. On the other hand, by Lemma 4.5, the matrix W_2 is compact in $L_T^2(\partial D) \times L_T^2(\partial D)$ and maps this space into $L_T^{2,\mathrm{Div}}(\partial D) \times L_T^{2,\mathrm{Div}}(\partial D)$. From this follows that the matrix $\mathrm{W}_1 + \mathrm{W}_2$ has index zero in $L_T^2(\partial D) \times L_T^2(\partial D)$. It also follows that if U and V are solutions of the system of boundary integral equations in $L_T^2(\partial D) \times L_T^2(\partial D)$, then U and V are in $L_T^{2,\mathrm{Div}}(\partial D)$ if and only if A and B are in $L_T^{2,\mathrm{Div}}(\partial D)$. In particular, the null space of the matrix $\mathrm{W}_1 + \mathrm{W}_2$ is the same in both spaces. If we can show that this matrix of operators is one-to-one in $L_T^2(\partial D) \times L_T^2(\partial D)$, we will have by the previous observation that it is invertible in $L_T^2(\partial D) \times L_T^2(\partial D)$ and also in $L_T^{2,\mathrm{Div}}(\partial D) \times L_T^{2,\mathrm{Div}}(\partial D)$. This will conclude the proof of the theorem.

Assume that U and V are solutions of the system with $A = B = 0$. Since U and V must be in $L_T^{2,\mathrm{Div}}(\partial D)$, E_i, H_i and E_e, H_e are solutions of the homogeneous version of problem (T), and by Theorem 3.4, they must be identically zero. In particular, on the boundary of the domain,

$$N \times E_i = N \times E_e = N \times \operatorname{curl} E_i = N \times \operatorname{curl} E_e = 0.$$

Now consider the new vector fields

$$E_e'(X) = -\operatorname{curl} \mathcal{S}_i U(X) - \frac{1}{\mu_i} \operatorname{curl}\, \operatorname{curl} \mathcal{S}_i V(X) \quad \text{in } \mathrm{R}^3 \backslash \overline{D},$$

$$E_i'(X) = \operatorname{curl} \mathcal{S}_e U(X) + \frac{1}{\mu_e} \operatorname{curl}\, \operatorname{curl} \mathcal{S}_e V(X) \quad \text{in } D.$$

Going to the boundary, we obtain the trace values,

$$N \times E_e' = \left(\frac{1}{2} I - M_i\right) U - \frac{1}{\mu_i} L_i V,$$

$$N \times E_i' = \left(\frac{1}{2} I + M_e\right) U + \frac{1}{\mu_e} L_e V,$$

and

$$N \times \operatorname{curl} E_e' = -L_i U - \frac{k_i^2}{\mu_i} \left(-\frac{1}{2} I + M_i\right) V,$$

$$N \times \operatorname{curl} E_i' = L_e U + \frac{k_e^2}{\mu_e} \left(\frac{1}{2} I + M_e\right) V.$$

It follows that on ∂D,

$$N \times E_e' - N \times E_i' = -\frac{1}{\mu_i} N \times E_i - \frac{1}{\mu_e} N \times E_e = 0,$$

and also

$$\frac{\mu_i}{k_i^2} N \times \operatorname{curl} E_e' - \frac{\mu_e}{k_e^2} N \times \operatorname{curl} E_i' = -\frac{1}{k_i^2} N \times \operatorname{curl} E_i - \frac{1}{k_e^2} N \times \operatorname{curl} E_e = 0.$$

Therefore, E_i' and E_e' are solutions of problem (T') and, by Theorem 3.5, they must be identically zero too. In particular,

$$N \times E_i' = N \times E_e' = N \times \operatorname{curl} E_i' = N \times \operatorname{curl} E_e' = 0$$

using the trace results of the previous section. Finally,

$$U = N \times E_i + \mu_i N \times E_e' = 0,$$

and

$$-V = N \times E_e - \mu_e N \times E_i' = 0. \qquad \square$$

In physical applications, the parameters μ_i and μ_e are usually assumed to be real numbers. In such a case, the conditions on the electromagnetic parameters take the following simpler form.

THEOREM 5.2. *Let D be a C^1 domain in R^3. Let* $\operatorname{Im} k_i > 0$ *and* $\operatorname{Im} k_e > 0$*, and assume that μ_i, μ_e, and ω are positive numbers. Then the transmission problem (T) has a unique solution for any A and B in $L_T^{2,\mathrm{Div}}(\partial D)$.*

Proof. First, observe that if μ_i and μ_e are positive numbers, then

$$\left|\frac{\mu_e + \mu_i}{\mu_e - \mu_i}\right| > 1,$$

and if ω is a positive number, then the conditions $\operatorname{Im} k_i > 0$ and $\operatorname{Im} k_e > 0$ imply that

$$\left|\frac{\epsilon_e + \epsilon_i}{\epsilon_e - \epsilon_i}\right| > 1.$$

Also, conditions (13) and (14) are trivially satisfied. It follows that to use Theorem 3.5, we need only to check that either

$$\operatorname{Im}(k_i \overline{k}_e^{2}) \leq 0 \tag{16}$$

or

$$\operatorname{Im}\left(\frac{\epsilon_i}{\epsilon_e}\right) > 0 \text{ and } \operatorname{Re}\left(\frac{\epsilon_i}{\epsilon_e}\right) > 0. \tag{17}$$

By writing

$$\epsilon_i = |\epsilon_i| \exp\left(i \arctan \frac{\sigma_i}{\omega \epsilon_{0i}}\right),$$

$$\epsilon_e = |\epsilon_e| \exp\left(i \arctan \frac{\sigma_e}{\omega \epsilon_{0e}}\right),$$

we see that (16) is equivalent to

$$\frac{\sigma_i}{2\epsilon_{0i}} \leq \frac{\sigma_e}{\epsilon_{0e}}. \tag{18}$$

On the other hand, for positive ω, the real part of ϵ_i/ϵ_e is always positive, and a computation shows that (17) becomes equivalent to

$$\frac{\sigma_i}{\epsilon_{0i}} > \frac{\sigma_e}{\epsilon_{0e}}. \tag{19}$$

Obviously, either (18) or (19) is satisfied, which concludes the proof. $\square$

Finally, the proof of Theorem 5.1 shows that for a Lipschitz domain D the following result holds.

THEOREM 5.3. *Let D be a Lipschitz domain in R^3. Assume that the electromagnetic parameters satisfy the conditions in Theorem 3.5. Assume also that $\frac{\mu_e+\mu_i}{\mu_e-\mu_i}$ and $\frac{\epsilon_e+\epsilon_i}{\epsilon_e-\epsilon_i}$ are not in the spectrum of $2M_0$ as an operator in $L_T^2(\partial D)$. Then the transmission problem (T) has a unique solution for any A and B in $L_T^{2,\mathrm{Div}}(\partial D)$.* $\square$

Acknowledgment. The author would like to thank Grant Welland for many useful conversations in the subject.

REFERENCES

[1] R. Brown and Z. Shen, *The initial-Dirichlet problem for a fourth-order parabolic equation in Lipschitz cylinders*, Indiana Univ. Math. J., 39 (1990), pp. 1313–1353.

[2] A. Calderón, *Cauchy integral on Lipschitz curves and related operators*, Proc. Nat. Acad. Sci. U.S.A., 74 (1977), pp. 1324–1327.

[3] R. Coifman, A. McIntosch, and Y. Meyer, *L'intégrale de Cauchy définit un opérateur borné sur L^2 pour les courbes Lipschitiziennes*, Ann. of Math., 116 (1982), pp. 361–387.

[4] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.

[5] M. Costabel and E. Stephan, *Strongly elliptic boundary integral equations for electromagnetic transmission problems*, Proc. Roy. Soc. Edinburgh Sect. A, 109 (1988), pp. 271–296.

[6] ———, *Integral equations for transmission problems in linear elasticity*, J. Integral Equations Appl., 2 (1990), pp. 211–223.

[7] B. Dahlberg, C. Kenig, and G. Verchota, *Boundary value problems for the systems of elastostatics in Lipschitz domains*, Duke Math. J., 57 (1988), pp. 795–818.

[8] L. Escauriaza, E. Fabes, and G. Verchota, *On a regularity theorem for weak solutions to transmission problems with internal Lipschitz boundaries*, Proc. Amer. Math. Soc., 115 (1992), pp. 1069–1076.

[9] E. Fabes, M. Jodeit, and N. Rivière, *Potential techniques for boundary value problems on C^1 domains*, Acta Math., 141 (1978), pp. 165–186.

[10] E. Fabes, M. Sand, and J. Seo, *The spectral radius of the classical layer potentials on convex domains*, in Partial Differential Equations with Minimal Smoothness and Applications, Springer-Verlag, New York, 1992, pp. 129–137.

[11] D. Jerison and C. Kenig, *The Neumann problem on Lipschitz domains*, Bull. Amer. Math. Soc., 4 (1981), pp. 203–207.

[12] W. Kupradse, *Randewertaufgaben der Schwngungstheorie und Integralgleichungen*, Deutscher Verlag der Wissenschafen, Berlin, 1956.

[13] M. Mitrea, R. Torres, and G. Welland, *Regularity and approximation results for the Maxwell problem in C^1 and Lipschitz domains*, in Clifford Algebras in Analysis and Related Topics, CRC Press, Boca Raton, 1995, pp. 297–308.

[14] ———, *Layer potential techniques in electromagnetism*, preprint.

[15] C. Müller, *Foundations of the Mathematical Theory of Electromagnetic Waves*, Springer-Verlag, Berlin, 1969.

[16] R. Torres and G. Welland, *The Helmholtz equation and transmission problems with Lipschitz interfaces*, Indiana Univ. Math. J., 42 (1993), pp. 1457–1485.

[17] G. Verchota, *Layer potentials and boundary value problems for Laplace's equation in Lipschitz domains*, J. Funct. Anal., 59 (1984), pp. 572–611.

[18] ———, *The Dirichlet problem for the biharmonic equation in C^1 domains*, Indiana Univ. Math. J., 36 (1987), pp. 867–895.

[19] T. von Petersdorff, *Boundary integral equations for mixed Dirichlet, Neumann and transmission problems*, Math. Methods Appl. Sci., 11 (1989), pp. 185–213.

[20] P. Wilde, *Transmission problems for the vector Helmholtz equation*, Proc. Roy. Soc. Edinburgh Sect. A, 105 (1987), pp. 61–76.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1424–1447, September 1996

013

DISSIPATION IN HAMILTONIAN SYSTEMS: DECAYING CNOIDAL WAVES*

G. DERKS† AND E. VAN GROESEN‡

Abstract. The uniformly damped Korteweg–de Vries (KdV) equation with periodic boundary conditions can be viewed as a Hamiltonian system with dissipation added. The KdV equation is the Hamiltonian part and it has a two-dimensional family of relative equilibria. These relative equilibria are space-periodic soliton-like waves, known as cnoidal waves.

Solutions of the dissipative system, starting near a cnoidal wave, are approximated with a long curve on the family of cnoidal waves. This approximation curve consists of a quasi-static succession of cnoidal waves. The approximation process is sharp in the sense that as a solution tends to zero as $t \to \infty$, the difference between the solution and the approximation tends to zero in a norm that sharply picks out their difference in shape. More explicitly, the difference in shape between a solution and a quasi-static cnoidal-wave approximation is of the order of the damping rate times the norm of the cnoidal-wave at each instant.

Key words. perturbed KdV equation, cnoidal waves, asymptotic behavior

AMS subject classifications. 35B20, 35Q53, 76B25

1. Introduction.

Consider the uniformly damped one-dimensional Korteweg–de Vries (KdV) equation with periodic boundary conditions

$$\begin{aligned} &u_t = -\partial_x \left[u_{xx} + u^2 \right] - \varepsilon\, u, && t > 0,\ \ x \in (0, 2\pi), \\ &u(0,t) = u(2\,\pi,t), \quad u_x(0,t) = u_x(2\,\pi,t), && t \geq 0. \end{aligned} \tag{1}$$

In this equation ε is a small parameter that gives the strength of the damping and the subscripts denote differentiation with respect to the given variable. Furthermore, we assume that the function $u(x,t)$ has mean value zero for all time:

$$\int_0^{2\pi} u(x,t)\, dx = 0, \qquad t \geq 0.$$

In [5] and [7], it is shown that the initial value problem of (1) with $\varepsilon = 0$ is well posed in H^s, $s \geq 1$. It is easy to see that if $\varepsilon \neq 0$, this property remains; see, e.g., [14].

If $\varepsilon = 0$, there is no damping present and the resulting equation is the KdV equation, which can be regarded as a Hamiltonian system with the Hamiltonian

$$H(u) = \int_0^{2\pi} \left[\frac{1}{2} u_x^2 - \frac{1}{3} u^3 \right] dx$$

and with the operator ∂_x as the structure map. The KdV equation was originally derived in 1895 as a model for planar, unidirectional waves propagating in shallow wa-

* Received by the editors August 6, 1993; accepted for publication (in revised form) March 13, 1995. This work is part of the research program of the Stichting voor Fundamenteel Onderzoek der Materie (FOM) and the Stichting Mathematisch Centrum (SMC), which are financially supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).

† Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Current address: Department of Mathematical and Computing Sciences, University of Surrey, Guildford, Surrey GV2 5XH, United Kingdom.

‡ Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

ter [21]. Over the last thirty years, the KdV equation has appeared as a model equation for many other physical situations that feature wave motion wherein nonlinearity and dispersion are comparable. For a review on the KdV equation, see [25] and [27].

The KdV equation is translation invariant. This invariance gives another first integral in the system besides the Hamiltonian, namely the L^2-norm of the solutions

$$I(u) = \frac{1}{2}\int_0^{2\pi} u^2\,dx.$$

(Moreover, the KdV equation is completely integrable, but here we use only the translation invariance.) The Hamiltonian I-flow is the translation operator (see [26])

$$(\Phi^I_\varphi(u))(x) = u(x+\varphi), \qquad x \in [0, 2\pi],\ \varphi \in [0, 2\pi].$$

The tangent vector to this flow is the Hamiltonian I-vector field, denoted by

$$X_I(u) = \partial_x I'(u) = \partial_x u.$$

Profiles of traveling-wave solutions of a translation-invariant Hamiltonian system can be found as critical points of the Hamiltonian for fixed values of I. In other words, they are relative equilibria (see [1]), and the family of all traveling-wave profiles is called the manifold of relative equilibria (MRE). In case of the periodic KdV equation, the relative equilibria are solitary-wave solutions, the so-called cnoidal waves. The cnoidal waves with minimal period 2π form a two-dimensional family which can be parameterized with the value of the integral I (a quantity related to the amplitude of the cnoidal wave) and the "position" of the cnoidal wave. The MRE consists of traveling-wave profiles, but for simplicity, the two-dimensional manifold consisting of the relative equilibrium solutions—hence the traveling-wave solutions—is also called the MRE. (Only when this can cause ambiguity, we will distinguish between these two manifolds by calling the second one the traveling-wave MRE.)

The cnoidal waves are orbitally stable solutions. In [4], this orbital stability is proved by using that in fact the cnoidal waves are constrained minima of the Hamiltonian for fixed values of the integral I. Here orbital stability means stability modulo translations. In other words, the profile of the cnoidal waves is dynamically stable; its "position" is ignored. This is the strongest kind of stability possible for this system because a small change in the speed or amplitude can cause a translational drift. For this reason, in this article, we consider only the profile of the waves and do not bother much about the "position" of the waves.

For the cnoidal waves, this implies that we are only interested in the one-dimensional family of wave profiles. For every fixed value of $I = \gamma$, we choose the cnoidal-wave profile that has its maximum at $x = 0$ (this profile is symmetric around $x = 0$) and denote it by $\bar{u}(\gamma)$. Then the set

$$\{\bar{u}(\gamma) \mid \gamma \geq 0\}$$

is a one-dimensional submanifold of the MRE from which the translations are divided out. The wave speed of the cnoidal wave with $I = \gamma$ is denoted by $\lambda(\gamma)$; it is also the Lagrange multiplier in the Euler–Lagrange equation of the constrained critical-point problem

$$(2) \qquad \partial_x(\bar{u}_{xx} + \bar{u}^2 + \lambda(\gamma)\,\bar{u}) = 0 \quad \text{or} \quad \bar{u}_{xx} + \bar{u}^2 + \lambda(\gamma)\,\bar{u} = \alpha(\gamma)$$

with $\alpha(\gamma)=\frac{\gamma}{2\pi}$, which follows by integrating the equation from 0 to 2π. The cnoidal waves can also be found as unconstrained critical points. To see this, for $\gamma \geq 0$, we define the modified KdV Hamiltonian

$$H_\gamma(u) = H(u) - \lambda(\gamma) I(u).$$

Then for every $\gamma \geq 0$, the cnoidal wave $\bar{u}(\gamma)$ is a critical point of H_γ.

The first general method for using the variational characterization of relative equilibria to draw conclusions about the stability was given in 1985 in [17]. Later, this method was extended to the energy-momentum method in [28] and [29]. In [15] and [23], the sufficient conditions for the stability of the relative equilibria were weakened. In this article, we will extend the use of the variational characterization of the relative equilibria to draw conclusions about the approximation with the cnoidal waves in the damped KdV equation.

For $\varepsilon \neq 0$, the cnoidal waves are no longer solutions of equation (1) and every solution decays to the zero state. This follows from the time behavior of the L^2-norm of u (which equals $I(u)$):

$$\frac{\mathrm{d}}{\mathrm{d}t} I(u) = (I'(u), \partial_x H'(u) - \varepsilon P(u)) = -\varepsilon \int_0^{2\pi} u^2\, dx = -2\varepsilon I(u). \tag{3}$$

(We use the notation $F'(u)$ to denote the variational derivative of a differentiable functional $F(u)$.) In other words, (3) states that $I(u(t)) = I(u(0))\, e^{-2\varepsilon t}$ and that $\lim_{t\to\infty} I(u(t)) = 0$, which implies that $\lim_{t\to\infty} u(t) = 0$.

Although a solution never stays in the neighborhood of one specific cnoidal-wave profile, the full MRE can be useful to approximate the behavior of a solution that starts near a cnoidal wave. This behavior is indicated by numerical experiments and analytical approximations; see [16]. A similar behavior can be found (numerically and experimentally) for the KdV equation with dissipation on an infinite interval. However, in this case, some problems arise in the derivation of an analytical approximation since the decay of the mass functional $M(u) = \int u$ then has to be taken into account; see [18]–[20].

In this article, we approximate a solution of the damped KdV equation on a periodic interval by a projection of the solution on the MRE. An important issue in this article is the justification of the approximation of a solution with this projection. We will use a norm in the Sobolev space H^1_{per} to derive this justification. The usual H^1_{per}-norm is given by

$$\|u\|^2_{H^1_{\mathrm{per}}} = \int_0^{2\pi} [\, u(x)^2 + u_x(x)^2\,]\, dx = \|u\|_0^2 + \|u_x\|_0^2, \qquad u \in H^1_{\mathrm{per}},$$

where $\|\,.\,\|_0$ denotes the L^2_{per}-norm. Because we consider only functions with mean value zero, by the Poincaré inequality, the following norm is equivalent to the H^1_{per}-norm:

$$\|u\|_1^2 = \int_0^{2\pi} u_x(x)^2\, dx = \|u_x\|_0^2, \qquad u \in H^1_{\mathrm{per},0},$$

where $H^1_{\mathrm{per},0}$ is the subspace of H^1_{per} consisting of 2π-periodic functions with mean value 0. Furthermore, we will often use the following Poincaré inequalities comparing the L^2_{per}-norm (respectively, the L^∞_{per}-norm) and the $H^1_{\text{per},0}$-norm:

$$\|u\|_0^2 = \int_0^{2\pi} \left| \int_{x_0}^{x} u_x(\xi)d\xi \right|^2 dx \leq \int_0^{2\pi} \left[\int_0^{2\pi} |u_x(\xi)|d\xi \right]^2 dx \leq (2\pi)^2 \|u\|_1^2, \tag{4}$$

$$\|u\|_\infty = \max_{x\in[0,2\pi]} |u(x)| = \max_{x\in[0,2\pi]} \left| \int_{x_0}^{x} u_x(\xi)d\xi \right| \leq \sqrt{2\pi}\, \|u\|_1. \tag{5}$$

Here x_0 denotes any zero of $u(t)$. (This zero exists because $u(t)$ has mean value zero.)

As we stated before, we are not interested in differences caused by translations. Therefore, we define (analogously to [3, 4]) translation-invariant distances related to the L^2_{per}- and $H^1_{\text{per},0}$-norm, denoted by ρ_0 and ρ_1, respectively, as

$$\rho_i(u_1, u_2) = \min_{\varphi\in[0,2\pi]} \|\Phi^I_\varphi(u_2) - u_1\|_i = \min_{\varphi\in[0,2\pi]} \|\Phi^I_\varphi(u_1) - u_2\|_i, \qquad i = 0, 1.$$

See also Figure 1.

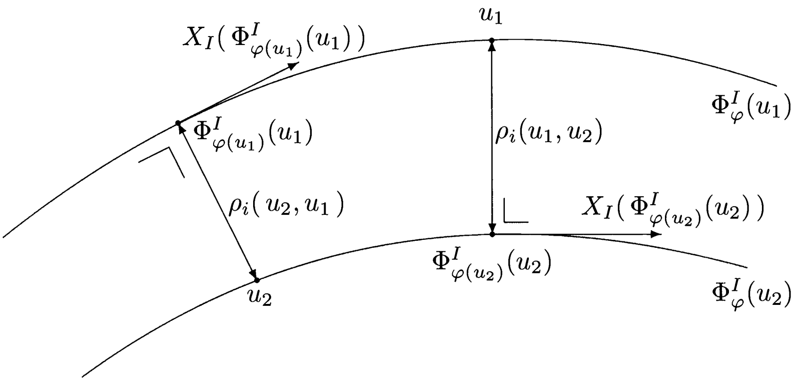

u_1
$X_I(\Phi^I_{\varphi(u_1)}(u_1))$
$\Phi^I_\varphi(u_1)$
$\Phi^I_{\varphi(u_1)}(u_1)$
$\rho_i(u_1, u_2)$
$\rho_i(u_2, u_1)$
$X_I(\Phi^I_{\varphi(u_2)}(u_2))$
$\Phi^I_{\varphi(u_2)}(u_2)$
u_2
$\Phi^I_\varphi(u_2)$

FIG. 1. *The translation-invariant distance $\rho_i(u_1, u_2) = \rho_i(u_2, u_1)$. The translations $\Phi^I_{\varphi(u_1)}$ and $\Phi^I_{\varphi(u_2)}$ are such that $\rho_i(u_1, u_2) = \|\Phi^I_{\varphi(u_1)}(u_1) - u_2\|_i = \|\Phi^I_{\varphi(u_2)}(u_2) - u_1\|_i$.*

To define a projection of a solution $u(t)$, we choose the wave profile on the MRE with an I-value equal to the I-value of the solution. Next, we define a position for this wave profile. It is obvious to choose the position such that the ρ_1-distance is as small as possible.

DEFINITION 1.1. *Let $u(t)$ be a solution of the damped KdV equation. Define the functions $\gamma(t) \in \mathbb{R}$ and $\varphi(t) \in \mathbb{R}$ such that*

$$\gamma(t) = I(u(t)),$$
$$\|\Phi^I_{-\varphi(t)}(u(t)) - \bar{u}(\gamma)\|_1 = \min_{\varphi\in\mathbb{R}} \|\Phi^I_\varphi(u(t)) - \bar{u}(\gamma)\|_1 = \rho_1(\bar{u}(\gamma), u(t)).$$

Finally, we define $\xi(t)$ to be the difference between $u(t)$ and its projection

$$\xi(t) = \Phi^I_{-\varphi(t)}(u(t)) - \bar{u}(\gamma(t)).$$

The projection of the solution $t \to u(t)$ onto the MRE is the curve $t \to \Phi^I_{\varphi(t)}(\bar{u}(\gamma(t)))$.

An important consequence of the choice of φ as given by Definition 1.1 is that $\|\xi\|_1^2 = \rho_1(\bar{u}(\gamma), u)$. Furthermore, the differential equation for the function $\gamma(t)$ is only in terms of γ and hence can be solved explicitly (see equation (3)):

$$\dot{\gamma} = -2\,\varepsilon\,\gamma \qquad \text{implying} \qquad \gamma(t) = \gamma(0)\, e^{-2\varepsilon t}.$$

The main result of this article can now be formulated.

THEOREM 1.2. *For every $\gamma_0 > 0$, there exists a $K > 0$ and an $\varepsilon_0 > 0$ such that any solution $u(t)$ of the damped KdV equation* (1) *with $\varepsilon \leq \varepsilon_0$ that starts with $\gamma(0) \leq \gamma_0$ and, within a distance ε of a cnoidal wave, stays in a relative ε-neighborhood of the family of cnoidal waves.*

Explicitly, if $u(0)$ is such that $\gamma(0) \leq \gamma_0$ and $\rho_1(\bar{u}(\gamma(0)), u(0)) \leq \varepsilon$, then for all $t \geq 0$,

$$\rho_1(\bar{u}(\gamma(t)), u(t)) \leq K\varepsilon e^{-\varepsilon t} = \bar{K}\varepsilon\|\bar{u}(\gamma(t))\|_0 = \tilde{K}\varepsilon\|\bar{u}(\gamma(t))\|_1, \tag{6}$$

where $\bar{K} = K/\sqrt{2\,\gamma(0)}$.

In (6), we use that on every compact γ-interval $\|\bar{u}(\gamma)\|_0$ and $\|\bar{u}(\gamma)\|_1$ are of the same order. Hence on $[0, \gamma_0]$, the quotient $\|\bar{u}(\gamma)\|_0/\|\bar{u}(\gamma)\|_1$ can be estimated by a constant independent of γ.

Remark 1. Notice that the estimate for the initial condition and the estimate for the time behavior are in the same norm.

Remark 2. From Theorem 1.2, it can be deduced that for every $\gamma_0 > 0$, there exists a $\hat{K}$ and an $\hat{\varepsilon}_0$ such that if $H(u(0)) - H(\bar{u}(\gamma(0))) \leq \hat{\varepsilon}_0$, then $H(u(t)) \leq \hat{K}\, I(u(t))$ for all t. This is sketched in Figure 2.

On the contrary, for the KdV–Burgers equation, i.e., $u_t = -\partial_x\,(\,u_{xx} + u^2\,) + \varepsilon\, u_{xx}$, we observe a "self-organization" towards the MRE. In an H-I-figure, this means that every solution decays to zero tangent to the MRE. Hence asymptotically every solution will be below the tangent line to the MRE at 0, hence below the line $H = I$. This behavior is sketched in Figure 3. This self-organization will not occur for the KdV equation with uniform damping. Hence the situation sketched in Figure 2 is also the best possible one. See [16] for more details.

Remark 3. In [14], the damped KdV equation with an additional forcing is considered and the existence of finite-dimensional attractors is investigated. In case there is no forcing, the attractor is trivial. The result in Theorem 1.2 gives more information than the existence of an attractor. It describes an approximation for the intermediate and asymptotic states. The asymptotic result also shows how the solution decays to the attractor 0.

In the next sections, we will prove Theorem 1.2. The proof uses the variational principle which underlies the stability result of the cnoidal waves (see [4]). More explicit, to prove the stability of the cnoidal waves, we can use the (Lyapunov) functional $L(u) = H(u) - H(\bar{u}(I(u)))$. This functional is also similar to the so-called amended Hamiltonian or energy-momentum functional as used in [28, 29]. To prove Theorem 1.2, we will analyze the time behavior of the function $L(u(t))$, where $u(t)$ is a solution of the damped KdV equation, and derive a Gronwall-type inequality for $L(u(t))$. However, it turns out that this inequality is not optimal. To obtain optimal results in time asymptotics, a shift of the MRE to a neighboring ($\mathcal{O}(\varepsilon)$) set has to be performed. The use of such a shift can also be found in [22]. However, in that article, only equilibria of finite-dimensional perturbed Hamiltonian systems are investigated on a finite time scale.

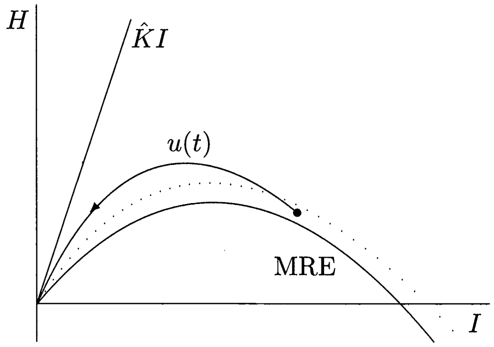


FIG. 2. *Sketch of the projection of a solution of the uniformly damped KdV equation, which starts near the MRE, in the H-I-plane.*

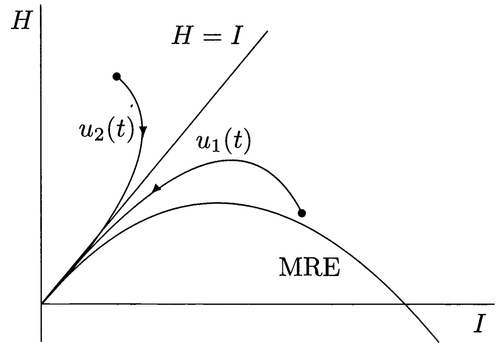


FIG. 3. *Sketch of the projection of two solutions $u_1(t)$ and $u_2(t)$ of the KdV–Burgers equation in the H-I-plane. Note the tangent behavior near zero.*

Remark 4. The shift of the MRE gives rise to an interesting question. The family of cnoidal waves is not invariant for the damped KdV equation. We will see that at every instance there is a forcing that drives a solution away from the MRE.

Keeping in mind the behavior as sketched in Figure 2, it is very unlikely that the MRE is stable for fixed values of ε. In other words, we cannot expect that for a fixed value of ε it yields that for every $\delta > 0$ there exist a $\delta_0 > 0$ and a $T > 0$ such that for every solution $u(t)$ which satisfies $\rho_1(u(0), \bar{u}(\gamma(0))) < \delta_0$, it holds that $\rho_1(u(t), \bar{u}(\gamma(t))) < \delta$ for all $t \geq T$.

The question remains as to if there is another manifold near the MRE that is stable in this sense. A possible candidate could be the shift of the MRE, but we will see later that it has a disadvantage similar to that of the MRE, although it approximates the solution up to higher order in ε. However, that does not help for fixed values of ε. Another possibility could be an iterated shift of the MRE. If it exists, it would give an invariant manifold. However, this question regarding existence is not obvious to answer.

In [13], ideas similar to those used in this article are exploited to analyze the relevance of a two-dimensional family of relative equilibria of a finite-dimensional mechanical system with one cyclic coordinate to which uniform friction is added. An extension to higher-dimensional manifolds of relative equilibria of (finite-dimensional) Hamiltonian systems with symmetries and their relevance under a dissipative perturbation was recently established in [12].

2. A first analysis of the damped KdV equation. As seen in §1, the time behavior of the additional first integral $\gamma(t) = I(u(t))$, where $u(t)$ is a solution of the damped KdV equation, is given by

$$\gamma(t) = I(u(t)) = I(u(0))e^{-2\varepsilon t} = \gamma(0)e^{-2\varepsilon t}.$$

This implies that every solution converges to 0 and hence to the MRE. By Definition 1.1 it holds that $I(\bar{u}(\gamma) + \xi) = I(\bar{u}(\gamma))$, implying that

$$(7) \qquad -\int_0^{2\pi} \bar{u}\,\xi = \frac{1}{2}\int_0^{2\pi} \xi^2;$$

therefore,

$$\|\xi(t)\|_0^2 = 2|(\bar{u}(\gamma(t)), \xi(t))| \le 2\|\bar{u}(\gamma(t))\|_0 \|\xi(t)\|_0,$$

and hence

$$(8) \qquad \|\xi(t)\|_0 \le 2\sqrt{2\gamma(0)}\, e^{-\varepsilon t}.$$

This implies that the translation-invariant L^2-distance between a solution and the MRE is less than or equal to a constant times the L^2-norm of the solution.

As we stated previously, to prove Theorem 1.2, we will make use of a similar (Lyapunov) functional as featured in the energy-momentum method to determine the stability of relative equilibria in an unperturbed Hamiltonian system (see [28, 29] or (for the KdV equation) [4]). To prove the stability of the cnoidal waves with such a technique, it is essential that the cnoidal wave $\bar{u}(\gamma)$ with minimal period 2π is a constrained minimum of the Hamiltonian of the (unperturbed) KdV equation on the level set with $I = \gamma$. This property is proved in Lemma A.1 in the appendix. This lemma implies that the following functional acts as a Lyapunov functional for a cnoidal wave in the case of the unperturbed KdV equation.

DEFINITION 2.1. *Let $u \in H^1_{\mathrm{per},0}$. Define the functional $L(u)$ on $H^1_{\mathrm{per},0}$ as*

$$(9) \qquad L(u) = H(u) - H(\bar{u}(\gamma)) = H_\gamma(u) - H_\gamma(\bar{u}(\gamma))$$

with $\gamma = I(u)$.

Furthermore, define the self-adjoint operator $\tilde{Q}(\gamma)$ on H^1_{per} as

$$\tilde{Q}(\gamma) = D^2\hat{L}(\bar{u}(\gamma)) = D^2H_\gamma(\bar{u}(\gamma))| = -D_{xx} - \lambda(\gamma) - 2\bar{u}(\gamma),$$

and let $Q(\gamma)$ be its restriction on $H^1_{\mathrm{per},0}$; hence

$$Q(\gamma) = D^2\hat{L}(\bar{u}(\gamma))|_{H^1_{\mathrm{per},0}} = D^2H_\gamma(\bar{u}(\gamma))|_{H^1_{\mathrm{per},0}}.$$

The Euler–Lagrange equation (2), i.e., $(\bar{u}_{xx} + \bar{u}^2 + \lambda\,\bar{u})$ is constant, the notation of Definition 1.1, and identity (8) imply that $L(u)$ can be written as

$$\begin{aligned} L(u) = H(\bar{u}+\xi) - H(\bar{u}) &= \int_0^{2\pi}\left[\bar{u}_x\,\xi_x + \frac{1}{2}\xi_x^2 - \bar{u}^2\,\xi - \bar{u}\,\xi^2 - \frac{1}{3}\,\xi^3\right] \\ &= \int_0^{2\pi} \xi\,[-\bar{u}_{xx} - \bar{u}^2 - \lambda\,\bar{u}] + \int_0^{2\pi}\left[\frac{1}{2}\,\xi_x^2 - \bar{u}\,\xi^2 + \lambda\,\bar{u}\,\xi - \frac{1}{3}\,\xi^3\right] \\ &= \int_0^{2\pi}\left[\frac{1}{2}\,\xi_x^2 - \bar{u}\,\xi^2 - \frac{1}{2}\lambda\,\xi^2 - \frac{1}{3}\,\xi^3\right] = \frac{1}{2}\,(Q(\gamma)\xi, \xi) - \frac{1}{3}\int_0^{2\pi}\xi^3. \end{aligned}$$

Define $\tilde{\mathbb{Y}}(\gamma)$ as the subspace of the tangent space to the I-level set that is orthogonal to the direction of the I-flow in H^1_{per}, and define $\mathbb{Y}(\gamma)$ to be the restriction to $H^1_{\text{per},0}$:

$$\tilde{\mathbb{Y}}(\gamma) = \{y \in H^1_{\text{per}} \mid (y, X_I(\bar{u}(\gamma))) = 0 \ \wedge \ (y, I'(\bar{u}(\gamma))) = 0\},$$
$$\mathbb{Y}(\gamma) = \{y \in H^1_{\text{per},0} \mid (y, X_I(\bar{u}(\gamma))) = 0 \ \wedge \ (y, I'(\bar{u}(\gamma))) = 0\}.$$

From the minimality of the cnoidal waves, it follows that $Q(\gamma)$ is strictly positive definite on $\mathbb{Y}(\gamma)$ for a fixed value of $\gamma > 0$.

For our purpose, we need a bit stronger property, namely that L is equivalent with the translation-invariant H^1_{per}-distance on every $L^2_{\text{per},0}$-compact set, which includes 0.

LEMMA 2.2. *For every compact interval $\mathcal{G} \in \mathbb{R}^+_0$ with $0 \in \mathcal{G}$, there exist $C \geq c > 0$ and a neighborhood $\mathcal{U} \subset H^1_{\text{per},0}$ of the MRE such that for all $u \in \mathcal{U}$ with $I(u) \in \mathcal{G}$, it holds that*

$$c\,\rho_1^2(\bar{u}(\gamma), u) \leq L(u) \leq C\,\rho_1^2(\bar{u}(\gamma), u) \tag{10}$$

with $\gamma = I(u)$.

Proof. Let $\gamma > 0$. First, we prove that L is bounded from above. Let $u \in H^1_{\text{per},0}$ and write $u = \Phi^I_\varphi(\bar{u}(\gamma) + \xi)$ as in Definition 1.1. Using the Poincarè inequalities (4) and (5), it is easy to calculate that there exists some $C_1(\gamma) > 0$ such that

$$\begin{aligned} L(u) &= \int_0^{2\pi} \left[\frac{1}{2}\xi_x^2 - \bar{u}\xi^2 - \frac{1}{2}\lambda\,\xi^2 - \frac{1}{3}\xi^3\right] \\ &\leq \frac{1}{2}\|\xi\|_1^2 + \|\bar{u}\|_\infty \|\xi\|_0^2 + \frac{1}{2}|\lambda|\|\xi\|_0^2 + \frac{1}{3}\|\xi\|_\infty\|\xi\|_0^2 \\ &\leq 2\,\pi^2\,\|\xi\|_1^2\left[\frac{1}{4\,\pi^2} + 2\,\|\bar{u}\|_\infty + |\lambda| + \frac{2\sqrt{2\pi}}{3}\|\xi\|_1\right] \\ &\leq C_1(\gamma)\|\xi\|_1^2 \end{aligned} \tag{11}$$

in a neighborhood of the MRE, e.g., if $\|\xi\|_1 \leq 1$. By definition, $\|\xi\|_1^2 = \rho_1(u, \bar{u})$; hence equation (11) implies that

$$L(u) \leq C_1(\gamma)\rho_1^2(u, \bar{u}(\gamma)) \tag{12}$$

and $C_1(\gamma)$ is bounded if $\gamma \to 0$.

Lemma A.3 in the appendix yields that $Q(\gamma)$ is strictly positive definite on $\mathbb{Y}(\gamma)$; explicitly,

$$(Q(\gamma)\,y, y) \geq c_1(\gamma)\|y\|_1^2 \quad \text{for all } y \in \mathbb{Y}(\gamma)$$

with $c_1(\gamma) > 0$ and $\lim_{\gamma\to 0} c_1(\gamma) > 0$ as well.

This inequality implies a lower bound on L. To see this, write $\xi = a\,X_I(\bar{u}) + b\,I'(\bar{u}) + y$, where $y \in \mathbb{Y}(\gamma)$. We will show that a and b are of the order $\|\xi\|_1^2$ if $\|\xi\|_1$ is small. First, we estimate a. We know that $(\,\xi, X_I(\Phi^I_{\varphi(u)}(u))\,) = 0$; therefore,

$$\begin{aligned} 0 &= (\xi, X_I(\Phi^I_{\varphi(u)}(u))) = (\xi, X_I(\bar{u})) + \mathcal{O}(\|\xi\|_0^2) \\ &= a\|X_I(\bar{u})\|_0^2 + \mathcal{O}(\|\xi\|_1^2). \end{aligned}$$

Next we estimate b, using the fact that $I(\Phi^I_{\varphi(u)}(u)) = I(u) = I(\bar{u})$, by

$$\begin{aligned} 0 &= I(\Phi^I_{\varphi(u)}(u)) - I(\bar{u}) = (I'(\bar{u}), \xi) + \mathcal{O}(\|\xi\|_0^2) \\ &= b\|I'(\bar{u})\|_0^2 + \mathcal{O}(\|\xi\|_1^2). \end{aligned}$$

Now we know that the "largest" part of ξ is in $\mathbb{Y}(\gamma)$, and we can derive a relation between $L(u)$ and $\|\xi\|_1^2$ for $\|\xi\|_1$ small:

$$\begin{aligned} L(u) &= L(\Phi^I_{\varphi(u)}(u)) = L(\bar{u}) + \frac{1}{2}(Q(\gamma)\xi,\xi) + \mathcal{O}(\|\xi\|_0^3) \\ &= \frac{1}{2}(Q(\gamma)y,y) + \mathcal{O}(|b|\|\xi\|_0 + |b|^2 + \|\xi\|_0^3) \\ &\geq \frac{1}{2}c_1(\gamma)\|y\|_1^2 + \mathcal{O}(\|\xi\|_0^3) = \frac{1}{2}c_1(\gamma)\|\xi\|_1^2 + \mathcal{O}(\|\xi\|_0^3). \end{aligned}$$

This means that there exists a $\delta_0 > 0$ such that for all u with $\rho_1(\bar{u},u) < \delta_0$ (recall that $\|\xi\|_1^2 = \rho_1(u,\bar{u})$),

$$L(u) \geq \frac{1}{4}c_1(\gamma)\rho_1(\bar{u},u).$$

To prove equation (10), we use the facts that $\lim_{\gamma\to 0} c_1(\gamma) \neq 0$ and $C_1(\gamma)$ is bounded. Hence in every compact interval $\mathcal{G}$ that includes 0, there exist $0 < c \leq C$ such that (10) holds. □

After these observations about the unperturbed KdV equation, we return to the damped KdV equation. The time behavior of the difference function $\xi(t)$ gives an idea about what causes the deviation of solutions of the damped KdV equation of the MRE. Using Definition 1.1, we see that

$$\begin{aligned} \dot{\xi} &= \frac{\mathrm{d}}{\mathrm{d}t}\left[\Phi^I_{-\varphi(t)}(u(t))\right] - \dot{\bar{u}}(\gamma(t)) \\ &= -\partial_x[\bar{u}_{xx} + \xi_{xx} + 2(\bar{u}+\xi)^2] - \varepsilon(\bar{u}+\xi) - \dot{\bar{u}} - \dot{\varphi}\partial_x(\bar{u}+\xi). \end{aligned}$$

To recognize more structure in this equation, we will rewrite it. By using the Euler–Lagrange equation (27) (in the appendix) and the differential equation for γ, it follows that

$$\begin{aligned} \dot{\xi} &= \partial_x[\xi_{xx} + 2\bar{u}\xi + \lambda\xi + \xi^2] + (\lambda - \dot{\varphi})\partial_x(\bar{u}+\xi) - \varepsilon[2\gamma\bar{u}'(\gamma) - \bar{u}] - \varepsilon\xi \\ &= \partial_x H_\gamma{}'(\bar{u}+\xi) + (\lambda - \dot{\varphi})\partial_x(\bar{u}+\xi) - \varepsilon\xi + R(\bar{u}(\gamma),\varepsilon) \end{aligned}$$

with the so-called residual

$$R(\bar{u}(\gamma),\varepsilon) = -\varepsilon[2\gamma\bar{u}'(\gamma) - \bar{u}]. \tag{13}$$

The first two terms of the equation for $\dot{\xi}$ have a Hamiltonian origin. The first term is the modified KdV Hamiltonian. The second term induces a translation of the wave profile; hence this term will be irrelevant for our analysis. The third and fourth terms are the most relevant for our analysis. The third term represents the damping. The fourth term does not depend on ξ. It is called the residual because it shows the effect of the damped KdV equation on the MRE, except for some irrelevant influence in the translation direction $X_I(\bar{u})$. The residual is an element of $\mathbb{Y}(\gamma)$; hence if the residual is not equal to 0, then a solution that starts on the MRE will soon deviate from it. This implies that the residual acts like a forcing in the ξ-equation. In other words, the third and fourth terms show a competition between a dissipation directed towards the MRE and a forcing away from the MRE.

We have seen that the functional L is equivalent to the H_1-norm of ξ. Therefore, we look at the time behavior of this functional to see how fast ξ can grow. Using the translation invariance of H and I, it follows immediately that $L(u) = L(\bar{u}+\xi)$. Using the second expression in (9) and the equation for $\dot{\xi}$, it follows that

$$\frac{\mathrm{d}}{\mathrm{d}t}L(u(t)) = \dot{\lambda}[I(\bar{u}+\xi) - I(\bar{u})] + (H_\gamma{}'(\bar{u}+\xi), -\varepsilon\xi + R(\bar{u},\varepsilon) - \dot{\bar{u}}).$$

The first term is zero because $I(\bar{u}+\xi) = I(\bar{u})$. Note that $H'_\gamma(\bar{u}+\xi) = Q(\gamma)\xi - \xi^2$; hence $(H_\gamma{}'(\bar{u}+\xi), -\varepsilon\xi) = -2\varepsilon L(u) + \frac{1}{3}(\xi^2, \xi)$. We again use the Euler–Lagrange equation to rewrite the inner product with $(R(\bar{u},\varepsilon) - \dot{\bar{u}})$. This yields

$$\frac{\mathrm{d}}{\mathrm{d}t}L(u(t)) = -2\varepsilon L(\gamma,\xi) + \varepsilon\int_0^{2\pi} \bar{u}^2\xi + \varepsilon\int_0^{2\pi} \bar{u}\xi^2 + \frac{1}{3}\varepsilon\int_0^{2\pi} \xi^3. \tag{14}$$

Note that in expression (14), for the time behavior of L, the first term is dissipative, the second is a forcing-like component, and the last two terms are small compared to the first two terms (if ξ and γ are small).

With this expression for $\frac{\mathrm{d}}{\mathrm{d}t}L(u(t))$, we derive a preliminary estimate for the functional L.

PROPOSITION 2.3. *For every $\varepsilon > 0$, there exist a $\delta > 0$ and a constant K_0, both depending on $\gamma(0)$, such that*

$$\sqrt{L(u(t))} \le K_0\left[\sqrt{L(u(0))} + 2\sqrt{2\pi}\gamma(0)\right]e^{-\varepsilon t} \tag{15}$$

as long as $\|\xi(t)\|_1 < \delta$.

Proof. For the last three terms in equation (14), it holds that

$$\left|\int_0^{2\pi} \bar{u}^2\xi\right| \le \|\xi\|_\infty \int_0^{2\pi} \bar{u}^2 \le 2\sqrt{2\pi}\gamma\|\xi\|_1,$$

$$\left|\int_0^{2\pi} \bar{u}\xi^2\right| \le \|\xi\|_\infty \int_0^{2\pi} \xi\bar{u} \le \sqrt{2\pi}\|\xi\|_1\|\bar{u}\|_0\|\xi\|_0 \le 4\pi\sqrt{\pi}\sqrt{\gamma}\|\xi\|_1^2,$$

$$\left|\int_0^{2\pi} \xi^3\right| \le \|\xi\|_\infty \int_0^{2\pi} \xi^2 \le \sqrt{2\pi}\|\xi\|_1\|\xi\|_0^2 \le 8\pi\sqrt{\pi}\sqrt{\gamma}\|\xi\|_1^2.$$

In these estimates, we use that $\|\xi\|_0 \le 2\sqrt{2\gamma}$ and the Poincaré inequalities (4) and (5).

To be able to switch from $\|\xi\|_1$ to the functional L, we will use Lemma 2.2. Let δ be such that the equivalence relation (10) holds for all $\|\xi\|_1 < \delta$. Substituting the relations above into (14) and using (10) gives the following estimate for the time behavior of L:

$$\frac{\mathrm{d}}{\mathrm{d}t}L \le \left[-2\varepsilon + \varepsilon\frac{20\sqrt{\gamma\pi\sqrt{\pi}}}{3c}\right]L + \varepsilon\frac{2\sqrt{2\pi}\gamma}{\sqrt{c}}\sqrt{L}.$$

Define $N(t) = e^{\varepsilon t}\sqrt{L}$; then this inequality implies that $N(t) = 0$ or

$$\dot{N} \le \left[\varepsilon\frac{10\sqrt{\gamma\pi\sqrt{\pi}}}{3c}\right]N + \varepsilon\frac{\sqrt{2\pi}\gamma}{\sqrt{c}}e^{\varepsilon t}.$$

Applying Gronwall's lemma to this equation gives

$$N(t) \le K_0 N(0) + \varepsilon K_0\int_0^t \frac{\sqrt{2\pi}e^{-\varepsilon\tau}\gamma(0)}{\sqrt{c}}d\tau = K_0\left[N(0) + \frac{\sqrt{2\pi}\gamma(0)}{\sqrt{c}}(1 - e^{-\varepsilon t})\right],$$

where $K_0 = \exp\left[\frac{10\pi\sqrt{\pi}\sqrt{\gamma(0)}}{3c}\right]$. □

The estimate of Proposition 2.3 only provides information about the approximation on a finite time scale. Even if we start on the MRE, hence with $\|\xi(0)\|_1 = 0$, after some time the norm of the right-hand side of estimate (15) is of order $\sqrt{\gamma}$ instead of order $\varepsilon\sqrt{\gamma}$. This effect is induced by the residual, which after an integration becomes of order 1 instead of order ε. In other words, we need a smaller residual. The present residual is induced by the projection on the MRE, which approximates the solution in zeroth order. If we have a better approximation than this projection, we can expect a smaller residual. In the next section, we will derive such a better approximation and prove Theorem 1.2 by using a functional related to this better approximation.

3. Justification of the approximation. The residual $R(\bar{u}(\gamma), \varepsilon)$ measures how well the curve $\bar{u}(\gamma(t))$ obeys the damped KdV equation. The function $t \to \bar{u}(\gamma(t))$ is a zeroth-order approximation of the damped KdV equation and therefore gives a residual of order ε. It can be expected that the residual for a first-order approximation of the damped KdV equation is smaller, of order ε^2. Using the knowledge that the L^2-norm $I(u)$ is $\gamma(0)e^{-2\varepsilon t}$, which is a slow time behavior, we try to find a better approximation of the form

$$u(t) = \Phi^I_{\lambda(\varepsilon)t}(\bar{u}(\gamma(t)) + \varepsilon\bar{v}_1(\gamma(t), \varepsilon)).$$

Substitution of this expression in the dynamical system (1) gives

$$\partial_x[H'(\bar{u} + \varepsilon\bar{v}_1) - \lambda(\varepsilon)I'(\bar{u} + \varepsilon\bar{v}_1)] - \varepsilon[(\bar{u} + \varepsilon\bar{v}_1) - 2\gamma(\bar{u}'(\gamma) + \bar{v}_1'(\gamma, \varepsilon))] = 0.$$

After taking first-order terms in ε of this equation, it remains (up to order-ε^2 terms)

$$H'(\bar{u} + \varepsilon\bar{v}_1) - \lambda(\varepsilon)I'(\bar{u} + \varepsilon\bar{v}_1) = \partial_x^{-1}[R(\bar{u}, \varepsilon)] + \alpha(\varepsilon). \tag{16}$$

(The operator ∂_x^{-1} is defined to act on the space $H^1_{\text{per},0}$ and $\alpha(\varepsilon)$ is a constant which is introduced by the integration.) If we can find a solution $(\bar{u}_1(\gamma, \varepsilon), \lambda(\varepsilon), \alpha(\varepsilon))$ of equation (16) (with $\bar{u}_1(\gamma, \varepsilon) = \bar{u}(\gamma) + \varepsilon\bar{v}_1(\gamma, \varepsilon)$), then we expect that the residual in $\bar{u}_1(\gamma, \varepsilon)$ is of order ε^2, an improvement compared to the residual in $\bar{u}(\gamma)$, which is of order ε.

Another way to interpret the definition of the function $\bar{u}_1(\gamma, \varepsilon)$ is by noticing that $\bar{u}_1(\gamma, \varepsilon)$ is a constrained critical point of a new Hamiltonian

$$H_{\text{new}}(u, \gamma, \varepsilon) = H(u) - (\partial_x^{-1}R(\bar{u}(\gamma), \varepsilon),\, u)$$

on the level set of $I(u) = \gamma$. Hence $\bar{u}_1$ is a kind of new relative equilibrium. However, the new Hamiltonian H_{new} is not translation invariant; hence neither can we find a two-parameter family of constrained critical points nor is $\Phi^I_{\lambda(\varepsilon)t}(\bar{u}_1)$ a solution of the new Hamiltonian system. Because we ignore all shifts in the solution, a curve of new relative equilibria $\bar{u}_1(\gamma, \varepsilon)$ is sufficient to give a better approximation for a solution of the damped KdV equation.

Remark 5. It is possible to define a new translation-invariant Hamiltonian which possesses a two-dimensional family of relative equilibria that give a residual of order ε^2. Analogously to the definition of the translation-invariant distance, we define this new Hamiltonian as

$$\tilde{H}_{\text{new}}(u, \gamma, \varepsilon) = H(u) - (\partial_x^{-1}R(\bar{u}, \varepsilon), \Phi^I_{\psi(u)}(u))$$

with the functional $\psi : H^1_{\text{per},0} \to [0, 2\pi)$ such that

$$(17) \qquad (\partial_x^{-1} R(\bar{u}, \varepsilon), \Phi^I_{\psi(u)}(u)) = \min_{\varphi \in [0,2\pi)} (\partial_x^{-1} R(\bar{u}, \varepsilon), \, \Phi^I_\varphi(u)).$$

If the minimum in (17) is attained at more than one value $\varphi \in [0, 2\pi)$, then $\psi(u)$ is the smallest one. Now H_{new} is a translation-invariant functional and its derivative with respect to u is

$$\tilde{H}'_{\text{new}}(u, \gamma, \varepsilon) = H'(u) - \Phi^I_{-\psi(u)}(\partial_x^{-1} R(\bar{u}, \varepsilon)).$$

Hence the constrained critical points $\tilde{u}_1$ of $\tilde{H}_{\text{new}}$ on the level set $I = \gamma$ satisfy

$$0 = H'(\tilde{u}_1) - \tilde{\lambda}_1 I'(\tilde{u}_1) - \Phi^I_{-\psi(u)}(\partial_x^{-1} R(\bar{u}, \varepsilon)).$$

Because $\psi(\bar{u}) = 0$, the approximation $\bar{u}_1$ gives rise to a residual of order ε^2. See [10] for more details.

First, we show that there indeed exists a curve of new relative equilibria $\bar{u}_1(\gamma, \varepsilon)$ in the neighborhood of $\varepsilon = 0$ by applying the implicit-function theorem to H_{new}, which is a perturbation of the original Hamiltonian H. To be able to apply the implicit-function theorem, it is important that $\partial_x^{-1} R(\bar{u}(\gamma), \varepsilon)$ is orthogonal to the kernel of $Q(\gamma)$ and hence orthogonal to $X_I(\bar{u})$. This will be shown in the proof of the next lemma. By taking the inner product with $X_I(\bar{u}_1)$ in equation (16), it follows that $\bar{u}_1$ is orthogonal to $R(\bar{u}(\gamma), \varepsilon)$.

LEMMA 3.1. *For every $\gamma > 0$ there exists an $\varepsilon_0(\gamma)$ and a unique curve*

$$\{\bar{u}_1(\gamma, \varepsilon) \mid |\varepsilon| \le \varepsilon_0(\gamma)\}$$

of minimal points of the Hamiltonian H_{new} on the level set $I = \gamma$ in $H^1_{\text{per},0}$. Explicitly, for every $|\varepsilon| \le \varepsilon_0(\gamma)$, there exist unique Lagrange multipliers $\lambda_1(\gamma, \varepsilon)$ and $\alpha_1(\gamma, \varepsilon)$ such that

$$(18) \qquad \begin{aligned} 0 &= H'_{\text{new}}(\bar{u}_1(\gamma, \varepsilon), \gamma, \varepsilon) - \lambda_1(\gamma, \varepsilon) I'(\bar{u}_1(\gamma, \varepsilon)) - \alpha_1(\gamma, \varepsilon), \\ 0 &= I(\bar{u}_1(\gamma, \varepsilon)) - \gamma. \end{aligned}$$

Furthermore, there exists a $K(\gamma) > 0$ such that for all $|\varepsilon| \le \varepsilon_0(\gamma)$, it holds that

$$\begin{aligned} \|\bar{u}(\gamma) - \bar{u}_1(\gamma, \varepsilon)\|_1 &\le K(\gamma) \|\partial_x^{-1} R(\bar{u}(\gamma), \varepsilon)\|_0 = \mathcal{O}(\varepsilon \|\bar{u}(\gamma)\|_0), \\ |\lambda(\gamma) - \lambda_1(\gamma, \varepsilon)| \; &\le K(\gamma) \|\partial_x^{-1} R(\bar{u}(\gamma), \varepsilon)\|_0 = \mathcal{O}(\varepsilon \|\bar{u}(\gamma)\|_0). \end{aligned}$$

Finally, $\lim_{\gamma \to 0} K(\gamma)$ and $\lim_{\gamma \to 0} \varepsilon_0(\gamma)$ exist and $\varepsilon_0(0) > 0$.

Proof. Let $\gamma > 0$. As we stated previously, we use the implicit-function theorem (see, e.g., [9]) to prove this lemma. First, we reformulate the problem. Instead of looking for a 2π-periodic solution of (18) with mean value, it is more convenient to add the mean-value zero condition to the equations and consider the problem in the space of all 2π-periodic functions. Hence we look for a 2π-periodic solution of

$$(19) \qquad 0 = \begin{pmatrix} H'(u) - \lambda I'(u) - \alpha \mathbf{1} \\ I(u) - \gamma \\ M(u) \end{pmatrix} - \begin{pmatrix} \partial_x^{-1} R(\bar{u}(\gamma), \varepsilon) \\ 0 \\ 0 \end{pmatrix},$$

where $\mathbf{1}$ is the function that equals 1 for all $x \in [0, 2\pi]$ and $M(u) = \int_0^{2\pi} u$. For $\varepsilon = 0$ (the unperturbed case), this problem does not have a unique solution in H^1_{per}. We have seen that a one-dimensional manifold of solutions can be formed: all translates of the cnoidal wave $\bar{u}(\gamma)$, hence $\{(\Phi^I_\varphi(\bar{u}(\gamma)), \lambda(\gamma), \alpha(\gamma)) \mid \varphi \in [0, 2\pi)\}$. Hence this will cause a problem in the application of the implicit-function theorem to these equations in H^1_{per}. To avoid this problem, we will use the fact that $\bar{u}_1$ is orthogonal to $R(\bar{u}, \varepsilon)$.

We distinguish two cases.

1. If $(R(\bar{u}, \varepsilon), \bar{u}_x)) \neq 0$, then we add the equation

$$0 = (\partial_x^{-1} R(\bar{u}, \varepsilon), \partial_x u) = -(R(\bar{u}, \varepsilon), u) \tag{20}$$

to our set of equations (19), and we add the term $\beta X_I(u)$ with the extra unknown β to the first equation. Explicitly, for $\varepsilon \geq 0$, we introduce the functions F and F_0 on $H^1_{\text{per}} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$:

$$F(u, \lambda, \alpha, \beta; \varepsilon) = \begin{pmatrix} H'(u) - \lambda I'(u) - \alpha \mathbf{1} - \beta X_I(u) - \partial_x^{-1} R(\bar{u}(\gamma), \varepsilon) \\ I(u) - \gamma \\ M(u) \\ (R(\bar{u}, \varepsilon), u) \end{pmatrix}$$

and $F_0(u, \lambda, \alpha, \beta) = F(u, \lambda, \alpha, \beta; 0)$ for all $u \in H^1_{\text{per}}$ and $\lambda, \alpha, \beta \in \mathbb{R}$. A solution of the equation

$$F(u, \lambda, \alpha, \beta; \varepsilon) = 0$$

gives a constrained critical point of the new Hamiltonian H_{new} on level sets of I. Indeed, take the inner product of $X_I(u)$ with the first equation in $F = 0$; then it follows that $\beta = 0$. In other words, the first equation in $F = 0$ is the Euler–Lagrange equation for the critical-point problem. Furthermore, if $\varepsilon = 0$, the critical-point problem for the KdV equation reappears; hence $F_0(\bar{u}(\gamma), \lambda(\gamma), \alpha(\gamma), 0) = 0$ and this solution is unique for $\varepsilon = 0$ in $H^1_{\text{per}} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ thanks to the last equation.

The function F_0 satisfies the following properties:

(i) F_0 is continuously differentiable. Indeed, it is a straightforward calculation to see that for all $(u, \lambda, \alpha, \beta), (\hat{u}, \hat{\lambda}, \hat{\alpha}, \hat{\beta}) \in H^1_{\text{per}} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$,

$$F_0(u, \lambda, \alpha, \beta) = F_0(\hat{u}, \hat{\lambda}, \hat{\alpha}, \hat{\beta}) + DF_0(\hat{u}, \hat{\lambda}, \hat{\alpha}, \hat{\beta}) \begin{pmatrix} u - \hat{u} \\ \lambda - \hat{\lambda} \\ \alpha - \hat{\alpha} \\ \beta - \hat{\beta} \end{pmatrix} - \begin{pmatrix} G(u, \hat{u}, \lambda, \hat{\lambda}) \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

where $G(u, \hat{u}, \lambda, \hat{\lambda}) = (u - \hat{u})^2 + (\lambda - \hat{\lambda})(u - \hat{u})$. This implies that

$$\begin{aligned} &\left\| F_0(u, \lambda, \alpha, \beta) - F_0(\hat{u}, \hat{\lambda}, \hat{\alpha}, \hat{\beta}) - DF_0(\hat{u}, \hat{\lambda}, \hat{\alpha}, \hat{\beta}) \begin{pmatrix} u - \hat{u} \\ \lambda - \hat{\lambda} \\ \alpha - \hat{\alpha} \\ \beta - \hat{\beta} \end{pmatrix} \right\|_0 \\ &\quad \leq \|(\lambda - \hat{\lambda})(u - \hat{u})\|_0 + \|(u - \hat{u})^2\|_0 \\ &\quad \leq 2\pi |\lambda - \hat{\lambda}| \|u - \hat{u}\|_1 + 8\pi^3 \|u - \hat{u}\|_1^2. \end{aligned}$$

Hence F_0 is continuously differentiable.

(ii) $DF_0(\bar{u}(\gamma), \lambda(\gamma), \alpha(\gamma), 0)$ is injective and surjective. We prove this property in Lemma A.5 in the appendix.

(iii) $DF_0(\bar{u}(\gamma),\lambda(\gamma),\alpha(\gamma),0)^{-1}[F(u,\lambda,\alpha,\beta;\varepsilon)-F_0(u,\lambda,\alpha,\beta)]$ is Lipschitz continuous. This last property follows from the facts that F and F_0 are Lipschitz continuous and $DF_0(\bar{u}(\gamma),\lambda(\gamma),\alpha(\gamma),0)^{-1}$ is bounded. This last observation is a consequence of the minimality of the cnoidal waves and hence of Lemma 2.2.

With (i)–(iii), all conditions for the application of the implicit-function theorem to the equation $F=0$ are satisfied.

2. If $(R(\bar{u},\varepsilon),\bar{u}_x))=0$, then we consider the equations (19) on a subspace of H^1_{per}, namely

$$\mathbb{X}(\gamma)=\{u\in H^1_{\text{per}}\mid (u,X_I(\bar{u}(\gamma)))=0\}.$$

Also, in a way similar to case 1, we can prove that all conditions for the application of the implicit-function theorem to equation (19) are satisfied.

The application of the implicit-function theorem implies that there exists a neighborhood $\mathcal{U}_1(\gamma)\subset H^1_{\text{per}}\times\mathbb{R}\times\mathbb{R}$ around the relative equilibrium $(\bar{u}(\gamma),\lambda(\gamma),\alpha(\gamma))$, a positive number $\varepsilon_0(\gamma)$, and a curve of points $\big((\bar{u}_1(\gamma,\varepsilon),\lambda_1(\gamma,\varepsilon),\alpha_1(\gamma,\varepsilon))\big)_{|\varepsilon|\le\varepsilon_0(\gamma)}$ in $\mathcal{U}_1(\gamma)$ such that

$$\begin{aligned}0&=H'(\bar{u}_1)-\lambda_1 I'(\bar{u}_1)-\alpha_1-\partial_x^{-1}R(\bar{u},\varepsilon),\\ 0&=I(\bar{u}_1)-\gamma,\\ 0&=M(\bar{u}_1).\end{aligned}\tag{21}$$

Furthermore, it follows also that

$$\begin{aligned}\|\bar{u}_1(\gamma,\varepsilon)-\bar{u}(\gamma)\|_1&=\mathcal{O}(\|\partial_x^{-1}R(\bar{u},\varepsilon)\|_0),\\ |\lambda_1(\gamma,\varepsilon)-\lambda(\gamma)|&=\mathcal{O}(\|\partial_x^{-1}R(\bar{u},\varepsilon)\|_0),\\ |\alpha_1(\gamma,\varepsilon)-\alpha(\gamma)|&=\mathcal{O}(\|\partial_x^{-1}R(\bar{u},\varepsilon)\|_0).\end{aligned}\tag{22}$$

From the definition of the residual $R(\bar{u},\varepsilon)$ (see (13)) and by using properties of elliptic functions (see [8]), it follows that there is a constant K_2 (independent of γ) such that $\|R(\bar{u}(\gamma),\varepsilon)\|_0\le\varepsilon K_2\sqrt{\gamma}$.

We must still show that $\bar{u}_1(\gamma,\varepsilon)$ is a constrained *minimum* on the level set with $I=\gamma$. Consider the linearization of $H'(u)-\lambda_1 I'(u)-\partial_x^{-1}R(\bar{u},\varepsilon)$ around $\bar{u}_1$:

$$\tilde{Q}_1(\gamma,\varepsilon)=D^2H(\bar{u}_1)-\lambda_1 D^2I(\bar{u}_1)=Q(\gamma)-(\lambda_1-\lambda)Id-2(\bar{u}_1-\bar{u}).$$

Using the fact that $Q(\gamma)$ is strictly positive definite on $\mathbb{Y}(\gamma)$ and (22), we will show that $\tilde{Q}_1(\gamma)$ is strictly positive definite on $\tilde{\mathbb{Y}}_1(\gamma)=\{\eta\in H^1_{\text{per}}\mid (\eta,X_I(\bar{u}_1))=0,(\eta,I'(\bar{u}_1))=0\}$.

Let $\eta\in\tilde{\mathbb{Y}}_1(\gamma)$. Then we have

$$\begin{aligned}(\tilde{Q}_1(\gamma)\eta,\eta)&=(Q(\gamma)\eta,\eta)-(\lambda_1-\lambda_0)\|\eta\|_0^2-2((\bar{u}_1-\bar{u})\eta,\eta)\\ &\ge(Q(\gamma)\eta,\eta)-|\lambda_1-\lambda_0|\|\eta\|_0^2-2\|\bar{u}_1-\bar{u}\|_1\|\eta\|_0^2\\ &\ge(Q(\gamma)\eta,\eta)-3K\|\partial_x^{-1}R_3(\bar{u},\varepsilon)\|_0\|\eta\|_0^2.\end{aligned}$$

From the minimality of the cnoidal waves, it follows that there exists a $c(\gamma)$ such that $(Q(\gamma)\xi,\xi)\ge c(\gamma)\|\xi\|_1^2$ for all $\xi\in\mathbb{Y}(\gamma)$. We will use this to prove that $(Q(\gamma)\eta,\eta)$ is strictly positive. Write

$$\eta=aX_I(\bar{u})+bI'(\bar{u})+y$$

with $y \in \mathbb{Y}(\gamma)$. Then we have

$$\begin{array}{lclclcl} |a| & = & |(\eta, X_I(\bar{u}))| & = & |0 + (\eta, \partial_x(\bar{u} - \bar{u}_1))| & \leq & K\|\partial_x^{-1}R_3(\bar{u},\varepsilon)\|_0\|\eta\|_0, \\ |b| & = & |(\eta, I'(\bar{u}))| & = & |0 + (\eta, (\bar{u} - \bar{u}_1))| & \leq & K\|\partial_x^{-1}R_3(\bar{u},\varepsilon)\|_0\|\eta\|_0. \end{array}$$

This implies that there exists some $\tilde{K} > 0$ such that

$$\begin{aligned} (Q(\gamma)\eta,\eta) &= b^2(Q(\gamma)\bar{u},\bar{u}) + (Q(\gamma)y,y) \\ &\geq c(\gamma)\|y\|_0^2 - K^2\|\partial_x^{-1}R_3(\bar{u},\varepsilon)\|_0^2\|\eta\|_0^2\|Q(\gamma)\|_0, \|\bar{u}\|_0^2 \\ &\geq \left(c(\gamma) - \tilde{K}\varepsilon\sqrt{\gamma}\right)\|\eta\|_0^2 \\ &\geq \frac{1}{2}c(\gamma)\|\eta\|_0^2 \end{aligned}$$

for ε sufficiently small and γ bounded. Hence

$$\begin{aligned} (\tilde{Q}_1(\gamma)\eta,\eta) &\geq \frac{1}{2}c(\gamma)\|\eta\|_0^2 - 3K\|\partial_x^{-1}R_3(\bar{u},\varepsilon)\|_0\|\eta\|_0^2 \\ &\geq \frac{1}{4}c(\gamma)\|\eta\|_0^2 \end{aligned}$$

for ε sufficiently small. With Lemma A.4, this implies that

$$(\tilde{Q}_1(\gamma)\eta,\eta) \geq \tfrac{1}{4}\tilde{c}_1(\gamma)\|\eta\|_1^2 \tag{23}$$

for some $\tilde{c}_1(\gamma) > 0$ with $\tilde{c}_1(0) > 0$. We can conclude that $\bar{u}_1(\gamma)$ is a constrained minimum of the new Hamiltonian H_{new}.

Finally, we consider the problem of the uniformness in γ. The procedure of the implicit-function theorem can be continued until the invertibility of the linearization fails. Because the lower bound on $\tilde{Q}(\gamma)$ is also strictly positive in the limit for $\gamma \to 0$, there is a uniform (in γ) neighborhood around the MRE near 0 for which a unique solution of (18) exists. In other words, the limit for $\gamma \to 0$ of $\varepsilon_0(\gamma)$ and $K(\gamma)$ exist and $\varepsilon_0(0) = 0$. □

Remark 6. At the new relative equilibrium $\bar{u}_1(\gamma,\varepsilon)$, the adapted residual is of order $\varepsilon^2\|\bar{u}(\gamma)\|_0$. Indeed,

$$\begin{aligned} R(\bar{u}_1(\gamma,\varepsilon),\varepsilon) &= \dot{\gamma}\bar{u}_1'(\gamma,\varepsilon) - \partial H_{\text{new}}'(\bar{u}_1,\gamma,\varepsilon) - R(\bar{u},\varepsilon) - \varepsilon P(\bar{u}_1) \\ &= -2\varepsilon\gamma[\bar{u}_1'(\gamma,\varepsilon) - \bar{u}'(\gamma)] - \varepsilon[\bar{u}_1 - \bar{u}] - \lambda_1 X_I(\bar{u}_1). \end{aligned}$$

With $\|\bar{u}(\gamma) - \bar{u}_1(\gamma,\varepsilon)\|_1 = \mathcal{O}(\varepsilon\|\bar{u}(\gamma)\|_0)$, we immediately see that $\|R(\bar{u}_1(\gamma,\varepsilon),\varepsilon)\|_1 = \mathcal{O}(\varepsilon^2\|\bar{u}(\gamma)\|_0)$.

With the new Hamiltonian and the new minima, we can define a functional to "measure" the distance to the new minima:

$$L_{\text{new}}(u,\varepsilon) = H_{\text{new}}(u,\gamma,\varepsilon) - H_{\text{new}}(\bar{u}_1(\gamma,\varepsilon),\gamma,\varepsilon)$$

with $\gamma = I(u)$. The functional L_{new} is equivalent to the translation-invariant H^1-distance between u and $\bar{u}_1(\gamma,\varepsilon)$ with $\gamma = I(u)$.

LEMMA 3.2. *For every $\gamma > 0$, there exist $C(\gamma) \geq c(\gamma) > 0$, $\delta(\gamma) > 0$, and $\varepsilon_0(\gamma) > 0$, such that for all ε with $|\varepsilon| \leq \varepsilon_0(\gamma)$ and for all η with $(\eta, X_I(\bar{u}_1 + \eta)) = 0$ and $I(\bar{u}_1(\gamma) + \eta) = \gamma$, it holds that*

$$c(\gamma)\|\eta\|_1^2 \leq L_{\text{new}}(\bar{u}_1(\gamma,\varepsilon) + \eta, \varepsilon) \leq C(\gamma)\|\eta\|_1^2$$

as long as $\|\eta\|_1 \leq \delta(\gamma)$.

For every compact γ*-interval* $\mathcal{G}$, *there exist* $C \geq c > 0$, $\delta > 0$, *and* $\varepsilon_0 > 0$ *such that for all* ε *with* $|\varepsilon| \leq \varepsilon_0$, *for all* $\gamma \in \mathcal{G}$, *and for all* η *with* $(\eta, X_I(\bar{u}_1 + \eta)) = 0$ *and* $I(\bar{u}_1(\gamma) + \eta) = \gamma$, *it holds that*

$$c\|\eta\|_1^2 \leq L_{\text{new}}(\bar{u}_1(\gamma, \varepsilon) + \eta, \varepsilon) \leq C\|\eta\|_1^2$$

as long as $\|\eta\|_1 \leq \delta$.

Proof. In the proof of Lemma 3.1, it is shown that $\tilde{Q}_1(\gamma, \varepsilon) = D_\eta^2 L_{\text{new}}(\bar{u}_1(\gamma, \varepsilon), \varepsilon)$ is bounded from below on $\tilde{\mathbb{Y}}(\gamma) = \{\eta \in L^2 \mid (\eta, X_I(\bar{u}_1)) = 0, (\eta, X_I(\bar{u}_1)) = 0\}$ (see equation (23)) and that this lower bound remains strictly positive if $\gamma \longrightarrow 0$. As in the proof of Lemma 2.2, we can show that (23) implies that there is some $c(\gamma) > 0$ such that

$$L_{\text{new}}(\bar{u}_1(\gamma, \varepsilon) + \eta, \varepsilon) \geq c(\gamma)\|\eta\|_1^2$$

with $c(0) > 0$.

For the upper bounds, we rewrite L_{new}:

$$\begin{aligned} L_{\text{new}}(\bar{u}_1(\gamma, \varepsilon) + \eta, \varepsilon) &= H_{\text{new}}(\bar{u}_1(\gamma) + \eta, \gamma, \varepsilon) - \lambda_1(\gamma, \varepsilon) I(\bar{u}_1(\gamma) + \eta) \\ &\quad - [H_{\text{new}}(\bar{u}_1(\gamma), \gamma, \varepsilon) - \lambda_1(\gamma, \varepsilon) I(\bar{u}_1(\gamma))] \\ &= (H'_{\text{new}}(\bar{u}_1(\gamma)) - \lambda_1(\gamma, \varepsilon) I'(\bar{u}_1(\gamma)), \eta) \\ &\quad + \frac{1}{2}(\tilde{Q}_1(\gamma, \varepsilon)\eta, \eta) - \frac{1}{3}\int_0^{2\pi} \eta^3(x, t)\,dx. \end{aligned}$$

It is a straightforward calculation to derive the following estimates for $\|\eta\|_1 \leq \delta$:

$$\begin{aligned} &(Q_1(\gamma, \varepsilon)\eta, \eta) \leq \|\eta_x\|_0^2 + (2\|\bar{u}_1\|_\infty + |\lambda_1|)\|\eta\|_0^2 \leq C_1(\gamma)\|\eta\|_1^2, \\ &\int_0^{2\pi} \eta^3(x, t)\,dx \leq \|\eta\|_\infty \|\eta\|_0^2 \leq \sqrt{2\pi}4\pi^2\delta\|\eta\|_1^2. \end{aligned} \tag{24}$$

Note that $C_1(\gamma)$ is bounded from above if $\gamma \longrightarrow 0$ and that the second estimate in (24) does not depend on γ at all. Substitution of the estimates in (24) gives the upper bounds in the lemma. □

With L_{new}, we investigate the time behavior of the distance between a solution $u(t)$ of the damped KdV equation and the new relative equilibrium $\bar{u}_1(\gamma(t), \varepsilon)$. For a solution of the damped KdV equation, we define

$$\eta(t) = \Phi^I_{\phi(t)}(u(t)) - \bar{u}_1(\gamma(t), \varepsilon)$$

or, equivalently,

$$u(t) = \Phi^I_{\phi(t)}(\bar{u}_1(\gamma(t), \varepsilon) + \eta(t))$$

with $\phi(t)$ such that $\|\eta(t)\|^2 = \rho_1(u(t), \bar{u}_1(\gamma(t)), \varepsilon))$. (This implies the property that $(\eta, X_I(\bar{u}_1 + \eta)) = 0$.) With this definition, the dynamical equation for η is

$$\dot{\eta} + \dot{\bar{u}}_1 = \partial_x H'_{\text{new}}(\bar{u}_1 + \eta, \gamma, \varepsilon) - \dot{\phi} X_I(\bar{u}_1 + \eta) + [\varepsilon P(\bar{u}_1 + \eta) + R(\bar{u}, \varepsilon)].$$

Next, we give an estimate for the growth of L_{new} which is essentially better than the one we derived for L in the previous section.

PROPOSITION 3.3. *For every $\gamma(0)$, there exists a constant K (depending on $\gamma(0)$) such that for all $t \geq 0$ and $|\varepsilon| \leq \varepsilon_0$, it holds that*

$$L_{\text{new}}(u(t),\varepsilon) \leq K L_{\text{new}}(u(0),\varepsilon)e^{-2\varepsilon t} + K\varepsilon e^{-2\varepsilon t}$$

as long as $\|\eta(t)\|_1 \leq \delta$ (δ is given by Lemma 3.2).

Proof. To prove this proposition, we consider the time derivative of L_{new}:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}[L_{\text{new}}(u(t),\varepsilon)] &= (H'_{\text{new}}(\bar{u}_1+\eta,\gamma,\varepsilon) - \lambda_1 I'(\bar{u}_1+\eta), \dot{\bar{u}}_1 + \dot{\eta}) \\ &\quad + \dot{\gamma}\left[\frac{\partial}{\partial\gamma}H_{\text{new}}(\bar{u}_1+\eta,\gamma,\varepsilon)) - \frac{\partial}{\partial\gamma}H_{\text{new}}(\bar{u}_1,\gamma,\varepsilon))\right] \\ &= \text{(I)} + \text{(II)}.\end{aligned} \tag{25}$$

We will elaborate the terms (I) and (II) separately.

$$\begin{aligned}\text{(I)} &= -\varepsilon\left(H'_{\text{new}}(\bar{u}_1+\eta,\gamma,\varepsilon) - \lambda_1 I'(\bar{u}_1+\eta), \bar{u}_1 + \eta - \frac{1}{\varepsilon}R(\bar{u},\varepsilon)\right) \\ &= -\varepsilon\left(\tilde{Q}_1(\gamma,\varepsilon)\eta - \eta^2, \bar{u}_1 + \eta - \frac{1}{\varepsilon}R(\bar{u},\varepsilon)\right) \\ &= -2\varepsilon L_{\text{new}} + \frac{\varepsilon}{3}\int_0^{2\pi}\eta^3 + (\eta^2,[\varepsilon\bar{u}_1 - R(\bar{u},\varepsilon)]) \\ &\quad -\varepsilon\left(\eta, \tilde{Q}_1(\gamma,\varepsilon)\left[\bar{u}_1 - \frac{1}{\varepsilon}R(\bar{u},\varepsilon)\right]\right).\end{aligned} \tag{26}$$

In the same way as we showed that $\|\xi(t)\|_0 \leq 2\sqrt{2\gamma(t)}$ (see (8)), it can be seen that $\|\eta(t)\|_0 \leq 2\sqrt{2\gamma(t)}$. Just as in the proof of Proposition 2.3, this implies that $\int_0^{2\pi}\eta^3 \leq 8\pi\sqrt{\pi}\sqrt{\gamma}\|\eta\|_1^2 = \mathcal{O}(\varepsilon^2\sqrt{\gamma}+\varepsilon\sqrt{\gamma})$.

Furthermore, $\varepsilon\bar{u}_1 - R(\bar{u},\varepsilon) = \varepsilon(\bar{u}_1 - \bar{u}) - \dot{\bar{u}}$; hence $\|\varepsilon\bar{u}_1 - R(\bar{u},\varepsilon)\|_0 \leq \varepsilon\|\bar{u}_1 - \bar{u}\|_0 + \|\dot{\bar{u}}\|_0$. In this estimate, we use the explicit expression for the cnoidal waves in terms of the Jacobi elliptic functions to conclude that $\|\dot{\bar{u}}\|_0 = 2\varepsilon\gamma\|\bar{u}'(\gamma)\| = \mathcal{O}(\varepsilon\sqrt{\gamma})$.

Using these estimates, we see that the second and third terms in (26) are bounded by $\varepsilon K_4\sqrt{\gamma}\|\eta\|_1^2$ (K_4 is a constant independent of γ and of ε). We will give more attention to the estimate of the last term because it will improve the estimate of Proposition 2.3. By definition,

$$\varepsilon\left(\eta, \tilde{Q}_1(\gamma,\varepsilon)\left[\bar{u}_1 - \frac{1}{\varepsilon}R(\bar{u},\varepsilon)\right]\right) = \varepsilon(\eta, \tilde{Q}_1(\gamma,\varepsilon)[(\bar{u}_1 - \bar{u}) - 2\gamma\bar{u}'(\gamma)]).$$

Furthermore (see Lemma A.2(ii) in the appendix),

$$\begin{aligned}\tilde{Q}_1(\gamma,\varepsilon)\bar{u}'(\gamma) &= Q(\gamma)\bar{u}'(\gamma) + 2(\bar{u} - \bar{u}_1)\bar{u}'(\gamma) \\ &= \lambda'(\gamma)\bar{u}_1 + (\bar{u} - \bar{u}_1)[\lambda'(\gamma) + 2\bar{u}'(\gamma)].\end{aligned}$$

Hence

$$\begin{aligned}\varepsilon\left(\eta, \tilde{Q}_1(\gamma,\varepsilon)\left[\bar{u}_1 - \frac{1}{\varepsilon}R(\bar{u},\varepsilon)\right]\right) &= \varepsilon(\eta, [\tilde{Q}_1(\gamma,\varepsilon) - 2\gamma(\lambda'(\gamma) + 2\bar{u}'(\gamma))](\bar{u}_1 - \bar{u})) \\ &\quad - 2\varepsilon\gamma\lambda'(\gamma)(\eta,\bar{u}_1).\end{aligned}$$

By using the facts that $2(\eta, \bar{u}_1) = -\|\eta\|_0^2$ (this follows from $I(\bar{u}_1 + \eta) = I(\bar{u}_1)$), $\gamma\lambda'(\gamma) = \mathcal{O}(\sqrt{\gamma})$, and $\|\bar{u}(\gamma) - \bar{u}_1(\gamma, \varepsilon)\|_1 = \mathcal{O}(\varepsilon\sqrt{\gamma})$ (see Lemma 3.1), it follows that there exist constants K_5 and K_6 such that

$$\varepsilon\left(\eta, \tilde{Q}_1(\gamma)\left[\bar{u}_1 - \frac{1}{\varepsilon}R(\bar{u}, \varepsilon)\right]\right) \leq \varepsilon^2 K_5\sqrt{\gamma}\|\eta\|_1 + \varepsilon K_6\sqrt{\gamma}\|\eta\|_1^2.$$

Next, we estimate the second term of equation (25). Using the definitions of H_{new} and $R(\bar{u}, \varepsilon)$, it follows that

$$\begin{aligned}(\text{II}) &= \dot{\gamma}\left(\eta, \partial_x^{-1}\frac{\partial}{\partial\gamma}(2\varepsilon\gamma\bar{u}'(\gamma) + \varepsilon\bar{u}(\gamma)\right) \\ &= -2\varepsilon^2\gamma\|\eta\|_0\|\partial_x^{-1}(3\bar{u}'(\gamma) + 2\gamma\bar{u}''(\gamma))\|_0 \\ &\leq K_7\varepsilon^2\|\eta\|_1\sqrt{\gamma}.\end{aligned}$$

(Again, we use the explicit expression for $\bar{u}$ in terms of the Jacobi elliptic functions.)

Finally, using the fact that $\|\eta\|^2 \leq L_{\text{new}}/c$ (for $\|\eta\|_1 \leq \delta$), we can estimate $\frac{\mathrm{d}}{\mathrm{d}t}L_{\text{new}}$ by

$$\frac{\mathrm{d}}{\mathrm{d}t}L_{\text{new}} \leq -2\varepsilon L_{\text{new}} + K_0\varepsilon\sqrt{\gamma}L_{\text{new}} + K_0\varepsilon^2\gamma\sqrt{L_{\text{new}}}$$

for some constant K_0. Integrating this equation and applying Gronwall's lemma, we have that there exists some constant K such that

$$L_{\text{new}}(u(t), \varepsilon) \leq KL_{\text{new}}(\gamma(0), \eta(0), \varepsilon)e^{-2\varepsilon t} + K\varepsilon e^{-2\varepsilon t}. \qquad \square$$

The proof of Theorem 1.2 is a corollary of Proposition 3.3.

Proof of Theorem 1.2. From Lemmas 3.2 and 3.3 and the fact that $\|\eta(t)\|^2 = \rho_1(u(t), \bar{u}_1(\gamma(t)), \varepsilon))$, it follows that

$$\rho_1(u(t), \bar{u}_1(\gamma(t)), \varepsilon)) \leq K\rho_1(u(0), \bar{u}_1(\gamma(0)), \varepsilon))e^{-\varepsilon t} + K\varepsilon e^{-\varepsilon t}$$

if $\|\eta(0)\|_1$ is sufficiently small.

Now we use the fact that $\|\bar{u}_1(\gamma, \varepsilon) - \bar{u}(\gamma)\|_1 \leq \hat{K}\varepsilon\|\bar{u}\|_0$ for some constant $\hat{K}$ (see Lemma 3.1), which yields

$$\begin{aligned}\rho_1(u(t), \bar{u}(\gamma(t))) &\leq \rho_1(u(t), \bar{u}_1(\gamma(t)), \varepsilon)) + \|\bar{u}_1(\gamma(t), \varepsilon) - \bar{u}(\gamma(t))\|_1 \\ &\leq K\rho_1(u(0), \bar{u}_1(\gamma(0)), \varepsilon))e^{-\varepsilon t} + \tilde{K}\varepsilon e^{-\varepsilon t}(1 + \sqrt{2\gamma(0)}) \\ &\leq K\rho_1(u(0), \bar{u}(\gamma(0)))e^{-\varepsilon t} + \tilde{K}\varepsilon e^{-\varepsilon t}(1 + \sqrt{2\gamma(0)}) \\ &\quad + K\|\bar{u}_1(\gamma(0), \varepsilon) - \bar{u}(\gamma(0))\|_1 e^{-\varepsilon t} \\ &\leq K\rho_1(u(0), \bar{u}(\gamma(0)))e^{-\varepsilon t} + \bar{K}\varepsilon e^{-\varepsilon t}\end{aligned}$$

for some constant $\bar{K}$. This completes the proof of Theorem 1.2. □

Appendix. Some properties of the unperturbed KdV equation and cnoidal waves. As indicated in [4], the cnoidal waves are constrained minima of the Hamiltonian on level sets of I.

LEMMA A.1. *Let* $\gamma > 0$. *For all* 2π*-periodic functions* u *with mean value zero that satisfy* $I(u) = \gamma$, *it holds that*

$$H(u) \geq H(\bar{u}(\gamma)).$$

If $H(u) = H(\bar{u}(\gamma))$ and $I(u) = \gamma$, then u is a cnoidal wave with 2π as a minimal period. Explicitly, we have

$$(I(u) = \gamma \wedge H(u) = H(\bar{u}(\gamma))) \Rightarrow \exists_{\phi \in \mathbb{R}}[u = \Phi^I_\phi(\bar{u}(\gamma))].$$

Furthermore, all cnoidal waves $\bar{u}(\gamma)$ satisfy the Euler–Lagrange equation

$$\partial_x(H'(u) - \lambda I'(u)) = \partial_x(\bar{u}_{xx} + \bar{u}^2 + \lambda \bar{u}) = 0, \tag{27}$$

and the cnoidal wave $\bar{u}(\gamma)$ is an unconstrained minimum of the modified KdV Hamiltonian $H_\gamma(u) = H(u) - \lambda(\gamma)I(u)$.

Proof. Let $\gamma > 0$. We start by proving that the minimum of H on the level set $I = \gamma$ exists. For this proof, we show that H is a weakly lower semicontinuous (w.l.s.c.) functional that is coercive on the level set $I = \gamma$.

(i) First, we show coerciveness. For all $u \in H^1_{\text{per},0}$ with $I(u) = \gamma$, it holds that

$$\left| \int_0^{2\pi} u^3(x)dx \right| \leq \|u\|_\infty \int_0^{2\pi} u^2(x)dx \leq \sqrt{2\pi}\|u\|_1 2\gamma \tag{28}$$

(we used the Poincaré inequalities (4) and (5)). This gives

$$H(u) \geq \frac{1}{2}\int_0^{2\pi} u_x^2(x)dx - \frac{1}{3}\left|\int_0^{2\pi} u^3(x)dx\right| \geq \|u\|_1 \left[\frac{1}{2}\|u\|_1 - \frac{\sqrt{2\pi}}{3}\gamma\right]. \tag{29}$$

The last expression grows to infinity for $\|u\|_1 \to \infty$.

(ii) Next, we prove weak lower semicontinuity. The norm is a w.l.s.c. functional; hence $\int_0^{2\pi} u_x^2(x)dx = \|u\|_1^2$ is w.l.s.c. The term $\int_0^{2\pi} u^3(x)dx$ is a functional that is even weakly continuous. To prove this, we use the fact that $H^1_{\text{per},0}$ is embedded in C^0 and the embedding operator is strongly continuous. (See [30, p. 82].) Hence if the sequence $(u_n)_{n\in\mathbb{N}}$ converges weakly to u in $H^1_{\text{per},0}$, then this sequence is uniformly convergent to u. This implies that

$$\lim_{n\to\infty} \int_0^{2\pi} u_n^3(x)dx = \int_0^{2\pi} u^3(x)dx, \tag{30}$$

which shows that the functional $\int_0^{2\pi} u^3(x)dx$ is weakly continuous. In the same way, it is proved that the set $\{u \in H^1_{\text{per},0} \mid \int_0^{2\pi} u^2(x)dx = 2\gamma\}$ is weakly closed.

A coercive w.l.s.c. functional defined on a (sequentially) weakly closed set attains its infimum on this set. (See [6, §6.1] and [31, Chap. 38].) This completes the proof that H has a minimum on the level set with $I = \gamma$.

From variational calculus, it follows that this minimum satisfies the Euler–Lagrange equation

$$H'(u) - \lambda I'(u) = \alpha \tag{31}$$

for some Lagrange multipliers λ and α. Using properties of elliptic functions (see, e.g., [8] or [10]), it follows that the cnoidal waves with minimal period are unique solutions of such a equation with minimal value of H. □

In several places, we use properties of the operators $Q(\gamma)$ and $\tilde{Q}(\gamma)$ and the cnoidal waves. We list some important ones.

LEMMA A.2. *The operators $Q(\gamma)$ and $\tilde{Q}(\gamma)$ and the cnoidal waves satisfy the following properties:*

(i) $Q(\gamma)\bar{u}_x(\gamma) = \tilde{Q}(\gamma)\bar{u}_x(\gamma) = 0$;
(ii) $Q(\gamma)\bar{u}'(\gamma) = \tilde{Q}(\gamma)\bar{u}'(\gamma) = \lambda'(\gamma)\bar{u}(\gamma) + \alpha'(\gamma)\mathbf{1}$ *and* $\alpha'(\gamma) = \frac{1}{\pi}$;
(iii) $\tilde{Q}(\gamma)\mathbf{1} = -2\bar{u} - \lambda(\gamma)\mathbf{1}$;
(iv) $\lambda'(\gamma) < 0$;
(v) $\operatorname{range}[\tilde{Q}(\gamma)] = \{u \in H^1_{\text{per}} \mid (X_I(\bar{u}), u) = 0\}$.

Proof. (i) The translation invariance of both H and I implies that $\tilde{Q}(\gamma)X_I(\bar{u}) = 0$; hence $\tilde{Q}(\gamma)\bar{u}_x(\gamma) = 0$. Also, because $\bar{u}_x \in H^1_{\text{per},0}$, $Q(\gamma)\bar{u}_x(\gamma) = 0$.

(ii) Differentiation of the Euler–Lagrange equation for the cnoidal waves, i.e.,

$$-\bar{u}_{xx}(\gamma) - \bar{u}^2 - \lambda(\gamma)\bar{u}(\gamma) - \alpha(\gamma) = 0,$$

with respect to γ shows statement (ii) of the lemma. Integration of this Euler–Lagrange equation yields $\alpha(\gamma) = \frac{1}{\pi}\gamma$ and hence $\alpha'(\gamma) = \frac{1}{\pi}$.

(iii) The equation follows immediately from the definition of $\tilde{Q}(\gamma)$.

(iv) The proof of this property can be found in [11].

(v) $\tilde{Q}(\gamma)$ is a self-adjoint operator, and in Lemma A.3 it is proved that $X_I(\bar{u})$ is the only eigenvector with eigenvalue 0. □

The cnoidal waves are minima of the modified KdV Hamiltonian H_γ. This implies that $Q(\gamma)$ is positive definite on $\mathbb{Y}(\gamma)$. In Lemma A.3, we show that a slightly stronger property holds.

LEMMA A.3. *The operator $Q(\gamma)$ is strictly positive definite on $\mathbb{Y}(\gamma)$. To be explicit, there is some $c_1(\gamma) > 0$ such that*

$$(Q(\gamma)y, y) \geq c_1(\gamma)\|y\|_1^2 \qquad \textit{for all } y \in \mathbb{Y}(\gamma) \tag{32}$$

and $\lim_{\gamma\to 0} c_1(\gamma) > 0$.

Proof. To prove this boundedness from below, we consider the eigenvalues of $Q(\gamma)$. These eigenvalues form a monotonically nondecreasing sequence in $\mathbb{R}$:

$$\lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \qquad \text{with } \lim_{n\to\infty} \lambda_n = \infty. \tag{33}$$

For the smallest eigenvalue λ_0, it holds that

$$\lambda_0 = \min\left\{(Q(\gamma)\xi, \xi) \mid \xi \in H^1_{\text{per},0}, \int_0^{2\pi} \xi^2 = 1\right\}. \tag{34}$$

This number is negative because

$$(Q(\gamma)\bar{u}(\gamma), \bar{u}(\gamma)) = -\int_0^{2\pi} \bar{u}(\gamma)^3 < 0. \tag{35}$$

The last inequality is based on the fact that $I(\bar{u}) = I(-\bar{u})$ and $H(\bar{u}) < H(-\bar{u})$ because of the minimality of $H(\bar{u})$ on the level set with $I = \gamma$. This implies that $-\frac{1}{3}\int_0^{2\pi}\bar{u}^3 < \frac{1}{3}\int_0^{2\pi}\bar{u}^3$; in other words, $-\frac{2}{3}\int_0^{2\pi}\bar{u}^3 < 0$.

The translation invariance of H and I implies that $Q(\gamma)\bar{u}_x = 0$; hence $Q(\gamma)$ has at least one eigenvalue which equals zero for all $\gamma > 0$.

At $\gamma = 0$, the operator $Q(0)$ is equal to $-(D_{xx} + Id)$. The eigenvalues of this operator on $H^1_{\text{per},0}$ are (k^2-1), $k \in \mathbb{N}$. All these eigenvalues are double. The continuity of $Q(\gamma)$ in γ implies that for $\gamma > 0$ in a neighborhood of 0, it holds that

$$\lambda_0(\gamma) < 0, \qquad \lambda_1(\gamma) = 0, \qquad \lambda_2(\gamma) > 0. \tag{36}$$

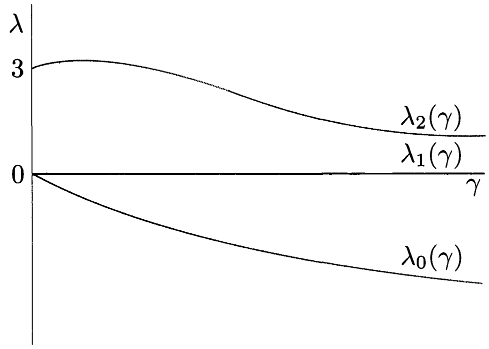


FIG. 4. *Sketch of the behavior of the eigenvalues* $\lambda_0(\gamma)$, $\lambda_1(\gamma)$, *and* $\lambda_2(\gamma)$.

See also Figure 4.

We prove $\lambda_2(\gamma) > 0$ for all $\gamma \geq 0$ by using a contradiction argument and $\tilde{Q}(\gamma)$, the extension of $Q(\gamma)$ on H^1_{per}. Assume that there is some $\gamma_0 > 0$ such that $\lambda_2(\gamma_0) = 0$. Then the operator $\tilde{Q}(\gamma)$ has a double eigenvalue zero. The differential equation for the eigenvalues and 2π-periodic eigenvectors of $\tilde{Q}(\gamma)$ is called Lamé's equation; see, e.g., [24]. It follows from Sturm–Liouville theory that zero is the second or the third eigenvalue of this equation because $\bar{u}_x$ is an eigenvector at zero. In [2], it is proved that the first three eigenvalues of this equation are single. Hence zero has to be a single eigenvalue of $\tilde{Q}(\gamma)$. This contradicts our assumption that $\lambda_2(\gamma_0) = 0$ and implies that $\lambda_2(\gamma_0) > 0$ for all $\gamma_0 \geq 0$. We have seen that $\lambda_2(0) = 3$; hence on every compact γ-interval, there is a positive lower bound for $\lambda_2(\gamma)$.

Using ideas similar to those of [23], this behavior of the eigenvalues of $Q(\gamma)$ implies that there is a $c_0(\gamma) > 0$ such that

$$\int_0^{2\pi} [y_x^2(x) - (2\bar{u}(x) - \lambda)y^2(x)]dx = (Q(\gamma)y, y) \geq c_0(\gamma)\|y\|_0^2 \tag{37}$$

for all $y \in \mathbb{Y}(\gamma)$. In Lemma A.4, we will prove that (37) implies that

$$(Q(\gamma)y, y) \geq c_1(\gamma)\|y\|_1^2 \qquad \text{for all } y \in \mathbb{Y}(\gamma)$$

with $c_1(\gamma) = \frac{c_0(\gamma)}{c_0(\gamma)+\|2\bar{u}(\gamma)+\lambda(\gamma)\|_\infty}$. □

In the proof of Lemma A.3, we used the fact that if $(Q(\gamma)y, y)$ is bounded from below in the L^2-norm for all $y \in \mathbb{Y}(\gamma)$, then it is bounded from below in the H^1-norm as well. This property can be concluded immediately from the following lemma.

LEMMA A.4. *Let* $p(x)$ *be a continuous function on* $[-\pi, \pi]$. *If for some* $\xi \in H^1_{\text{per},0}$ *it holds that*

$$\int_0^{2\pi} [\xi_x^2(x) + p(x)\xi^2(x)]dx \geq c\|\xi\|_0^2, \tag{38}$$

then

$$\int_0^{2\pi} [\xi_x^2(x) + p(x)\xi^2(x)]dx \geq c_1\|\xi\|_1^2 \tag{39}$$

with $c_1 = \frac{c}{c+\|p\|_\infty}$.

Proof. Assume that (38) holds for some $\xi \in H^1_{\mathrm{per},0}$. By rewriting the integral in equation (40), we see that

$$\begin{aligned}\int_0^{2\pi}[\xi_x^2(x)+p(x)\xi^2(x)]dx &= c_1\int_0^{2\pi}\xi_x^2(x)dx+(1-c_1)\int_0^{2\pi}\xi_x^2(x)dx\\ &\quad+\int_0^{2\pi}p(x)\xi^2(x)dx.\end{aligned}\tag{40}$$

Now we use inequality (38); it follows that

$$\begin{aligned}&\int_0^{2\pi}[\xi_x^2(x)+p(x)\xi^2(x)]dx\\ &\quad\geq c_1\int_0^{2\pi}\xi_x^2(x)dx+(1-c_1)c\|\xi\|_0^2-c_1\|p\|_\infty\|\xi\|_0^2\\ &\quad= c_1\int_0^{2\pi}\xi_x^2(x)dx+c\|\xi\|_0^2-c_1(c+\|p\|_\infty)\|\xi\|_0^2\\ &\quad= c_1\|\xi\|_1^2. \qquad \square\end{aligned}\tag{41}$$

The last subject in this appendix is the operator $DF_0(\bar{u}(\gamma),\lambda(\gamma),\alpha(\gamma),0)$ as defined in the proof of Lemma 3.1.

LEMMA A.5. *The operator $DF_0(\bar{u}(\gamma),\lambda(\gamma),\alpha(\gamma),0)$ is injective and surjective, for all $\gamma > 0$.*

Proof. Let $\gamma > 0$. Define A_0 on $H^1_{\mathrm{per}}\times\mathbb{R}\times\mathbb{R}\times\mathbb{R}$ as

$$A_0 = DF_0(\bar{u}(\gamma),\lambda(\gamma),\alpha(\gamma),0) = \begin{pmatrix} \tilde{Q}(\gamma) & -\bar{u}(\gamma) & -\mathbf{1} & -\bar{u}_x(\gamma)\\ \bar{u}(\gamma) & 0 & 0 & 0\\ \mathbf{1} & 0 & 0 & 0\\ R(\bar{u}(\gamma),\varepsilon) & 0 & 0 & 0 \end{pmatrix}.$$

(i) First, we prove that A_0 is an injective map. Assume that $A_0(v,l,a,b)=0$ for some $(v,l,a,b)\in H^1_{\mathrm{per}}\times\mathbb{R}\times\mathbb{R}\times\mathbb{R}$; hence

$$0=\tilde{Q}(\gamma)v-l\bar{u}(\gamma)-a\mathbf{1}-b\bar{u}_x(\gamma),\tag{42}$$

$$0=(\bar{u}(\gamma),v),\tag{43}$$

$$0=(\mathbf{1},v),\tag{44}$$

$$0=(R(\bar{u}(\gamma),\varepsilon),v).\tag{45}$$

Taking the inner product of (42) with $\bar{u}_x$ yields $b\|\bar{u}_x\|_0^2=0$ and hence $b=0$. Write $v = c_1\bar{u}'(\gamma)+c_2\bar{u}_x+y$, with $y\in\tilde{\mathbb{Y}}(\gamma)$. This decomposition is unique because $(\bar{u}'(\gamma),\bar{u}) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\gamma}\|\bar{u}\|_0^2 = 1$ and hence $\bar{u}'(\gamma)\notin\tilde{\mathbb{Y}}(\gamma)$. From (43), it follows that $c_1=(v,\bar{u})=0$. Taking the inner product of (42) with v shows that $0=(\tilde{Q}(\gamma)v,v)=(\tilde{Q}(\gamma)y,y)$. Hence $y=0$ because $\tilde{Q}(\gamma)$ is strictly positive definite on $\tilde{\mathbb{Y}}(\gamma)$. From (45), it follows that $0=c_2(R(\bar{u}(\gamma),\varepsilon),\bar{u}_x)$. Because of the assumption that $(R(\bar{u}(\gamma),\varepsilon),\bar{u}_x)\neq 0$, this implies that $c_2=0$ and hence $v=0$. Finally, substituting $v=0$ and $b=0$ in (42) yields $l\bar{u}(\gamma)+a\mathbf{1}=0$. Because $\bar{u}(\gamma)$ and $\mathbf{1}$ are linearly independent, this implies that $l=0$ and $a=0$.

(ii) Next, we show that A_0 is surjective. Assume that $A_0(v,l,a,b) = (w,m,c,d)$ for some $(v,l,a,b), (w,m,c,d) \in H^1_{\text{per}} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$; hence

$$w = \tilde{Q}(\gamma)v - l\bar{u}(\gamma) - a\mathbf{1} - b\bar{u}_x(\gamma), \tag{46}$$

$$m = (\bar{u}(\gamma), v), \tag{47}$$

$$c = (\mathbf{1}, v), \tag{48}$$

$$d = (R(\bar{u}(\gamma), \varepsilon), v). \tag{49}$$

Taking the inner product of (46) with $\bar{u}_x$ yields $b\|\bar{u}_x\|_0^2 = (w, \bar{u}_x)$. This defines b because $\|\bar{u}_x\|_0^2 \neq 0$. Take the inner product of (46) with $\mathbf{1}$ and use Lemma A.2(iii), which yields $2\pi a = -(w,\mathbf{1}) - 2m - \lambda c$. This defines a. Take the inner product of (46) with $\bar{u}'(\gamma)$ and use Lemma A.2(ii), which yields $l = -(w, \bar{u}'(\gamma)) - b(\bar{u}_x, \bar{u}'(\gamma)) + \lambda'(\gamma)m + \alpha'(\gamma)c$. This defines l. From Lemma A.2(v), it follows that $\tilde{Q}(\gamma)$ is invertible on $\{\bar{u}_x\}^{\perp}$; hence

$$v = \tilde{Q}(\gamma)^{-1}[w + l\bar{u} + a\mathbf{1} + b\bar{u}_x] + f\bar{u}_x.$$

The value of f follows from (49). □

REFERENCES

[1] V. I. Arnold, *Dynamical Systems* III, Encyclopaedia of Mathematical Sciences, vol. 3, Springer-Verlag, Berlin, Heidelberg, 1988.

[2] F. M. Arscott, *Periodic Differential Equations*, Monographs in Pure and Applied Mathematics, vol. 66, Pergamon Press, Oxford, 1964.

[3] T. B. Benjamin, *The stability of solitary waves*, Proc. Roy. Soc. London Ser. A, 328 (1972), pp. 153–183.

[4] ———, *Lectures on nonlinear wave motion*, in Lectures in Applied Mathematics, vol. 15, American Mathematical Society, Providence, RI, 1974, pp. 3–47.

[5] T. B. Benjamin, J. L. Bona, and J. J. Mahony, *Model equations for long waves in nonlinear dispersive systems*, Philos. Trans. Roy. Soc. London Ser. A, 272 (1972), pp. 47–78.

[6] M. Berger, *Nonlinearity and Functional Analysis*, Pure and Applied Mathematics, vol. 74, Academic Press, New York, 1977.

[7] J. L. Bona and R. Smith, *The initial value problem for the Korteweg-de Vries equation*, Philos. Trans. Roy. Soc. London Ser. A, 278 (1975), pp. 555–604.

[8] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Scientist*, Grundlehren der mathematischen Wissenschaften 67, 2nd ed., Springer-Verlag, Berlin, 1971 (revised version).

[9] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, 1985.

[10] G. Derks, *Coherent structures in the dynamics of perturbed Hamiltonian systems*, Ph.D. thesis, University of Twente, Twente, The Netherlands, 1992.

[11] G. Derks and S. A. van Gils, *On the uniqueness of traveling waves in perturbed Korteweg–de Vries equations*, Japan J. Indust. Appl. Math., 10 (1993), pp. 413–430.

[12] G. Derks, D. Lewis, and T. Ratiu, *Approximations with curves of relative equilibria in Hamiltonian systems with dissipation*, Nonlinearity, 8 (1995), pp. 1087–1113.

[13] G. Derks and T. P. Valkering, *Approximation in a damped Hamiltonian system by successive relative equilibria*, Japan J. Indust. Appl. Math., 9 (1992), pp. 141–161.

[14] J. M. Ghidaglia, *Weakly damped forced Korteweg–de Vries equations behave like a finite dimensional dynamical system in the long time*, J. Differential Equations, (1988), pp. 369–390.

[15] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry* I, J. Funct. Anal., 74 (1987), pp. 160–197.

[16] E. van Groesen, F. P. H. van Beckum, and T. P. Valkering, *Decay of travelling waves in dissipative Poisson systems*, Z. Angew. Math. Phys., 41 (1990), pp. 501–523.

[17] D. D. Holm, J. E. Marsden, T. Ratiu, and A. Weinstein, *Nonlinear stability of fluid and plasma equiliberia*, Phys. Rep., 123 (1985), pp. 1–116.

[18] V. I. Karpman and E. M. Maslov, *A perturbation theory for the Korteweg–de Vries equation*, Phys. Lett., 60A (1977), pp. 307–308.

[19] D. J. KAUP AND A. C. NEWELL, *Solitons as particles, oscillators, and in slowly changing media: a singular perturbation theory*, Proc. Roy. Soc. London Ser. A, 361 (1978), pp. 413–446.

[20] C. J. KNICKERBOCKER AND A. C. NEWELL, *Shelves and the Korteweg–de Vries equation*, J. Fluid Mech., 98 (1980), pp. 803–818.

[21] D. J. KORTEWEG AND G. DEVRIES, *On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves*, Philos. Magazine, 39 (1895), pp. 422–443.

[22] N. R. LEBOVITZ AND A. NEISHTADT, *Slow evolution in perturbed Hamiltonian systems*, Stud. in Appl. Math., 92 (1994), pp. 127–144.

[23] J. H. MADDOCKS AND R. L. SACHS, *On the stability of KdV multi-solitons*, Comm. Pure Appl. Math., 46 (1993), pp. 867–901.

[24] W. MAGNUS AND S. WINKLER, *Hill's Equation*, Interscience Tracts in Pure and Applied Mathematics, vol. 20, Interscience Publishers, New York, 1966.

[25] J. W. MILES, *The Korteweg–de Vries equation: A historical essay*, J. Fluid Mech., 106 (1981), pp. 131–147.

[26] P. J. OLVER, *Applications of Lie groups to differential equations*, Springer-Verlag, New York, 1986.

[27] A. C. SCOTT, F. Y. F. CHU, AND D. W. MCLAUGHLIN, *The soliton: A new concept in applied science*, Proc. IEEE, 61 (1973), pp. 1443–1483.

[28] J. C. SIMO, D. LEWIS, AND J. E. MARSDEN, *Stability of relative equilibria, part* I: *The reduced energy momentum method*, Arch. Rational Mech. Anal., 115 (1991), pp. 15–60.

[29] J. C. SIMO, T. A. POSBERGH, AND J. E. MARSDEN, *Stability of relative equilibria, part* II: *Application to nonlinear elasticity*, Arch. Rational Mech. Anal., 115 (1991), pp. 60–100.

[30] S. L. SOBOLEV, *Applications of Functional Analysis in Mathematical Physics*, Americal Mathematical Society, Providence, RI, 1963.

[31] E. ZEIDLER, *Nonlinear Functional Analysis and Its Applications, Part* III: *Variational Methods and Optimization*, Springer–Verlag, Berlin, New York, Heidelberg, 1985.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1448–1453, September 1996

014

A REFINED WIENER–LEVINSON METHOD IN FREQUENCY ANALYSIS*

K. PAN†

Abstract. This paper is concerned with the problem of determining unknown frequencies $\omega_1, \ldots, \omega_I$, using the first N observed values of a discrete-time signal $\{x(m)\}_{m=0}^{N-1}$ arising from a continuous waveform that is the superposition of a finite number of sinusoidal waves with well-defined frequencies ω_j, $j = 1, 2, \ldots, I$. In [K. Pan and E. B. Saff, *J. Approx. Theory*, 71 (1992), pp. 239–251] (see also [W. B. Jones, O. Njåstad, W. J. Thron, and H. Waadeland, *J. Comput. Appl. Math.*, 46 (1993), pp. 217–228]), we proved that unknown frequencies ω_j, $j = 1, 2, \ldots, I$, in a periodic discrete-time signal can be determined by zeros of Szegö polynomials with respect to some distribution function by using the first N samples with a rate of convergence of $1/N$. We introduce a refined way to obtain a rate of convergence of $1/N^p$ by using about pN samples of the signals, where p is any given positive integer.

Key words. frequency analysis, orthogonal polynomials

AMS subject classifications. 33C45, 40A15, 41A21

1. Introduction.

We denote a doubly infinite sequence $x = \{x(m)\}_{-\infty}^{\infty}$ of real numbers as a signal. We consider signals of the form

$$x(m) = \sum_{j=-I}^{I} \alpha_j e^{i\omega_j m}, \quad x(0) \neq 0, \tag{1.1}$$

where

$$\alpha_0 \geq 0, \quad \alpha_{-j} = \bar{\alpha}_j, \quad \omega_{-j} = -\omega_j, \quad 0 = \omega_0 < \omega_1 < \cdots < \omega_I < \pi. \tag{1.2}$$

The problem of determining the frequencies ω_j from the first N samples $\{x(m)\}_{m=0}^{N-1}$ has important applications to science and engineering. Recently, a method for solving this problem was introduced by Jones, Njåstad, and Saff based upon the techniques of Wiener and Levinson. The starting points for this method are the autocorrelation coefficients

$$\mu_k^{(N)} := \sum_{m=0}^{N-1-k} x(m)\overline{x(m+k)}, \quad k = 0, 1, 2, \ldots, \quad \text{and } \mu_{-k} = \mu_k.$$

They form a positive definite Hermitian sequence (cf. [JNS]); that is,

$$D_n^{(N)} := \det(\mu_{i-j}^{(N)})_0^n > 0, \quad n = 0, 1, 2, \ldots.$$

We consider the monic Szegö polynomials $\phi_{n,N}(z), n \geq 1$, as follows,

$$\phi_{n,N}(z) := \frac{1}{D_{n-1}^{(N)}} \begin{vmatrix} \mu_0^{(N)} & \mu_{-1}^{(N)} & \cdots & \mu_{-n}^{(N)} \\ \mu_1^{(N)} & \mu_0^{(N)} & \cdots & \mu_{-n+1}^{(N)} \\ \vdots & \vdots & \cdots & \vdots \\ \mu_{n-1}^{(N)} & \mu_{n-2}^{(N)} & \cdots & \mu_{-1}^{(N)} \\ 1 & z & \cdots & z^n \end{vmatrix}.$$

* Received by the editors September 7, 1994; accepted for publication (in revised form) April 14, 1995.

† Department of Mathematics, Barry University, Miami Shores, FL 33161.

The following conjecture on the asymptotics of zeros of $\phi_{n,N}(z)$ was introduced in [JNS].

CONJECTURE (see [JNS]). *As $n \to \infty$ and $N \to \infty$, the $2I + L$ zeros of $\phi_{n,N}(z)$ of largest modulus approach the point $e^{i\omega_j}$, $j = \pm 1, \pm 2, \ldots, \pm I$, and also* 1 *if $L = 1$. Here $L = 1$ if $\alpha_0 > 0$ and $L = 0$ if $\alpha_0 = 0$.*

This conjecture has recently been verified by the results of [PS] and [JNTW] for n fixed, $n \geq 2I + L$, and $N \to \infty$. In [PS] (see also [JNTW]), we proved that the rate of convergence is $\mathcal{O}(1/N)$ for $n = 2I + L$. That is, let $\alpha_{N,j,n}$ denote the zero of $\phi_{n,N}(z)$ that is closest to $e^{i\omega_j}$; then $|\alpha_{N,j,n} - e^{i\omega_j}| = \mathcal{O}(1/N)$ as $N \to \infty$ and $1/N$ is the best possible. Thus, for N large enough, those zeros of $\phi_{n,N}(z)$ can be used to approximate the unknown frequencies.

From numerical experiments (cf. [JNS]), we can see that it may take 1000 samples to get only two significant digits. We want to find some alternative ways to improve this method.

In this paper, we introduce a new idea to approximate the unknown frequencies by creating a "window" to compute the autocorrelation coefficients. This method will give us an asymptotic rate of convergence of $1/N^p$ by using about pN samples of x for any positive integer $p \geq 1$.

The outline of the paper is as follows. In §2, we state our main results, and the proofs of these results are given in §3. The numerical results can be found in §4.

2. Main theorems. For convenience we let $\beta_j := e^{i\omega_j}, j = -I, \ldots, I$. Set

$$\mu_k := \sum_{-I}^{I} |\alpha_j|^2 \beta_j^k, \quad k = 0, \pm 1, \ldots.$$

Then the sequence $\{\mu_k\}_{-\infty}^{\infty}$ is a positive $(2I + L)$-definite hermitian sequence (cf. [PS]). This means that

$$b_n > 0 \quad \text{for} \quad 0 \leq n \leq 2I + L - 1, \qquad b_{2I+L} = 0,$$

where

$$b_n := \det(\mu_{i-j})_0^n.$$

For any integers $N > 0$ and $p \geq 1$, we define $a(N, p, k)$, $k = 0, \ldots, pN$, as follows:

$$\left(\frac{1 - R^{N+1}}{1 - R}\right)^p = (1 + R + \cdots + R^N)^p = \sum_{k=0}^{pN} a(N, p, k) R^k.$$

Also, we define

$$\nu_m^{(N,p)} := \sum_{k=0}^{pN} a(N, p, k) x(k) \overline{x(k + m)}, \nu_{-m}^{(N,p)} = \overline{\nu_m^{(N,p)}}, \quad m = 0, 1, 2, \ldots.$$

Although we cannot prove that $\{\nu_m^{(N,p)}\}$ is a positive definite hermitian sequence, we can prove the following.

THEOREM 2.1. *For $N \to \infty$, we have*

$$c_n^{(N,p)} := \det(\nu_{i-j}^{(N,p)})_{i,j=0}^n > 0, \quad n = 0, 1, \ldots, 2I + L - 1.$$

Furthermore, we have

$$\frac{1}{(N+1)^{p(n+1)}}\, c_n^{(N,p)} = b_n + \mathcal{O}\left(\frac{1}{N^p}\right), \quad n = 0, 1, \ldots, 2I+L-1.$$

Next, set

$$\psi_{2I+L,N,p}(z) := \frac{1}{c_{2I+L-1}^{(N,p)}} \begin{vmatrix} \nu_0^{(N,p)} & \nu_{-1}^{(N,p)} & \cdots & \nu_{-(2I+L)}^{(N,p)} \\ \nu_1^{(N,p)} & \nu_0^{(N,p)} & \cdots & \nu_{-(2I+L)+1}^{(N,p)} \\ \vdots & \vdots & \cdots & \vdots \\ \nu_{(2I+L)-1}^{(N,p)} & \nu_{(2I+L)-2}^{(N,p)} & \cdots & \nu_{-1}^{(N,p)} \\ 1 & z & \cdots & z^{2I+L} \end{vmatrix}.$$

THEOREM 2.2. *For any integer $p \geq 1$, we have*

$$\lim_{N\to\infty} \psi_{2I+L,N,p}(z) = \phi_{2I+L}(z) := (z-1)^L \sum_{j=1}^{I} (z-\beta_j)(z-\beta_{-j}), \quad z \in C.$$

The convergence is uniform on compact subsets of $\mathbf{C}$. *More precisely, we have for each compact set $K \subset \mathbf{C}$,*

$$|\psi_{2I+L,N,p}(z) - \phi_{2I+L}(z)| \leq \frac{A}{N^p}, \quad z \in K, \quad N \to \infty,$$

where A is a constant that depends on K.

COROLLARY 2.3. *For each N large, let $\beta_{N,j,p}$ denote the zero of $\psi_{2I+L,N,p}(z)$ that is closest to β_j. Then for $j = \pm 1, \pm 2, \ldots, \pm I$ and $j = 0$ if $L = 1$,*

$$|\beta_{N,j,p} - \beta_j| = \mathcal{O}\left(\frac{1}{N^p}\right), \quad N \to \infty.$$

Remark 1. The rate of convergence in the corollary is the best possible. This can be seen from the following example for the case when $\alpha_0 = 0$.

Let

$$x(m) := \beta_1^m + \beta_{-1}^m = 2\cos(\pi m/2),$$

where $\beta_1 = i$, and $\phi_2(z) = z^2 + 1$. On computing the moments $\nu_k^{(N,p)}$ and using the determinant representation for the orthogonal polynomial $\psi_{2,N,p}(z)$, we find

$$\psi_{2,N,p}(z) = z^2 + 1 \quad \text{for } N \text{ even}$$

and

$$\psi_{2,N,p}(z) = z^2 + 1 + \frac{1}{N^p} \quad \text{for } N \text{ odd}.$$

Thus the zeros of $\psi_{2,N,p}(z)$ approach $\pm i$ with exact rate $1/N^p$.

Remark 2. From this example, we can see that the zeros of $\psi_{2I+L,N,p}(z)$ do not lie in $|z| < 1$ since $\{\nu_m^{(N,p)}\}$ is not a positive definite hermitian sequence.

3. Proofs of theorems. Before we give the proof of Theorem 2.1, we need the following lemma.

LEMMA 3.1. *For $N \to \infty$, we have*

$$\frac{1}{(N+1)^p}\nu_m^{(N,p)} = \mu_m + \mathcal{O}\left(\frac{1}{N^p}\right), \quad m = 0, \quad \pm 1, \ldots.$$

Proof. Notice that

$$\begin{aligned}
\nu_m^{(N,p)} &= \sum_{k=0}^{pN} a(N,p,k)x(k)\overline{x(k+m)} = \sum_{k=0}^{pN} a(N,p,k) \sum_{j=-I}^{I} \alpha_j \beta_j^k \sum_{l=-I}^{I} \overline{\alpha}_l \overline{\beta}_l^{k+m} \\
&= \sum_{j,l=-I}^{I} \alpha_j \overline{\alpha}_l \overline{\beta}_l^m \sum_{k=0}^{pN} a(N,p,k)\beta_j^k \overline{\beta}_l^k \\
&= \sum_{j=-I}^{I} |\alpha_j|^2 \overline{\beta}_j^m \sum_{k=0}^{pN} a(N,p,k) + \sum_{j \neq l} \alpha_j \overline{\alpha}_l \overline{\beta}_l^m \sum_{k=0}^{pN} a(N,p,k)\beta_j^k \overline{\beta}_l^k \\
&= \sum_{j=-I}^{I} |\alpha_j|^2 \overline{\beta}_j^m (N+1)^p + \sum_{j \neq l} \alpha_j \overline{\alpha}_l \overline{\beta}_l^m \left[\frac{1-(\beta_j\overline{\beta}_l)^{N+1}}{1-\beta_j\overline{\beta}_l}\right]^p.
\end{aligned}$$

Thus we have

$$\frac{1}{(N+1)^p}\,\nu_m^{(N,p)} = \mu_m + \mathcal{O}\left(\frac{1}{N^p}\right). \qquad \square$$

Proof of Theorem 2.1. It follows from Lemma 3.1. □

Proof of Theorem 2.2. From Lemma 3.1, we have

$$\begin{aligned}
&\psi_{2I+L,N,p}(z) \\
&= \frac{1}{c_{2I+L-1}^{(N,p)}} \begin{vmatrix} \nu_0^{(N,p)} & \nu_{-1}^{(N,p)} & \cdots & \nu_{-(2I+L)}^{(N,p)} \\ \nu_1^{(N,p)} & \nu_0^{(N,p)} & \cdots & \nu_{-(2I+L)+1}^{(N,p)} \\ \vdots & \vdots & \cdots & \vdots \\ \nu_{(2I+L)-1}^{(N,p)} & \nu_{(2I+L)-2}^{(N,p)} & \cdots & \nu_{-1}^{(N,p)} \\ 1 & z & \cdots & z^{2I+L} \end{vmatrix} \\
&= \frac{1}{b_{2I+L-1} + \mathcal{O}(1/N^p)} \left\{ \begin{vmatrix} \mu_0 & \mu_{-1} & \cdots & \mu_{-(2I+L)} \\ \mu_1 & \mu_0 & \cdots & \mu_{-(2I+L)+1} \\ \vdots & \vdots & \cdots & \vdots \\ \mu_{(2I+L)-1} & \mu_{(2I+L)-2} & \cdots & \mu_{-1} \\ 1 & z & \cdots & z^{2I+L} \end{vmatrix} \right\} + \mathcal{O}(1/N^p) \\
&= \phi_{2I+L}(z) + \mathcal{O}(1/N^p),
\end{aligned}$$

where $\mathcal{O}(1/N^p)$ is uniform in z on any compact subset of $\mathbf{C}$. Here we use the fact that (see [PS])

$$\frac{1}{b_{2I+L-1}}\begin{vmatrix} \mu_0 & \mu_{-1} & \cdots & \mu_{-(2I+L)} \\ \mu_1 & \mu & \cdots & \mu_{-(2I+L)+1} \\ \vdots & \vdots & \cdots & \vdots \\ \mu_{(2I+L)-1} & \mu_{(2I+L)-2} & \cdots & \mu_{-1} \\ 1 & z & \cdots & z^{2I+L} \end{vmatrix} = \phi_{2I+L}(z).$$

This completes the proof of Theorem 2.2. □

Proof of Corollary 2.3. From the proof of Theorem 2.2, we obtain that

$$\psi_{2I+L,N,p}(\beta_j) = \mathcal{O}(1/N^p), \quad j = \pm 1, \dots, \pm I \text{ and } j = 0 \text{ if } L = 1.$$

As previously remarked, $\beta_{N,j,p} \to \beta_j$. Thus for $N \to \infty$, we have

$$|\beta_j - \beta_{N,j,p}| = \left| \frac{\psi_{2I+L,N,p}(\beta_j)}{\Pi_{s \neq j}(\beta_j - \beta_{N,s})} \right| = \mathcal{O}(1/N^p), \quad N \to \infty,$$

for $j = \pm 1, \dots, \pm I$ and $j = 0$ if $L = 1$. □

4. Numerical results. We can use Levinson's algorithm to compute $\psi_{2I+L,N,p}(z)$. After we find $\nu_0^{(N,p)}$, $\nu_1^{(N,p)}, \dots, \nu_{2I+L}^{(N,p)}$, we compute δ_0, E_0, δ_1, $E_1, \dots, \delta_{2I+L}$, E_{2I+L} successively. Initially, set

$$\delta_0 = 1, \qquad E_0 = \nu_0^{(N,p)}, \qquad \delta_1 = -\nu_1^{(N,p)}/\nu_0^{(N,p)}, \qquad q_0^{(1)} = \delta_1, \qquad q_1^{(1)} = 1.$$

Then for $k = 2, 3, \dots, 2I + L$, compute

$$E_{k-1} = \sum_{j=0}^{k-1} q_j^{(k-1)} \nu_{k-1-j}^{(N,p)},$$

$$\delta_k = -\frac{\sum_{j=0}^{k-1} q_j^{(k-1)} \nu_{j+1}^{(N,p)}}{E_{k-1}},$$

$$q_j^{(k)} = \delta_k q_{k-1}^{(k-1)} + q_{j-1}^{(k-1)}, \qquad j = 1, 2, \dots, k-1,$$

$$q_k^{(k)} = 1, \qquad q_0^{(k)} = \delta_k.$$

Finally,

$$E_k = \sum_{j=0}^{2I+L} q_j^{(k)} \nu_{2I+L-j}^{(N,p)}.$$

Then $\psi_{2I+L,N,p}(z) = \sum_{j=0}^{2I+L} q_j^{(2I+L)} z^j$.

Example 1. (See Table 1.)

$$x(m) = \beta_{-1}^m + \beta_1^m = 2\cos(m\pi/7), \quad \beta_1 = e^{i\pi/7} = 0.9009688680 + 0.4338837393i.$$

$$\beta_{N,1,p} := \text{ the zero of } \psi_{2,N,p}(z) \text{ that is closest to } \beta_1.$$

TABLE 1

	p	$\beta_{N,1,p}$	$\|\beta_{N,1,p} - \beta_1\|$
$N = 10$	1	0.6812315316 + 0.4115244315i	0.22087198933
	2	0.8056936270 + 0.4281280070i	0.095448933
	3	0.8897385565 + 0.4337942395i	0.11230668133
	4	0.9035357004 + 0.4338789879i	0.00256683679
	5	0.9022590749 + 0.4338825418i	0.00129020745
$N = 100$	1	0.8837853613 + 0.4336756199i	0.01718476698
	2	0.9003845124 + 0.4338834935i	0.0005843556517
	3	0.9009557144 + 0.4338837362i	0.00001315360037
	4	0.9009686365 + 0.4338837423i	$0.2315194376 \times 10^{-6}$
	5	0.9009688896 + 0.4338837430i	$0.21914607 \times 10^{-7}$

TABLE 2

	p	$\|\beta_{N,1,p} - \beta_1\|$	$\|\beta_{N,2,p} - \beta_2\|$
$N = 100$	3	0.0002508	0.001165689
	4	0.0001524	0.000173923
$N = 500$	2	0.000409366	0.0168823
	3	0.000010432	0.0001343

Remark 3. We can see from Example 1 that we need only 10 samples to get two significant digits by taking $p = 5$.

Example 2. (See Table 2.)

$$x(m) = 4\cos(m\pi/4) + 2\cos(m\pi/7),$$

$$\beta_1 = e^{i\pi/4} = 0.7071067810 + 07071067810i,$$

$$\beta_2 = e^{i\pi/7} = 0.9009688680 + 0.4338837393i,$$

$$\beta_{N,1,p} := \text{ the zero of } \psi_{4,N,p}(z) \text{ that is closest to } \beta_1,$$

$$\beta_{N,2,p} := \text{ the zero of } \psi_{4,N,p}(z) \text{ that is closest to } \beta_2.$$

Acknowledgment. The author thanks the referee for valuable remarks and suggestions concerning this research.

REFERENCES

[JNS] W. B. JONES, O. NJÅSTAD, AND E. B. SAFF, *Szegö polynomials associated with Wiener–Levinson filters*, J. Comput. Appl. Math., 32 (1990), pp. 387–406.

[JNTW] W. B. JONES, O. NJÅSTAD, W. J. THRON, AND H. WAADELAND, *Szegö polynomials applied to frequency analysis*, J. Comput. Appl. Math., 46 (1993), pp. 217–228.

[PS] K. PAN AND E. B. SAFF, *Asymptotics for zeros of the Szegö polynomials associated with trigonometric polynomial signals*, J. Approx. Theory, 71 (1992), pp. 239–251.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1454–1481, September 1996

FAMILIES OF ORTHOGONAL TWO-DIMENSIONAL WAVELETS*

PETER MAASS†

Abstract. We construct orthonormal wavelet bases of $L^2(\mathbb{R}^2)$ with compact support for dilation matrices of determinant 2. The key idea is to describe the set $\mathcal{H}_2$ of all two-dimensional (2D) scaling coefficients satisfying the orthogonality condition as an implicit function. This set includes the scaling coefficients for induced 1D wavelets. We compute the tangent space of $\mathcal{H}_2$ at H_N, the scaling coefficients for induced 1D Daubechies wavelets. The structure of the tangent space allows us to build nonseparable wavelets by starting at H_N and tracing $\mathcal{H}$ along its tangent lines. Various families of compactly supported orthogonal 2D wavelets for the quincunx grid are explicitly given.

Key words. wavelets, dilation equations, multiresolution analysis

AMS subject classifications. 42A52, 65D20

1. Introduction.

The wavelet transform has by now proved to be a reliable tool for a wide range of applications in signal processing. The success of this method often relies on the special properties of orthogonal wavelets with compact support. In short, an orthogonal one-dimensional (1D) wavelet is a function ψ such that

$$\{\psi_{mk}(x) = 2^{-m/2}\psi(2^{-m}x - k) \mid k, m \in \mathbb{Z}\}$$

forms an orthonormal basis for $L^2(\mathbb{R})$. Fast algorithms require compactly supported wavelets. Besides the Haar wavelets

$$\psi(x) = \chi_{[0,1/2]}(x) - \chi_{[1/2,1]}(x),$$

no compactly supported orthogonal wavelets were known before I. Daubechies [6] succeeded in merging the wavelet idea, originally defined via group representations on the affine group [10], with the concept of a multiresolution analysis, stemming from signal processing [12]. The outcome was a family $\{\psi_N\}$ of compactly supported orthogonal wavelets with linearily increasing regularity. This family of functions has found applications in such diverse fields as data compression, numerical solutions of partial differential equations, the construction of multigrid methods [15], the examination of electrocardiograms, and many more. For a long but still incomplete list of applications, see the references in [7, 17, 2, 16].

When constructing orthogonal 2D wavelets, the dilation parameter 2 is replaced by a matrix A satisfying certain restrictions. For example, tensor products of orthogonal 1D wavelets lead to 2D wavelets for $A = \mathrm{diag}(2,2)$. In applications such as image compression, these wavelet bases lead to artifacts in directions parallel to the coordinate axis. In other words, these separable wavelets are not isotropic. Moreover, we need $|\det(A)| - 1 = 3$ wavelets in order to obtain a basis of $L^2(\mathbb{R}^2)$. Hence we are lead to study dilation matrices with $\det(A) = 2$. Here the standard example is

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix},$$

which corresponds to a rotation of $\pi/4$. Thus far, three types of orthogonal 2D wavelets are known. First of all, a beautiful generalization of the Haar wavelet leads

* Received by the editors April 2, 1993; accepted for publication (in revised form) May 30, 1995. This work was partially supported by Deutsche Forschungsgemeinschaft grant MA 1657/1–1.

† Fachbereich Mathematik, Universität Potsdam, Postfach 60 15 53, D-14415 Potsdam, Germany.

to 2D wavelets which are characteristic functions of certain sets [9]. Besides the tensor-product wavelets for diagonal dilation matrices one can also lift 1D wavelets to higher-dimensional wavelets for dilation matrices A with $|\det(A)| = 2$ [4]. Both types are called separable, they have an inherent 1D structure. Moreover, nonseparable quadrature mirror filters (QMFs) have been constructed in [11]; they lead to continuous 2D wavelets [18].

Our aim is to develop a general construction procedure for nonseparable compactly supported orthogonal 2D wavelets for dilation matrices with $|\det(A)| = 2$. As usual, we exploit the connection between wavelets and multiresoluton analysis. This leads us to study the Fourier series H of the coefficients of finite scaling equations. Instead of constructing a particular H, we examine the set $\mathcal{H}_2$ of all trigonometric polynomials H satisfying the orthogonality condition. The key idea is to use the description of $\mathcal{H}_2$ as an implicit function. Starting at a known point on $\mathcal{H}_2$, namely the Fourier series H_N associated with the lifted 1D Daubechies wavelet ψ_N, we can trace part of $\mathcal{H}_2$ via its tangential space.

In this paper, we are primarily interested in constructing wavelets. Hence the problem of investigating the regularity or the number of vanishing moments of these wavelets will only be touched upon in §4.

The paper is organized as follows. Section 2 contains the relevant theory for 1D wavelets and multiresolution analysis. Section 3 starts from the 2D scaling equation and constructs the tangent space of $\mathcal{H}_2$ at H_N. The structure of the tangent space allows us to examine $\mathcal{H}_2$ by starting at H_N and tracing $\mathcal{H}$ along its tangent lines. In order to ensure that such an $H \in \mathcal{H}_2$ actually leads to an orthogonal wavelet, we also have to check the Cohen criterion. This is done in §4, leading to families of nonseparable orthogonal wavelets.

2. Multiresolution analysis and orthogonal 1D wavelets.

Wavelets are most conveniently described within the framework of multiresolution analysis. This concept was introduced in 1986 by Y. Meyer and S. Mallat; since then, it has become the main tool for constructing compactly supported wavelets. A multiresolution analysis in $\mathbb{R}^d$ for a dilation matrix A consists of a series of nested linear spaces $\{V_m | m \in \mathbb{Z}\}$

$$\cdots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset \cdots$$

such that

$$\overline{\cup V_m} = L^2(\mathbb{R}^d), \qquad \cap V_m = \{0\},$$

$$f(x) \in V_m \Longleftrightarrow f(Ax) \in V_{m-1},$$

where A, the dilation matrix, is a matrix whose eigenvalues λ have modulus greater than 1, $|\lambda| > 1$, and which maps the integer vectors $k \in \mathbb{Z}^d$ to integer vectors, i.e.,

$$\Gamma = A\mathbb{Z}^d \subset \mathbb{Z}^d.$$

The decisive condition that makes this scheme work is the requirement that there exists a function φ such that

$$\{\varphi(x-k) \mid k \in \mathbb{Z}^d\} \text{ is an orthonormal basis of } V_0.$$

The requirement that φ and its integer translates are orthonormal can be replaced by alternative conditions [7], e.g., by the stability requirement that

$$\{\varphi(x-k) \mid k \in \mathbb{Z}^d\} \text{ forms a Riesz basis of } V_0.$$

Note that the subspace V_m is spanned by $\{\varphi(A^{-m}x-k) \mid k \in \mathbb{Z}^d\}$. The name *multiresolution analysis* stems from the interpretation that different subspaces V_m contain functions representing details of increasing size as $m \longrightarrow \infty$. The projection of a signal f onto V_m can thus be interpreted as applying a low-pass filter with diminishing bandwidth as $m \longrightarrow \infty$. From $\varphi \in V_0 \subset V_{-1}$, it follows that φ satisfies a scaling equation

$$\varphi(x) = |\det(A)| \sum h_k \varphi(Ax-k). \tag{1}$$

The set of scaling coefficients $\{h_k\}$ is called the associated discrete filter. Most of this article deals with properties of the Fourier series of such discrete filters.

The Fourier series of the scaling coefficients is denoted by $H(\omega)$, $\omega \in \mathbb{R}^d$:

$$H(\omega) = sum_{k\in\mathbb{Z}^d} h_k e^{-ik\cdot\omega}, \tag{2}$$

where $\omega \cdot k$ denotes the standard scalar product.

Let A^{-t} denote the transpose of the inverse matrix A^{-1} and define the Fourier transform of a function $f \in L^2(\mathbb{R}^d)$ by

$$\hat{f}(\omega) = (2\pi)^{-d/2} \int\limits_{\mathbb{R}^d} f(x) e^{-i\omega\cdot x}\, dx.$$

Then taking the Fourier transform on both sides of the scaling equation (1) leads to

$$\begin{aligned} \hat{\varphi}(\omega) &= H(A^{-t}\omega)\hat{\varphi}(A^{-t}\omega) \\ &= \prod_{m\geq 1} H((A^{-t})^m\omega). \end{aligned} \tag{3}$$

By W_m we denote the orthogonal component of V_m in V_{m-1}:

$$V_m \oplus W_m = V_{m-1}.$$

In other words, W_m is the complement of a low-pass filtered function space V_m in a space V_{m-1} of functions with a larger bandwidth, i.e., the projection of a signal onto W_m amounts to applying a band-pass filter. Y. Meyer [13] proved that there exist functions $\psi_1, \ldots, \psi_{|\det(A)|-1} \in L^2(\mathbb{R})$, the associated wavelets, with mean value 0 such that

$$\overline{\operatorname{span}\{\psi_j(x-k) \mid k \in \mathbb{Z}^d,\ j = 1, \ldots, |\det(A)|-1\}} = W_0,$$

$$\langle \psi_i(\cdot), \psi_j(\cdot - k)\rangle_{L^2(\mathbb{R}^d)} = \delta_{0k}\delta_{ij}.$$

In other words, only dilation matrices with $|\det(A)| = 2$ allow the construction of a single function ψ such that

$$\{\psi_{mk}(x) = 2^{-m/2}\, \psi(A^{-m}x-k) \mid m \in \mathbb{Z}, k \in \mathbb{Z}^2\} \tag{4}$$

forms an orthonormal basis for $L^2(I\!R^d)$.

The most interesting dilation matrix in the 1D case, $d = 1$, is therefore $A = 2$, and ϕ satisfies the scaling relation

$$\varphi(x) = 2\sum h_k \; \varphi(2x - k). \tag{5}$$

An orthogonal wavelet ψ can then be constructed explicitly from a given orthogonal scaling function ϕ, i.e., $\phi(x) \perp \phi(x-k)$ if $k \neq 0$, by, e.g.,

$$\psi(x) = 2\sum g_k \varphi(2x - k), \tag{6}$$

$$g_k = (-1)^k h_{1-k}. \tag{7}$$

All properties of ψ can be deduced from the scaling coefficients $\{h_k\}$ [6]. For example, in our search for compactly supported wavelets, we need only look at scaling equations with a finite number of nonzero scaling coefficients. After introducing the Fourier series

$$H(\omega) = \sum_{k=0}^{n} h_k e^{-ik\omega}, \qquad G(\omega) = \sum_{k=0}^{n} g_k e^{-ik\omega},$$

the main step in constructing compactly supported orthogonal wavelets requires us to find a trigonometric polynomial H satisfying the orthogonality condition [6].

THEOREM 2.1. *If $\varphi \in L^2(I\!R)$ is a solution of* (5) *which is orthogonal to its integer translates, i.e.,*

$$k \neq 0 : \varphi(x) \perp \varphi(x - k),$$

then

$$|H(\omega)|^2 + |H(\omega + \pi)|^2 = 1, \qquad H(0) = 1. \tag{8}$$

A set of coefficients $\{h_k, g_k\}$ satisfying $(3, 4)$ is called a QMF in the language of signal processing. A QMF is the discrete analogue of the function pair (φ, ψ). Thus far, we have reviewed the procedure which starts with a multiresolution analysis and the related scaling function φ (resp. the wavelets ψ) and ends at a QMF, i.e., a set of discrete coefficients $\{h_k\}$ satisfying an orthogonality condition (8). The construction of orthogonal wavelets proceeds in the reverse direction. However, there exist pathological examples of QMFs where the corresponding solutions φ (resp. ψ) of (5) (resp. (6)) are not orthogonal to there integer translates [6]. This can be avoided, e.g., by checking the following Cohen criterion, which we state for dimensions $d = 1$ and $d = 2$.

CRITERION 2.2 (Cohen criterion). *The Fourier series $H(\omega) = \sum_{k \in Z\!\!\!Z} h_k \; e^{-ik\omega}$, $\omega \in I\!R$, satisfies the 1D Cohen criterion if there exists a set $K \subset I\!R$ such that*

- *K contains a neighborhood of the origin,*
- *$|K| = 2\pi$ and for all ω in $[-\pi, \pi]$, there exists a $k \in Z\!\!\!Z$ such that $\omega + 2k\pi \in K$,*
- *for all $m > 0$, $H(2^{-m}\omega)$ does not vanish on K.*

The Fourier series $H(\omega) = \sum_{k \in Z\!\!\!Z^2} h_k \; e^{-ik\cdot\omega}$, $\omega = (\omega_1, \omega_2) \in I\!R^2$, satisfies the 2D Cohen criterion if there exists a set $K \subset I\!R^2$ such that

- *K contains a neighborhood of the origin,*

- $|K| = 4\pi^2$ *and for all* ω *in* $[-\pi,\pi]^2$, *there exists a* $k \in \mathbb{Z}^2$ *such that* $\omega + 2\pi k \in K$,
- *for all* $m > 0$, $H(A^{-m}\omega)$ *does not vanish on* K.

In [3], it was proved that this criterion in addition to the orthogonality condition (8) is necessary and sufficient for the L^2 convergence of the infinite product (3) and the orthogonality of the limit function.

THEOREM 2.3. *Suppose* $H(\omega) = \sum_{k\in\mathbb{Z}} h_k\, e^{ik\omega}$ *satisfies* (8). *Then the infinite product* (3) *converges in* $L^2(\mathbb{R})$ *to an orthogonal solution of* (5) *if and only if* H *satisfies the Cohen criterion* (*Criterion* 2.2).

Our approach for constructing orthogonal wavelets starts by introducing $\mathcal{H}_1$, the set of all trigonometric polynomials in one variable satisfying the orthogonality condition (8), and its subset $\mathcal{H}_1^N \subset \mathcal{H}_1$,

$$\mathcal{H}_1^N = \left\{ H(\omega) = \sum_{k=0}^{2N-1} h_k e^{-ik\omega} \mid H \text{ satifies } (8) \right\},$$

consisting of the polynomial solutions of (8) of degree $(2N-1)$. We choose to label the subspaces by N instead of using the filter length $n = 2N-1$ because the Daubechies wavelets ψ_N obey scaling relations with $2N$ coefficents and their related Fourier series H_N are therefore elements of H_1^N. These polynomials have $2N$ coefficients; hence $\mathcal{H}_1^N$ can be viewed as a manifold in $\mathbb{R}^{2N}$.

Let us further define

$$q(\omega) = |H(\omega)|^2. \tag{9}$$

For any $H \in \mathcal{H}_1$, the corresponding q is an even positive trigonometric polynomial with $q(0) = 1$ solving the simple equation

$$q(\omega) + q(\omega + \pi) = 1. \tag{10}$$

This is a linear equation with the general solution

$$\mathcal{K}_1 = \left\{ q(\omega) = 1/2 + \sum_{k\in\mathbb{Z}} \alpha_k \cos\left((2k+1)\omega\right) \mid \sum \alpha_k = 1/2,\ q \geq 0 \right\}.$$

$\mathcal{K}_1$ is the intersection of a linear affine space with the convex cone of positive functions, i.e., $\mathcal{K}_1$ is a convex set. We introduce the map

$$q : \mathcal{H}_1 \longrightarrow \mathcal{K}_1 \tag{11}$$

$$H \longmapsto |H|^2. \tag{12}$$

We will often use the shorthand notation $q(\omega)$ for $(q(H))(\omega)$; we will identify the trigonometric polynomial H with the set of its coefficients whenever appropriate. $\mathcal{K}_1$ has a simple structure; hence our construction follows a four-step procedure:

1. choose a $q \in \mathcal{K}_1$;
2. solve $|H|^2 = q$;
3. check the Cohen criterion;
4. solve the scaling equations (5) and (6).

In general, it is not possible to give a solution of (1) in closed form, but the graphical iteration process described in [6] converges to a solution under rather general conditions. Moreover, the values of φ and ψ at dyadic values can be computed efficiently

[7]. This solves step 4. Step 2, i.e., the question of inverting the map q, is answered by the following theorem.

THEOREM 2.4 (Fejer–Riesz theorem). *Let q be a real positive polynomial in $\cos(\omega)$, i.e.,*

$$q(\omega) = \sum_{k=0}^{n} \alpha_k \cos(k\omega), \quad q(\omega) \geq 0.$$

Then there exists a trigonometric polynomial H of the same degree, i.e.,

$$H(\omega) = \sum_{k=0}^{n} h_k e^{-ik\omega},$$

with real coefficients, such that

$$q(\omega) = |H(\omega)|^2. \tag{13}$$

This result is due to Riesz. The proof is based on the ability to factorize polynomials in one variable, i.e.,

$$q(\omega_0) = 0 \Longrightarrow (\omega - \omega_0) \mid q(\omega);$$

see [14]. This is not possible for polynomials in more than one variable. The major step in constructing orthogonal 2D wavelets will be a characterization of some trigonometric polynomials in two variables which allow taking the root in the sense of the above theorem.

The Fejer–Riesz theorem (Theorem 2.4) implies that any $q \in \mathcal{K}_1$ has a root in the sense of (13). Another way to view this result is as follows.

COROLLARY 2.5. $q(\mathcal{H}_1) = \mathcal{K}_1$.

The main advantage of investigating $\mathcal{K}_1$ instead of $\mathcal{H}_1$ is that $\mathcal{K}_1$ is a flat linear manifold while $\mathcal{H}_1$ is defined by a set of quadratic equations. Moreover, $\mathcal{K}_1$ is convex. To end this section, we determine a special set of extremal points of $\mathcal{K}_1$ by constructing a set of supporting hyperplanes. This will lead to the well-known family of orthogonal 1D Daubechies wavelets. We define the subset of trigonometric polynomials of degree $2N-1$:

$$\mathcal{K}_1^N := \{q \in \mathcal{K}_1 \mid \deg(q) \leq 2N - 1\}.$$

LEMMA 2.6. *For fixed N, let $\{d_k \mid k = 0, \ldots, 2N-1\}$ denote the scaling coefficients associated with the Daubechies wavelets ψ_N and let*

$$H_N(\omega) = \sum_{k=0}^{2N-1} d_k \; e^{-ik\omega}$$

denote the corresponding Fourier series. Then

$$|H_N(\omega)|^2 = q_N(\omega) = 1 - c_N \int_0^{\omega} \sin(t)^{2N-1} \, dt,$$

$$c_N^{-1} = \int_0^{\pi} \sin^{2N-1}(t) dt$$

is an extremal point of $\mathcal{K}_1^N$.

Proof. Define

$$u(\omega) = 1 - c_N \int_0^\omega \sin(t)^{2N-1} dt.$$

We begin by establishing that

$$q_N(\omega) = \sum_{k=0}^{N-1} \begin{pmatrix} N-1+k \\ k \end{pmatrix} (\sin(\omega/2))^{2k},$$

as defined in [6], is indeed equal to $u(\omega)$. With the above definition of c_N, we have

$$\begin{aligned} u(\omega+\pi) &= -c_N \int_\pi^{\pi+\omega} \sin(t)^{2N-1}\, dt \\ &= c_N \int_0^\omega \sin^{2N-1}(t) dt. \end{aligned}$$

u is a positive polynomial of degree $2N-1$ in $\cos(t)$ which solves (10); therefore, $u \in \mathcal{K}_1^N$. Moreover,

$$u'(\omega) = -c_N \ \sin^{2N-1}(\omega),$$

which implies that u is the unique element in $\mathcal{K}_1^N$ with

$$u^{(k)}(\pi) = 0, \qquad k = 0, 1, \ldots, 2N-1.$$

On the other hand, the Daubechies wavelet ψ_N was constructed by requiring that $H_N \in \mathcal{H}_1^N$ or, equivalently, $q_N = |H_N|^2 \in \mathcal{K}_1^N$ has a zero at $\omega = \pi$ of the highest possible order. It follows that $u = q_N$. This part of the lemma, with a different reasoning, can be found in [13].

It remains to show that u is an extremal point of $\mathcal{K}_1^N$. We construct supporting hyperplanes. Any

$$q(\omega) \ = \ 1/2 + \sum_{k=1}^{2N-1} \alpha_k \ \cos((2k-1)\omega) \in \mathcal{K}_1^N$$

is an even positive smooth function which solves (10); therefore, $0 \le q(\omega) \le 1$. Since $q(0) = 1$, we immediately have

$$q'(0) = 0, \qquad q''(0) = \sum_{k=1}^{2N-1} (2k-1)^2 \alpha_k \le 0.$$

Hence the condition $q''(0) = 0$ defines a supporting hyperplane $\mathcal{E}_1$:

$$\mathcal{E}_1 : \sum_{k=1}^{N} (2k-1)^2 \alpha_k = 0.$$

Note that $q(H_N) = q_N$ lies on this hyperplane.

$\mathcal{K}_1^N \cap \mathcal{E}_1$ is again a convex set. Among the elements $q \in \mathcal{K}_1 \cap \mathcal{E}_1$, we select the functions on the plane $\mathcal{E}_2$ defined by $q^{(4)}(0) = 0$. This again gives a condition on $\{\alpha_k\}$:

$$\mathcal{E}_2 : \sum_{k=1}^{N} (2k-1)^4 \alpha_k = 0.$$

We can continue in this vein, defining successive $\mathcal{E}_\ell$, $\ell = 1, \dots, N$, each defined by

$$\mathcal{E}_\ell : \sum_{k=1}^{N} (2k-1)^{2\ell} \alpha_k = 0,$$

and each $\mathcal{E}_\ell$ is a supporting hyperplane for

$$\mathcal{K}_1 \cap \mathcal{K}_2 \cap \cdots \cap \mathcal{K}_{\ell-1}.$$

The corresponding linear equations for $\{\alpha_k\}$ are linearly independent. Hence there is a unique element in $\mathcal{K}_1^N$ lying in the intersection of all hyperplanes

$$q \in \cap_{k=1}^{N} \mathcal{E}_k.$$

By construction, $q_N \in \cap_{k=1}^{N} \mathcal{E}_k$, so that q_N is this unique element. □

3. 2D wavelets. Our aim is to construct compactly supported orthogonal 2D wavelets, i.e., we search for functions $\psi \in L^2(\mathbb{R}^2)$ such that

$$\{\psi_{mk}(x) = 2^{-m/2}\ \psi(A^{-m}x - k) \mid m \in \mathbb{Z},\ k \in \mathbb{Z}^2\}$$

forms an orthonormal basis for $L^2(\mathbb{R}^2)$. Hence we consider dilation matrices A with $|\det(A)| = 2$. Obviously, the scaling function φ and the related wavelet ψ depend very much on the dilation matrix A; see (3). For example, the same set of coefficients $\{h_k\}$ may lead to a wavelet with arbitrary high regularity for one dilation matrix A and to a discontinuous wavelet for another A [4].

However, the 2D equivalent to the 1D orthogonality condition of Theorem 2.1 depends only on a set of representatives of the cosets of the adjoint grid

$$\tilde{\Gamma} = A^t \mathbb{Z}^2;$$

see Lemma 3.1 below. If $|\det(A)| = 2$, then the set of representatives consists of a single vector $z \in \mathbb{Z}^2 \backslash \tilde{\Gamma}$. There are exactly three different grids Γ stemming from dilation matrices A with $|\det(A)| = 2$:

- the line grid, i.e., $m = (m_1, m_2) \in \Gamma \Longleftrightarrow m_2$ is even;
- the column grid, i.e., $m = (m_1, m_2) \in \Gamma \Longleftrightarrow m_1$ is even;
- the quincunx grid, i.e., $m = (m_1, m_2) \in \Gamma \Longleftrightarrow m_1 + m_2$ is even.

Since the first two grids—and their adjoint grids $\tilde{\Gamma}$—are related to each other by a simple exchange of variables, we may assume without loss of generality that

$$z = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

As in the 1D case, the construction of orthogonal wavelets centers around the Fourier series H of the scaling coefficients. The following results treat all dilation matrices with $|\det(A)| = 2$, but to illustrate the construction, we will always refer to the standard example where A incorporates a rotation by $\pi/4$:

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

This matrix has been discussed, e.g., in [4], [17]. It is the most basic dilation matrix with two complex eigenvalues; hence it produces nonseparable wavelets.

3.1. Orthogonality relation. As in the 1D case, the orthogonality of $\{\varphi(x-m)\}$ leads to an orthogonality condition expressed in terms of the Fourier series (2). The proof of the following lemma is a straightforward generalization of the corresponding result in $L^2(\mathbb{R})$; see [13].

LEMMA 3.1. *Given a dilation matrix A with an arbitrary value of $|\det(A)|$, choose a complete set of representatives of $\tilde{\Gamma}$*

$$\{z_1 = 0, z_2, \dots, z_{|\det(A)|}\},$$

i.e.,

$$\mathbb{Z}^2 = \bigcup_{k=1}^{|\det(A)|} (z_k + \tilde{\Gamma}).$$

Suppose the scaling equation (1) has a solution $\varphi \in L^2(\mathbb{R}^d)$ which is orthogonal to its integer translates

$$\varphi(x) \perp \varphi(x-m) \quad \textit{for all} \quad m \neq 0 \in \mathbb{Z}^d.$$

Let $z \in \mathbb{Z}^2$ be a representative of the coset of the grid $\tilde{\Gamma}$, i.e., $(z+\tilde{\Gamma}) \cup \tilde{\Gamma} = \mathbb{Z}^2$.

Then the Fourier series H satisfies the orthogonality relation

$$|H(\omega)|^2 + |H(\omega + 2\pi A^{-t}z)|^2 = 1, \qquad H(0) = 1. \tag{14}$$

If we specify this result, then we may choose a representative z such that

$$2\pi A^{-1} z = \begin{pmatrix} \pi \\ \pi \end{pmatrix}.$$

COROLLARY 3.2. *For $d = 2$ and $|\det(A)| = 2$, the orthogonality condition of Lemma 3.1 is given by*

$$|H(\omega)|^2 + \left| H\left(\omega + \begin{pmatrix} \pi \\ \pi \end{pmatrix}\right)\right|^2 = 1, \qquad H(0) = 1. \tag{15}$$

The 2D orthogonality condition is very similar to its 1D counterpart, and we can easily lift Fourier series which satisfy (8) to higher dimensions.

COROLLARY 3.3. *If $H_1(\omega)$ satisfies the 1D orthogonality condition, then*

$$H(\omega_1, \omega_2) = H_1(\omega_1)$$

obeys (15).

This leads to so-called induced wavelets, which have been studied in [4]. They are not suitable for application since they are not isotropic and have a poor smoothness.

As in the 1D case, we have to check in addition the Cohen criterion, which insures that an H satisfying (15) leads to an orthogonal scaling function which generates a multiresolution analysis [4]. Moreover, we still need to know how to construct the wavelet ψ given the scaling function φ.

LEMMA 3.4. *Let A denote a dilation matrix with $|\det(A)| = 2$ and let z be a representative of the coset of $\tilde{\Gamma}$. Given an orthogonal scaling function φ with scaling coefficients $\{h_k \mid k \in \mathbb{Z}^2\}$, define*

$$g_k = (-1)^{e(k)} h_{z-k}, \quad k \in \mathbb{Z}^2,$$

where the exponent $e(k)$ is given by

$$e(k) = \begin{cases} 0 & \text{if} \quad k \in \tilde{\Gamma}, \\ 1 & \text{if} \quad k \in \mathbb{Z}^2 \backslash \tilde{\Gamma}, \end{cases}$$

e.g., for the quincunx grid, we may choose $e(k) = k_1 + k_2$. Then

$$\psi(x) := 2 \sum g_k \varphi(Ax - k)$$

defines an orthogonal 2D wavelet, i.e.,

$$\{2^{-m/2} \ \psi(A^{-m}x - k) \mid m \in \mathbb{Z}, k \in \mathbb{Z}^2\}$$

is an orthonormal basis for $L^2(\mathbb{R}^2)$.

Proof. The proof proceeds analogously to the 1D-case. □

3.2. The implicit-function approach. As in the 1D case, we examine the set of trigonometric polynomials H which solve (15):

$$\mathcal{H}_2 := \{H \mid H \text{ solves (17), } H \text{ is a trigonometric polynomial}\}.$$

By $\mathcal{H}_2^N \subset \mathcal{H}_2$ we denote the subspace of trigonometric polynomials of degree $\deg(H) \leq 2N - 1$. We define

$$q(\omega) = |H(\omega)|^2. \tag{16}$$

The general solution of $q(\omega) + q(\omega + \pi \begin{pmatrix}1\\1\end{pmatrix}) = 1$ is given by

$$\mathcal{K}_2 = \left\{ q(\omega) = 1/2 + \sum_{\substack{k \in \mathbb{Z}^2 \\ k_1+k_2 \text{ odd}}} \alpha_k \cos(k \cdot \omega) \mid q(0) = 1, \ q \geq 0 \right\}. \tag{17}$$

$\mathcal{K}_2$ is the intersection of an affine linear space with the convex cone of positive functions, i.e., $\mathcal{K}_2$ is a convex set. In contrast to the 1D case, not every $q \in \mathcal{K}_2$ has a polynomial root in the sense of (16). We define the subset

$$q(\mathcal{H}_2) = \mathcal{K}_{\text{orth}},$$

$$\mathcal{K}_{\text{orth}} = \{q \in \mathcal{K}_2 \mid q = |H|^2, \ H \text{ is a trigonometric polynomial}\}.$$

In a slight abuse of notation, we write H for the set of scaling coefficients $\{h_k^N \mid k \in \mathbb{Z}^2\}$ and we use the same symbol for the Fourier series

$$H(\omega) = \sum_{k \in \mathbb{Z}^2} h_k e^{-ik \cdot \omega}.$$

As described in Corollary 3.3, we can lift elements $q_1 \in \mathcal{K}_1$ (resp. $H_1 \in \mathcal{H}_1$) to higher dimensions:

$$q_1 \in \mathcal{K}_1 : q(\omega_1, \omega_2) = q_1(\omega_1) \in \mathcal{K}_{\text{orth}}.$$

For example, we can lift the scaling coefficients $\{d_k \mid k = 0, \dots, 2N-1\}$ of the Daubechies wavelet ψ_N:

$$H_N = \{h_k^N \mid k = (k_1, k_2),\ 0 \le k_1,\ k_2 \le 2N-1\}. \tag{18}$$

Note that

$$h^N_{(k_1,0)} = d_{k_1},$$

$$h^N_{(k_1,k_2)} = 0 \quad \text{if} \quad k_2 \neq 0.$$

Notational remark. The length of the filters is generally denoted by n, but in most cases, we deal with the dimension $n = 2N-1$ related to the Daubechies filters.

The starting point for the following is the observation

$$H_N \in \mathcal{H}_1 \subset \mathcal{H}_2. \tag{19}$$

$$q(H_N) \in \mathcal{K}_1 \subset \mathcal{K}_{\text{orth}} \subset \mathcal{K}_2. \tag{20}$$

As in the 1D case, we expect that $\mathcal{K}_{\text{orth}}$ has a more convenient structure than $\mathcal{H}_2$, at least $\mathcal{K}_{\text{orth}}$ contains the flat subset $\mathcal{K}_1$.

Our procedure for investigating $\mathcal{H}_2$ and $\mathcal{K}_2$ proceeds as follows:

1. Describe $\mathcal{H}_2$ as an implicit function, i.e.,

$$H(\omega) = \sum_{0 \le k_1, k_2 \le n} h_k e^{-ik\cdot\omega} \in \mathcal{H}_2 \Longleftrightarrow F(H) = 0.$$

The function $F : I\!R^{(n+1)^2} \to I\!R^{(n+1)^2-n+1}$ is determined either via the description (17) of $\mathcal{K}_2$ or by expressing the orthogonality condition (15) directly in terms of the coefficients $\{h_k\}$.

2. Choose $n = 2N-1$ and apply the implicit-function theorem to compute the tangent space of $\mathcal{H}_2^N$ at H_N. This requires three steps:

(a) Compute the Jacobian J of F at H_N.

(b) Split the coefficients in two sets $x \in I\!R^{n-1}$ and $y \in I\!R^{(n+1)^2}$ such that

$$\frac{\partial F}{\partial y}(H_N) \text{ is not singular.}$$

(c) Compute the tangent vectors of $\mathcal{H}_2$ at H_N with the help of

$$\left(\frac{\partial F}{\partial y}(H_N)\right)^{-1} \frac{\partial F}{\partial x}(H_N) = -\nabla g,$$

where $g(x)$ is the function implicitly defined by $F(x, g(x)) = 0$.

3. Choose $n = 2N-1$, start at $H_N \in \mathcal{H}_2$, follow a tangent vector t_N, and add a correction term in order to stay on $\mathcal{H}_2$:

$$H = H_N + s\, t_N + \text{correction term}.$$

We begin by writing out the defining equations for $\mathcal{H}_2$ and $\mathcal{K}_{\mathrm{orth}}$ in detail. We are interested in compactly supported wavelets. Hence we investigate scaling equations with a finite number of scaling coefficients $\{h_k\}$. Let

$$H(\omega) = \sum_{k_1,k_2=0}^{n} h_k \, e^{-ik\cdot\omega} \tag{21}$$

be a trigonometric polynomial of degree n with $(n+1)^2$ coefficients. The grid points corresponding to coefficients of the polynomial are marked in Figure 1. From (15), we get the first condition for the coefficients $\{h_k\}$: $H(0) = 1$, i.e.,

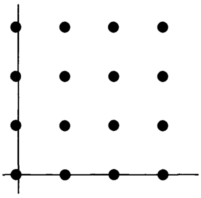


FIG. 1.

$$\sum_k h_k = 1. \tag{22}$$

The coefficients of $q = |H|^2$ are given by

$$\begin{aligned} q(\omega) &= \sum_{k,l} h_k h_l \, e^{-i(k-l)\cdot\omega} = \sum_m \left\{ \sum_k h_k h_{k-m} \right\} e^{-im\cdot\omega} \\ &= \sum_k h_k^2 + 2 \sum_{m_1=1}^{n} \left\{ \sum_{k_1=0}^{n} h_{(k_1,0)} h_{(k_1-m_1,0)} \right\} \cos(m_1\omega_1) \\ &\quad + 2 \sum_{m_1=-n}^{n} \sum_{m_2=1}^{n} \left\{ \sum_k h_k h_{k-m} \right\} \cos(m\cdot\omega). \end{aligned}$$

$q \in \mathcal{K}_2$ implies that the coefficients of $\cos(m\cdot\omega)$, m_1+m_2 even, have to vanish and that the constant term has to equal $1/2$; see (17). Together with (22), this poses n^2+n+2 conditions. In other words, the coefficients $H = \{h_k\}$ satisfying (15) are the zeros of

$$F : \mathbb{R}^{(n+1)^2} \longrightarrow \mathbb{R}^{n^2+n+2},$$

where the n^2+n+2 equations are given by

$$F_1(H) = \sum_k h_k - 1,$$

$$F_2(H) = \sum_k h_k^2 - 1/2,$$

$$F_m(H) = \sum_k h_k h_{k-m} \quad \text{for } m \neq (0,0),\ m_1+m_2 \text{ even}.$$

We are interested in filters of length $n = 2N - 1$; $\mathcal{H}_2^N$ is therefore given here as an implicitly defined function

$$\mathcal{H}_2^N = \{H \in I\!R^{(2N)^2} \mid F(H) = 0\}.$$

The next step requires us to compute the Jacobian of F evaluated at the scaling coefficients $\{h_k^N\}$ for the induced 1D Daubechies wavelets (18). These coefficients vanish if $k_2 \neq 0$; in order to distinguish between the 2D array of coefficients h_k^N and its 1D vector of nonzero coefficients, we define for fixed N

$$d_k = h^N_{(k,0)}, \quad k = 0, \dots, 2N-1. \tag{23}$$

The partial derivatives of F_1 and F_2 are given by

$$\frac{\partial F_1}{\partial h_k}(H_N) = 1, \qquad \frac{\partial F_2}{\partial h_k}(H_N) = 2h_k^N = \begin{cases} 2d_{k_1} : k_2 = 0, \\ \ \ 0 \ \ : k_2 \neq 0. \end{cases}$$

For $m = (m_1, 0)$, $m_1 \neq 0$ even, we obtain

$$\frac{\partial F_m}{\partial h_k}(H_N) = \begin{cases} d_{k_1 - m_1} + d_{k_1 + m_1} : k_2 = 0, \\ \qquad 0 \qquad\quad : \text{otherwise}. \end{cases}$$

We combine these partial derivatives in an $(N-1) \times (2N)$ submatrix J_0 with coefficients $(J_1)_{ij}$, $1 \le i \le (N-1)$, $0 \le j \le (2N-1)$:

$$(J_0)_{ij} = \frac{\partial F_{(2i,0)}}{\partial h_{(j,0)}}(H_N) = d_{j-2i} + d_{j+2i}. \tag{24}$$

For example, for $N = 2$, i.e., $n = 3$, this submatrix is simply

$$J_0 = (d_2, d_3, d_0, d_1).$$

The remaining partial derivatives for $1 \le m_2 \le n$, $-n \le m_1 \le n$, $m_1 + m_2$ even, are determined by

$$\frac{\partial F_m}{\partial h_k}(H_N) = \begin{cases} d_{k_1 - m_1} : \text{if } k_2 = m_2,\ m_1 \le k_1 \le n + m_1, \\ \ \ 0 \ \ \ : \text{otherwise}. \end{cases}$$

For a fixed $m_2 > 0$, m_2 odd, we gather the relevant derivatives in a $(2N) \times (2N)$ submatrix J_1 with coefficients $(J_1)_{ij}$, $0 \le i, j \le (2N-1)$:

$$(J_1)_{ij} = \frac{\partial F_{(-n+2i, m_2)}}{\partial h_{(j, m_2)}}(H_N) = d_{j+n-2i}, \tag{25}$$

$$J_1 = \begin{pmatrix} d_n & 0 & 0 & 0 \cdots 0 \\ d_{n-2} & d_{n-1} & d_n & 0 \cdots 0 \\ & \cdot & & \cdot \\ & \cdot & & \cdot \\ d_1 & d_2 & \cdots & d_n & 0 \\ 0 & d_0 & d_1 & \cdots & d_{n-1} \\ & \cdot & & & \cdot \\ & \cdot & & & \cdot \\ 0 & \cdots & 0 & 0 & d_0 \end{pmatrix};$$

for a fixed $m_2 > 0$, m_2 even, we obtain a $(2N-1)\times(2N)$ submatrix J_2 with coefficients $(J_2)_{ij},\ 0 \le i \le (2N-2),\ 0 \le j \le (2N-1)$:

$$(J_2)_{ij} = \frac{\partial F_{(-n+1+2i,m_2)}}{\partial h_{(j,m_2)}}(H_N) = d_{j+n-1-2i}, \tag{26}$$

$$J_2 = \left(\begin{array}{cccccc} d_{n-1} & d_n & 0 & 0 & 0 & \cdots 0 \\ d_{n-3} & \cdots & d_{n-1} & d_n & 0 & \cdots 0 \\ & \cdot & & & & \cdot \\ & \cdot & & & & \cdot \\ d_0 & d_1 & \cdots & \cdots & \cdots & d_n \\ 0 & 0 & d_0 & \cdots & d_{n-3} & d_{n-2} \\ & \cdot & & & & \cdot \\ & \cdot & & & & \cdot \\ 0 & \cdots & \cdots & 0 & d_0 & d_1 \end{array}\right).$$

For $N = 2$, these matrices reduce to

$$J_1 = \begin{pmatrix} d_3 & 0 & 0 & 0 \\ d_1 & d_2 & d_3 & 0 \\ 0 & d_0 & d_1 & d_2 \\ 0 & 0 & 0 & d_0 \end{pmatrix}, \qquad J_2 = \begin{pmatrix} d_2 & d_3 & 0 & 0 \\ d_0 & d_1 & d_2 & d_3 \\ 0 & 0 & d_0 & d_1 \end{pmatrix}.$$

If we arrange the 2D array of coefficients $\{h_k\}$ in a vector $h'_{k'}$ by

$$h_{(k_1,k_2)} = h'_{k'}, \quad k' = k_1 + (n+1)k_2,$$

i.e., the subvectors of length $n+1$ correspond to coefficents h_k on the same line in Figure 1, then the full Jacobian J has the following block structure:

$$J = \left(\begin{array}{ccccccc} 1 & \cdots & \cdots & \cdots & \cdots & \cdots & 1 \\ 2D_0,\ldots,2D_n & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ & & & & & & \\ J_0 & & & & & & \\ & J_1 & & & & & \\ & & J_2 & & & & \\ & & & J_1 & & & \\ & & & \cdots & \cdots & \cdots & \\ & & & & & J_2 & \\ & & & & & & J_1 \end{array}\right).$$

For $N = 2$, we obtain a 14×16 matrix. We collect some results concerning J_1 and J_2. Similar matrices occur in the dyadic construction of $\varphi(x)$ [5].

LEMMA 3.5. *Let J_1 and J_2 be defined as above, $n = 2N-1$.*

(a) *J_1 is a regular matrix, and*

(b) *J_2 has rank n and $x = (x_j)_{j=0,\ldots,n}$, $x_j = (-1)^j d_{n-j}$, satisfies*

$$J_2 x = 0.$$

Proof. Let b^j, $j = 1,\ldots,(n+1)/2$, denote the jth line vector of J_1. Since $d_n \neq 0$, these vectors are linearly independent. Since $d_0 \neq 0$, we conclude that the other line vectors, where c^j, $j = 1, \ldots, (n+1)/2$, denotes the $(n+2-j)$th line vector of J_1, are also linearly independent.

The spans of the two sets of vectors are orthogonal since ($d_\ell = 0$ if $\ell < 0$ or $\ell > n$)

$$\begin{aligned}\langle b^j, c^i \rangle &= \sum_{k=0}^{n} d_{k+n-2(j-1)} d_{k+n-2(n+1-i)} \\ &= \sum_{k=0}^{n} d_k d_{k+2(j-1-n-1+i)},\end{aligned}$$

and the orthogonality of the Daubechies scaling function gives ($\ell \neq 0$)

$$\begin{aligned}0 &= \langle \varphi_N(\cdot), \varphi_N(\cdot - \ell) \rangle \\ &= 4 \sum_{k,m=0}^{n} d_k d_m \langle \varphi_N(2 \cdot -k), \varphi_N(2 \cdot -2\ell - m) \rangle \\ &= 4 \sum_{k=0}^{n} d_k d_{k-2\ell}.\end{aligned}$$

The indices i and j run between 1 and $(n+1)/2$; hence $\ell = j - n + i - 2 \leq -1$ and

$$\langle b^j, c^i \rangle = 0.$$

Hence the $(n+1)$ line vectors $\{b^j, c^j\}$ of J_1 are linearly independent; this proves (a).

In the same manner, we prove that the line vectors of J_2 are linearly independent. It remains to construct a nonzero vector in the kernel of J_2. Here we exploit the fact that

$$\forall \ell : \langle \varphi_N(\cdot),\ \psi_N(\cdot - \ell) \rangle = 0.$$

Inserting the 1D scaling relations yields

$$\begin{aligned}0 &= \sum_{m,k} d_k (-1)^m d_{1-m} \langle \varphi_N(2 \cdot -k), \varphi_N(2 \cdot -2\ell - m) \rangle \\ &= \sum (-1)^k d_k d_{1+2\ell-k}.\end{aligned}$$

The ith coefficient of $J_2 x$ is

$$\begin{aligned}(J_2 x)_i &= \sum d_{k+n-1-2i} (-1)^k d_{n-k} \\ &= \sum_{k=0}^{n} (-1)^{n-k} d_{-k+2n-2i-1} d_k \\ &= 0.\end{aligned}$$

This proves (b). □

In order to apply the implicit-function theorem to

$$F : I\!R^{(n+1)^2} \longrightarrow I\!R^{n^2+n+2}, \qquad F(H_N) = 0,$$

we need to find $n-1$ "free" coefficients $x = (h_{f_1}, \ldots, h_{f_{N-1}})$ such that the remaining "dependent" coefficients $y = (h_{e_1}, \ldots, h_{e_{N^2+N+2}})$ have a regular Jacobian $\frac{\partial F}{\partial y}(H_N)$. Unfortunately, this is not possible.

LEMMA 3.6. *For any choice of $(n-1)$ coefficients $x = (h_{f_1}, \dots, h_{f_{N-1}})$, the Jacobian of F at H_N with respect to the remaining dependent coefficients $y = (h_{d_1}, \dots, h_{d_{N^2+N+2}})$ is singular.*

Proof. $H_N(\omega)$ satisfies the orthogonality condition (15); hence $H(0) = 1$ and

$$H_N(\pi) = \sum_{k=0}^{n} d_k e^{-ik\pi} = 0.$$

It follows that the odd and even coefficients of H_N both sum up to $1/2$.

In each column of J_1 and J_2, either all odd or all even d_n's occur exactly once. Therefore, the sum of the line vectors of J_1 or J_2 is the vector $(1/2, \dots, 1/2)$. The same is true if we combine the second row of J with J_0. We see that the first row vector of J equals the sum of the other rows.

But $\frac{\partial F}{\partial y}(H_N)$ is the submatrix of J formed by those columns of F which correspond to coefficients in y. This does not change the sum of the rows of $\frac{\partial F}{\partial y}(H_N)$, i.e., the first row of $\frac{\partial F}{\partial y}(H_N)$ is a linear combination of the other rows and $\frac{\partial F}{\partial y}(H_N)$ cannot be regular. □

Nevertheless, the hard implicit-function theorem [1] ensures the existence of the implicit function under the given conditions. However, the hard implicit-function theorem is not constructive, i.e., it does not help to compute the tangent vectors of the implicit function. To circumvent this problem, we first solve for the implicit function without the first condition $F_1 : \sum h_k = 1$. $\tilde{F}$ denotes F without the first equation, i.e., we examine

$$\tilde{F} : I\!R^{(n+1)^2} \longrightarrow I\!R^{n^2+n+1} \tag{27}$$

and compute the tangent space at $H_N \in \mathcal{H}_2^N$ of the implicit function defined by

$$\tilde{F}(h) = 0.$$

If we later intersect this tangent space with the hyperplane $\sum h_k = 1$, we will obtain the tangent space of $\mathcal{H}_2$ at H_N.

There is not much choice for the n free coefficients: J_2 occurs $(n-1)/2$ times in J. J_2 has rank deficit 1, i.e., we need to choose one free coefficient for each occurrence of J_2. Combining the first $(n+1)$ entries of the first row of $\tilde{J}$ with J_0 gives a $(n+1)/2$-by-$(n+1)$ matrix, i.e., we have at least $(n+1)/2 = N$ free coefficients among the first $(n+1) = 2N$ coefficients. With our convention for numbering, the first $(n+1)$ coefficients are $m = (m_1, 0), m_1 = 0, \dots, n$. As free coefficients, we choose

$$x = (h_{(0,0)}, \dots, h_{N,0)}, h_{(0,2)}, h_{(0,4)}, \dots, h_{(0,2N-2)});$$

the remaining coefficients are collected in y. Let J_2' denote the matrix J_2 without the first column, and let J_0' denote J_0 without the first N columns: J_0' is an $(N-1) \times N$ matrix; J_2' is a $(2N-1) \times (2N-1)$ matrix; $\frac{\partial \tilde{F}}{\partial y}(H_N)$ is a square matrix with dimension

$4N^2 - 2N + 1 = n^2 + n + 1$:

$$
(28) \qquad \frac{\partial \tilde{F}}{\partial y}(H_N) = \left(\begin{array}{cccccccc}
d_{(n+1)/2}, \dots, d_n & 0 & \cdots & \cdots & \cdots & \cdots & & 0 \\
J_0' & & & & & & & \\
 & J_1 & & & & & & \\
 & & J_2' & & & & & \\
 & & & J_1 & & & & \\
 & & & \cdots & \cdots & \cdots & & \\
 & & & & & & J_2' & \\
 & & & & & & & J_1
\end{array} \right).
$$

We now consider the function

$$g : I\!R^n \longrightarrow I\!R^{n^2+n+1}$$

implicitly defined by $\tilde{F}(x, g(x)) = 0$. In order to obtain the tangent vectors of $\mathcal{H}_2$ at H_N, we compute the Jacobian g according to

$$\nabla g = -\left(\frac{\partial \tilde{F}}{\partial y}\right)^{-1} \frac{\partial \tilde{F}}{\partial x}.$$

$\frac{\partial \tilde{F}}{\partial x}$ consists of those columns of the Jacobian of $\tilde{F}$ which were not used in $\frac{\partial \tilde{F}}{\partial y}$; hence $\frac{\partial \tilde{F}}{\partial x}$ and ∇g are $(n^2 + n + 1) \times n$ matrices. We observe that $\frac{\partial \tilde{F}}{\partial x}$ does not use any columns involving the submatrix J_1, i.e., the rows of $\frac{\partial \tilde{F}}{\partial x}$ corresponding to coefficinets h_k with $k = (k_1, k_2)$, k_2 odd, are all zero. For $N = 2$, we obtain

$$J_2' = \begin{pmatrix} d_3 & 0 & 0 \\ d_1 & d_2 & d_3 \\ 0 & d_0 & d_1 \end{pmatrix},$$

$$\frac{\partial \tilde{F}}{\partial y} = \begin{pmatrix} d_2\, d_3 & 0 & 0 & 0 \\ d_0\, d_1 & 0 & 0 & 0 \\ 0 & J_1 & 0 & 0 \\ 0 & 0 & J_2' & 0 \\ 0 & 0 & 0 & J_1 \end{pmatrix}, \qquad \frac{\partial \tilde{F}}{\partial x} = \begin{pmatrix} d_0 & d_1 & 0 \\ d_2 & d_3 & 0 \\ & \mathcal{O} & \\ 0 & 0 & d_2 \\ 0 & 0 & d_0 \\ 0 & 0 & 0 \\ & \mathcal{O} & \end{pmatrix};$$

here $\mathcal{O}$ denotes a 4×3 matrix with zero entries.

Once we have computed ∇g, the tangent vectors t of $\mathcal{H}_1^N$ at H_N are obtained in the following way. We write the vector t naturally as an array $t = \{t_k \mid k = (k_1, k_2),\ 0 \leq k_1, k_2 \leq (2N-1)\}$. Then $dx = (dx_1, \dots, dx_n)$ corresponds to an array $dx = \{dx_k \mid k = (k_1, k_2),\ 0 \leq k_1, k_2 \leq (2N-1)\}$ whose entries are zero whenever its index k corresponds to a dependent variable y; similarly, the $(n^2 + n + 1)$ vector $\nabla g dx$ corresponds to an array whose entries are zero whenever its index corresponds to a free variable x. With this convention,

$$t = dx + \nabla dx$$

gives a desired tangent vector. In addition, we have to obey the omitted equation $0 = F_1(h) = \sum h_k - 1$. This is satisfied if we restrict the tangent vectors t to the hyperplane

$$\sum_{k\in\mathbb{Z}^2} t_k = 0.$$

LEMMA 3.7. *The tangent vectors $t = \{t_k \mid k = (k_1, k_2)\}$ of $\mathcal{H}_2$ at H_N satisfy*

$$t_k = 0 \quad \textit{if } k_2 \textit{ is odd.}$$

Proof. The derivatives of the implicit function g are the solution of the linear system

$$-\frac{\partial\tilde{F}}{\partial y}\,\nabla g = \frac{\partial\tilde{F}}{\partial x}.$$

(The columns of ∇g are the tangent vectors of g.) This large linear system splits up into smaller linear systems with matrices

$$\begin{pmatrix} d_{\frac{n+1}{2}},\ldots,d_n \\ J_0' \end{pmatrix},\ J_1,\ J_2'.$$

The components of the tangent vectors of g corresponding to odd lines—i.e., k_2 odd—are the solution of the subsystem

$$J_1 z = \mathcal{O}.$$

Lemma 3.6 proves $z = 0$, and hence $t = dx + \nabla g dx$ has zero entries whenever $k = (k_1, k_2)$, k_2 odd. □

Moreover, we can solve the linear systems with matrix J_2' explicitly. This gives the components of the tangent vectors corresponding to coefficients on even lines (k_2 even).

LEMMA 3.8. *Let $z = (z_1, \ldots, z_n)$ denote the solution vector of*

$$J_2' z = \begin{pmatrix} d_{n-1} \\ d_{n-3} \\ \cdot \\ d_0 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{pmatrix}.$$

Then $z_k = (-1)^{k-1}(d_{n-k}/d_n)$, $k = 1, \ldots, n$.

Proof. The Daubechies wavelet ψ_N is orthogonal to the scaling function φ_N:

$$\begin{aligned} 0 &= \langle\psi_N,\ \varphi_N(\cdot-\ell)\rangle \\ &= 4\sum_{m,k} d_k\ (-1)^m d_{1-m}\ \ \langle\varphi_N(2\cdot-m),\ \varphi_N(2\cdot-2\ell-k)\rangle \\ &= 2\sum_{k=0}^{n}(-1)^k d_k d_{1-2\ell-k}. \end{aligned}$$

On the other hand, the entries of the $n \times n$ matrix J_2' are given by

$$(J_2')_{\ell k} = d_{n+1-2\ell+k}, \quad 1 \leq \ell,\ k \leq n.$$

Let z be defined as above:

$$\begin{aligned}(d_n\ J_2'\ z)_\ell &= \sum_{k=1}^{n} d_{n+1-2\ell+k}(-1)^{k-1} d_{n-k} \\ &= \sum_{k=0}^{n-1}(-1)^k\ d_k d_{1-2(\ell-n)-k} \\ &= (-1)^{n+1} d_n d_{1-2(\ell-n)-n} = d_n d_{1+n-2\ell}.\end{aligned}$$

This gives—up to the factor d_n—the desired right-hand side. □

Thus far, we have obtained a complete description of the tangent vectors of the implicit function $g(x)$, i.e., the tangent directions of $\mathcal{H}_2$ at H_N are given by $dx + \nabla g dx$.

THEOREM 3.9. *Let $\mathcal{H}_2^N$ denote the set of coefficients*

$$h = \{h_k \mid k = (k_1, k_2),\ 0 \leq k_1, k_2 \leq n)$$

satisfying the 2D orthogonality relation. Let H_N denote the scaling coefficients associated with the induced 1D Daubechies wavelet ψ_N. The tangent space $\mathcal{T}_N$ of $\mathcal{H}_2^N$ at H_N is the direct sum

$$\mathcal{T}_N = T_1 \bigoplus T_2$$

of the linear spaces T_1 and T_2. T_1 is the induced tangent space of $\mathcal{H}_1^N$, i.e., it contains tangent vectors $t = \{t_k\}$ with $t_{(k_1,k_2)} = 0$ for $k_2 \neq 0$. T_2 is spanned by $(n-1)/2$ vectors t^m, $m = 1, \ldots, (n-1)/2$:

$$t^m = \left\{ t_{(k_1,k_2)} \mid t_{(k_1,k_2)} = \left\{ \begin{array}{cc} (-1)^{k_1-1} d_{n-k_1}, & k_2 = 2m, \\ 0 & elsewhere \end{array} \right\} \right\}.$$

COROLLARY 3.10. *In terms of Fourier polynomials, the tangent vectors at H_N, which are orthogonal to the induced 1D tangent vectors, are linear combinations of*

$$e^{-i2m\omega_2} \cdot G_N(\omega_1), \quad m = 1, \ldots, (n-1)/2,$$

where $G_N(\omega) = \sum_{k=0}^{n}(-1)^k d_{n-k} e^{-ik\omega}$ is the Fourier polynomial of the Daubechies wavelet coefficients.

For $N = 2$, the Daubechies scaling coefficients are explicitly given as

$$8 \begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ d_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 3 \\ 1 \end{pmatrix} + \sqrt{3} \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix}.$$

Remark. These are the coefficients of the scaling equation for the scaling function φ_2. The scaling equation for the wavelet ψ has coefficients $g_k = (-1)^k d_{1-k}$. The connection between wavelet coefficients and central difference quotients can be seen here. The g_k's are a simple combination of the central difference quotient of order 3 and the central difference quotient of order 2 on four points. In a similar way, one can easily parametrize all orthogonal 1D-wavelets by difference quotients.

The tangent vectors of $\mathcal{H}_2$ at H_2 are given by the induced tangent vectors s^1 and s^2,

$$s^1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & -\sqrt{3} & 1+\sqrt{3} \end{pmatrix}, \quad s^2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 1-\sqrt{3} & \sqrt{3} \end{pmatrix},$$

and the additional tangent vector t^1,

$$t^1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -1-\sqrt{3} & 3+\sqrt{3} & -3+\sqrt{3} & 1-\sqrt{3} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The tangent vectors are given as arrays, with the convention that, e.g., $s^1_{(0,3)} = -t^1_{(2,0)} = 1+\sqrt{3}$, $s^2_{(0,2)} = t^1_{(2,3)} = 1-\sqrt{3}$.

For the proofs of the results above, we used only that H_N satisfies the orthogonality condition (8). Hence the tangent spaces look the same for any induced $H \subset \mathcal{H}_1 \subset \mathcal{H}_2$ with $h_0 \neq 0 \neq h_n$, n odd. No further properties of the Daubechies wavelets were necessary.

Now we start computing the tangent vectors at

$$q(H_N) \in q(\mathcal{H}_2) = \mathcal{K}_{\text{orth}} \subset \mathcal{K}_2,$$

where q is defined by $q(H) = |H|^2$. The tangent vectors of $\mathcal{K}_{\text{orth}}$ at $q(H_N)$ are therefore computed as the linearization of

$$q(H_N(\omega_1) + se^{-i2k\omega_2}\, G_N(\omega_1)).$$

Since $\mathcal{K}_1$ is a flat subset of $\mathcal{K}_{\text{orth}}$, it follows that the Fourier polynomials of the induced 1D tangent vectors are linear combinations of

$$\cos(n\omega_1) - \cos((n-2m)\omega_1), \quad m = 1, \dots, (N-1)/2.$$

THEOREM 3.11. *Let $\{d_k\}$ denote the scaling coefficients of the Daubechies wavelet ψ_N, and define*

$$H_N(\omega_1) = \sum_{k=0}^{n} d_k e^{-ik\omega_1}, \qquad G_N(\omega_1) = \sum_{k=0}^{n} (-1)^k d_{n-k} e^{-ik\omega_1} = e^{-in\omega_1}\overline{H}_N(\omega_1).$$

The Fourier polynomials of the tangent vectors of $\mathcal{K}_{\text{orth}}$ at $q(H_N)$ are linear combinations of the induced 1D tangent vectors with

$$e^{i2k\omega_2}[H_N(\omega_1)\overline{G}_N(\omega_1)] + e^{-i2k\omega_2}[\overline{H}_N(\omega_1)G_N(\omega_1)], \quad k = 1, \dots, (n-1)/2.$$

Our idea was to compute the tangent spaces of $\mathcal{K}_{\text{orth}}$ with the hope that $\mathcal{K}_1$ is not the only flat subspace of $\mathcal{K}_{\text{orth}}$. But numerical tests did not reveal any other flat subspaces. Nevertheless, the simple structure of the tangents at $H_N \in \mathcal{H}_2$ allows us to trace part of H_2 by following the tangent lines.

4. Families of orthogonal 2D wavelets. In the last section, we studied the tangent space of the set $\mathcal{H}_2$ of coefficients $\{h_k\}$ satisfying the 2D orthogonality condition (15). $\mathcal{H}_2$ is not a flat manifold; therefore, starting at $H_N \in \mathcal{H}_2$ and following a tangent t does not lead directly to new wavelets. In this section, we discuss different correction terms c such that

$$H(\omega_1, \omega_2) = H_N(\omega_1) + \mu t + \mu c(\omega_1, \omega_2) \text{ correction term} \in \mathcal{H}_2.$$

Moreover, the orthogonality condition ensures only discrete orthogonality of the coefficient set. In order to prove that the solution of the scaling equation with coefficients $H = \{h_k\} \in \mathcal{H}_2$ is orthogonal to its translates, we also need to check the condition stated in Criterion 2.2; see [4]. However, not only do the Fourier series H_N of the scaling coefficients for the 1D Daubechies wavelets satisfy the 1D Cohen criterion with $K = [-\pi, \pi]$, but we also have

$$\forall m > 0,\ 0 \leq \epsilon < \pi : H_N(2^{-m}\omega_1) \text{ does not vanish on } [-\pi - \epsilon, \pi + \epsilon].$$

This follows directly from Lemma 2.6. This implies for our construction that the 2D Cohen criterion will be satisfied at least for small μ. For a special class of wavelets, see Lemma 4.3, we will check the orthogonality directly.

4.1. Twisted wavelets. Let $n = 2N - 1$ be an odd integer and let $H_0(\omega) \in \mathcal{H}_2$ denote the Fourier polynomial of an orthogonal 1D wavelet with $(n+1)$ coefficients, e.g., the Daubechies wavelets H_N. As shown in the previous section, the Fourier filter of the tangent vector is given by

$$G_0(\tau) = e^{-in\tau}\overline{H}_0(\pi + \tau)$$

—this denotes complex conjugation. The linear combination $H_0(\omega_1) + \mu e^{-i2k\omega_2}G_0(\omega_1)$ proceeds along a tangent line of $\mathcal{H}_2$. In order to stay on $\mathcal{H}_2$, we must add a correction term.

LEMMA 4.1. *Let H_0 and G_0 be defined as above. Define H_1 and G_1 by*

$$H_1(\tau) = (1/2)e^{-in\tau}\{e^{-i\tau}[\overline{H}_0(\tau) - \overline{H}_0(\pi + \tau)] + e^{i\tau}[\overline{H}_0(\tau) + \overline{H}_0(\pi + \tau)]\},$$

$$G_1(\tau) = (1/2)\{e^{i\tau}[H_0(\pi + \tau) - H_0(\tau)] \ + \ e^{-i\tau}[H_0(\pi + \tau) + H_0(\tau)]\}.$$

For any $k \in \mathbb{Z}$, $\mu \in \mathbb{R}$, the Fourier filter $m(\omega_1, \omega_2)$ defined by

$$(1 + \mu^2)\ m(\omega_1, \omega_2) = H_0(\omega_1) + \mu e^{-i2k\omega_2}G_0(\omega_1) - \mu\{G_1(\omega_1) - \mu e^{-i2k\omega_2}H_1(\omega_1)\}$$

is a solution of the 2D orthogonality condition, i.e., it constitutes a QMF.

Remark. H_1 also satisfies the 1D orthogonality condition, and G_1 is the corresponding tangent vector. The coefficients of H_1 come from a simple twist of the coefficients of H_0: if we pair the coefficients of H_0, i.e., $(h_{(0,0)}, h_{(1,0)}), (h_{(2,0)}, h_{(3,0)}), \ldots,$ $(h_{(n-1,0)}, h_{(n,0)})$, and reverse these pairs and their order, i.e.,

$$(h_{(n-1,0)}, h_{(n,0)}),\ (h_{(n-3,0)}, h_{(n-2,0)}), \ldots, (h_{(0,0)}, h_{(1,0)}),$$

then we obtain the coefficients of H_1.

Proof. It is a lengthy but easy calculation to show that m obeys the 2D orthogonality relation: the summands of $|m|^2$ are, e.g., of the form

$$2G_0\overline{H}_1(\tau) = e^{i\tau}\{H_0(\tau)\overline{H}_0(\pi+\tau) + |H_0(\pi+\tau)|^2\} + e^{-i\tau}\{H_0(\tau)\overline{H}_0(\pi+\tau) - |H_0(\pi+\tau)|^2\}.$$

It follows that

$$(G_0\overline{H}_1 + \overline{G}_0 H_1)(\tau) + (G_0\overline{H}_1 + \overline{G}_0 H_1)(\tau + \pi) = 0.$$

The other terms are treated in the same way. □

For these Fourier series, the coefficients are concentrated on two lines, namely $(m_1, 0)$ and $(m_1, 2k)$. Alterations of the above procedure lead to more equally spread coefficients, e.g.,

$$\begin{aligned}&(1+\mu^2)m(\omega_1,\omega_2)\\ &\quad = H_0(\omega_1) + \mu e^{-i(3\omega_2+\omega_1)}G_0(\omega_1) - \mu\{e^{-i(\omega_1+\omega_2)}G_1(\omega_1) - \mu e^{-i2\omega_2}H_1(\omega_1)\}.\end{aligned}$$

These series also solve the orthogonality relation.

Another class of trigonometric polynomials satisfying the orthogonality condition with a more convenient description can be found as follows.

THEOREM 4.2. *Let $H(\omega) \in \mathcal{H}_1$ denote the Fourier series of the scaling coefficients $\{h_k \mid k = 0, \dots, n\}$, n odd, for an orthogonal 1D wavelet. Denote the tangent vector by $G(\tau)$. Let P and Q be (π, π)-periodic trigonometric polynomials in (ω_1, ω_2) satisfying*

$$|P|^2 + |Q|^2 = 1, \qquad P(0) = 1.$$

Then

$$m(\omega_1, \omega_2) = P(\omega_1, \omega_2)\ H(\omega_1) \ +\ Q(\omega_1, \omega_2) G(\omega_1)$$

satisfies the 2D orthogonality relation, i.e., m constitutes a QMF.

Proof. Obviously

$$|m|^2 = |P|^2|H|^2 + |Q|^2|G|^2 + PH\overline{QG} + \overline{PH}QG.$$

Therefore, exploiting the periodicity of P and Q and the orthogonality of H leads to

$$\begin{aligned}&|m(\omega_1,\omega_2)|^2 + |m(\omega_1+\pi,\omega_2+\pi)|^2\\ &\quad = |P(\omega_1,\omega_2)|^2 + |Q(\omega_1,\omega_2)|^2 + (P\overline{Q})(\omega_1,\omega_2)[H\overline{G}(\omega_1) + H\overline{G}(\pi+\omega_1)]\\ &\qquad\qquad + (\overline{P}Q)(\omega_1,\omega_2)[\overline{H}G(\omega_1) + \overline{H}G(\pi+\omega_1)].\end{aligned}$$

n is odd and $G(\tau) = e^{-in\tau}\overline{H}(\pi + \tau)$. Therefore, the terms in the brackets vanish. □

Of course, this family of nonseparable QMF filters could have been found directly, but the structure of the tangent space inspired this choice. The relation for P and Q looks rather similar to the 2D orthogonality condition, but P and Q do not have to be related by a shift of the argument. For example, we can choose $|P|^2$ to be a power of $\cos(x)$ and construct the corresponding Q with the help of the Fejer–Riesz theorem (Theorem 2.4). This opens a way of controlling the zeros of m and might be helpful in constructing smooth wavelets for the quincunx grid.

We have to show that these Fourier series satisfy the 2D Cohen criterion. We will single out a special class and prove the orthogonality of the wavelets directly.

LEMMA 4.3. *Let H_N denote the Fourier series of the scaling coefficients for the 1D Daubechies wavelet ψ_N. As usual, $G_N(\omega)$ denotes the tangent vector. Let*

$$P(\omega_2) = H_N(2\omega_2) \quad \textit{and} \quad Q(\omega_1, \omega_2) = e^{i(\omega_1+\omega_2)}G_N(2\omega_2).$$

Then

$$m(\omega_1,\omega_2) = P(\omega_2)H_N(\omega_1) \ + \ Q(\omega_1,\omega_2)G_N(\omega_1)$$

is the Fourier series of the scaling coefficients for a compactly supported orthogonal 2D wavelet for the dilation matrix $A = \left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$.

Remarks. The exponential factor $e^{i(\omega_1+\omega_2)}$ can be replaced by any $e^{i(k_1\omega_1+k_2\omega_2)}$, where $k_1, k_2 \in \mathbb{Z}$, $k_1 + k_2$ are even, and $2k_1 - k_2 = 1$ or 2. Moreover, generalizations to other dilation matrices A with $|\det(A)| = 2$ are obvious.

Proof. We have to check the Cohen criterion. m has a set of trivial zeros at

$$(2n\pi, \pi/2 \ + k\pi), \ ((2n+1)\pi, \ k\pi), \quad k, n \in \mathbb{Z}.$$

If (ω_1, ω_2) is a nontrivial zero of m, then

$$|H_N(2\omega_2)|^2 |H_N(\omega_1)|^2 = |H_N(\pi + 2\omega_2)|^2 |H_N(\pi + \omega_1)|^2$$

or

$$\frac{|H_N(2\omega_2)|^2}{|H_N(\pi + 2\omega_2)|^2} = \frac{|H_N(\pi + \omega_1)|^2}{|H_N(\omega_1)|^2}.$$

$q = |H_N|^2$ is monotonically decreasing on $[0, \pi]$, $q(0) = 1$, $q(\pi) = 0$, q is an even function, and q' is a multiple of $\sin^{2N-1}(\omega)$. Hence there is no zero outside the lines

$$\omega_2 = \pm(1/2)(\omega_1 - \pi) + k\pi, \quad k \in \mathbb{Z}.$$

The lines $\omega_2 = -(1/2)(\omega_1 - \pi) + k\pi$ may be excluded since N is odd, $\overline{H}(\omega) = H(-\omega)$:

$$m(\omega_1, -(1/2)(\omega_1 - \pi) + k\pi) = H_N(\omega_1)H_N(\pi - \omega_1)[1 - e^{i(\omega_1/2+\pi/2+k\pi)}],$$

which is nonzero unless ω_1 is an odd multiple of π.

We have to find a set K congruent to $[-\pi, \pi]^2$ where $\hat{\varphi}$, a solution of the scaling equation with coefficients given by m, does not vanish. The Fourier transform of the scaling equation gives

$$\hat{\varphi}(\omega_1, \omega_2) = \prod_{j>0} m\left((A^t)^{-j} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}\right). \tag{29}$$

The previous consideration restricts zeros to lines

$$\omega_2 = (1/2)(\omega_1 - \pi) + k\pi$$

and their images under $(A^t)^j$. These lines intersect $[-\pi, \pi]^2$ at four line segments. Translating those line segments by $(0, 2\pi)$ yields a set K such that m does not vanish on $(A^t)^{-j}K$ with the possible exception of the critical points

$$\left(\frac{4}{5}\pi, \frac{2}{5}\pi\right), \ \left(\frac{2}{5}\pi, -\frac{4}{5}\pi\right), \ \left(-\frac{2}{5}\pi, \frac{4}{5}\pi\right), \ \left(-\frac{4}{5}\pi, -\frac{2}{5}\pi\right). \tag{30}$$

At these points,

$$|P(\omega_2)H_N(\omega_1)| \neq |Q(\omega_1, \omega_2)G(\omega_1)|, \quad \text{i.e.,} \quad m \neq 0.$$

Hence we have constructed a set K where $\hat{\varphi}$ does not vanish. A standard argument as in the proof of Lemma 3.1 of [6] shows that $\hat{\varphi}$ is bounded away from 0 on K. □

Naturally, we are interested in differentiable wavelets. The critical points in (30) form an orbit under the action of $A = \left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$. Hence we can obtain upper bounds for the smoothness of φ by applying the techniques of [4, Chaps. II.3.b and IV.3.b]. Also, generalizations of the methods introduced by T. Eirola [8] and L. Villemoes [18] to higher dimension will lead to Sobolev estimates for φ. However, preliminary calculation did not reveal any canonical candidates for smooth wavelets among the members of the family of wavelets described above. However, following Theorem 4.2, we have some freedom in choosing P and Q such that, e.g., m has a zero of a specified order at (π, π).

4.2. α-wavelets. In this section, we develop an iterative procedure for constructing QMFs, i.e., solutions of the 2D orthogonality condition. We again begin with an induced 1D wavelet and proceed along the tangential direction at this point. This implies that we obtain orthogonal wavelets for at least small values of the tangent parameter.

We begin by investigating Fourier filters with 16 coefficients $\{h_m \mid m = (m_1, m_2),\ 0 \le m_1, m_2 \le 3\}$. The results of §3 show that if we proceed along the tangential directions t, then

$$t_m = 0 \quad \text{if } m_1 \text{ is odd.}$$

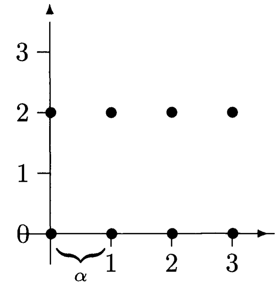


FIG. 2.

The position of the remaining eight coefficients in the grid $\mathbb{Z}^2$ are marked in Figure 2. These eight coefficients are grouped in pairs of two. In particular, the first two coefficicients on the baseline form the pair $\alpha = (h_{(0,0)}, h_{(1,0)})$.

Up to normalization the four coefficients on the baseline should satisfy the 1D orthogonality condition; this implies

$$(h_{(2,0)}, h_{(3,0)}) = \lambda\ \alpha^{\perp} = \lambda(-h_{(1,0)}, h_{(0,0)}).$$

Now we add a multiple of the tangent vector in line $m_2 = 2k$ and adjust α in order to satisfy the 2D orthogonality condition.

THEOREM 4.4. *Let $\alpha = (h_{(0,0)}, h_{(1,0)})$ and $\lambda, \mu \in \mathbb{R} \cup \{\pm\infty\}$ and define $\alpha^{\perp} = (-h_{(1,0)}, h_{(0,0)})$,*

$$h_2^{\alpha} = \begin{pmatrix} 0 & 0 \\ -\lambda\mu\alpha & \mu\alpha^{\perp} \\ 0 & 0 \\ \alpha & \lambda\alpha^{\perp} \end{pmatrix}. \tag{31}$$

If

$$\alpha = \frac{1}{2(1+\lambda^2)(1+\mu^2)}(1-\lambda\mu+\lambda+\mu,\ 1-\lambda\mu-\lambda-\mu),$$

then the Fourier series of h satisfies the 2D orthogonality relation.

Proof. Let H denote the Fourier filter of h and let $q = |H|^2$. We have to check that $q \in \mathcal{K}_2$, i.e., the coefficients of $\cos(k \cdot \omega)$, $k = (k_1, k_2) \neq (0,0)$ even, have to vanish. We can phrase this condition differently: the discrete convolution of h_2^α with itself has to vanish except at the origin. Since $\alpha \cdot \alpha^\perp = 0$, this is obviously the case.

We have two degrees of freedom left, i.e., $(h_{(0,0)}, h_{(1,0)})$, and we have to satisfy two more conditions:

$$\sum_m h_m = 1, \qquad \sum_m h_m^2 = 1/2.$$

The first conditon restricts $(h_{(0,0)}, h_{(1,0)})$ to a line in $\mathbb{R}^2$:

$$h_{(0,0)}(1-\lambda\mu+\lambda+\mu) + h_{(1,0)}(1-\lambda\mu-\lambda-\mu) = 1;$$

the second condition restricts $(h_{(0,0)}, h_{(1,0)})$ to a circle:

$$(h_{(0,0)}^2 + h_{(1,0)}^2)(1+\lambda^2)(1+\mu^2) = 1/2.$$

We show that the line is a tangent to the circle, i.e., the point on the line with minimal distance from the origin is also a point on the circle. Obviously, the α given in the theorem is the point on the line with minimal distance from the origin. We compute its norm:

$$\begin{aligned} \| \alpha \|_2^2 &= \frac{1}{4(1+\lambda^2)^2(1+\mu^2)^2}\{(1-\lambda\mu+\lambda+\mu)^2 + (1-\lambda\mu-\lambda-\mu)^2\} \\ &= \frac{1}{2(1+\lambda^2)(1+\mu^2)}. \qquad \square \end{aligned}$$

The pair of parameters $(\lambda, \mu) = (2+\sqrt{3}, 0)$ leads to the Daubechies wavelet ψ_2.

Remark. We can generate more QMFs by shifting the indices of the coefficients $(-\lambda\mu\alpha, \mu\alpha^\perp)$ in the line $m_2 = 2$ of h_2^α. Shifts of $\ell = (\ell_1, \ell_2)$ lead to coefficients satisfying the orthogonality relation as long as $\ell \neq (-2, 2k)$ and $\ell_1 + \ell_2$ is even. For example, let $\alpha = (\alpha_1, \alpha_2)$ and $\ell = (-1, 1)$; then

$$h = \begin{pmatrix} 0 & -\lambda\mu\alpha_1 & -\lambda\mu\alpha_2 & -\mu\alpha_2 & \mu\alpha_1 \\ \alpha_1 & \alpha_2 & -\lambda\alpha_2 & \lambda\alpha_1 & 0 \end{pmatrix}$$

defines a QMF. More possibilities arise from splitting $(-\lambda\mu\alpha, \mu\alpha^\perp)$ into its two components

$$\beta = -\lambda\mu\alpha \quad \text{and} \quad \gamma = \mu\alpha^\perp$$

and shifting both components separately by ℓ_β and ℓ_γ. The resulting set of coefficients h still satisfies the 2D orthogonality condition as long as $\ell_\beta \neq (-2, 2k)$, $\ell_\gamma \neq (-2, 2k)$, $\ell_\beta \neq \ell_\gamma + (2, 0)$. For example, $\ell_\beta = (1, -1)$ and $\ell_\gamma = (-1, -3)$ lead to

$$\tilde{h} = \begin{pmatrix} 0 & -\lambda\mu\alpha_1 & -\lambda\mu\alpha_2 & 0 \\ \alpha_1 & \alpha_2 & -\lambda\alpha_2 & \lambda\alpha_1 \\ 0 & -\mu\alpha_2 & \mu\alpha_1 & 0 \end{pmatrix}.$$

COROLLARY 4.5. *This set of coefficients $\tilde{h}$ is—up to normalization—identical to the family of QMF filters studied in* [11]. *In particular, the choice $(\lambda,\mu) = (-2+\sqrt{3}, -\sqrt{3})$ leads to a wavelet for the quincunx grid which is known to be continuous* [18].

Now we start iterating the procedure above. First, we choose arbitrary nonzero values for $\lambda = \mu_1, \mu_2, \ldots, \mu_l$ and compute. Let $h_2^\alpha = [h_2^0, h_2^1]$ denote the full set of coefficients, where $h_2^0 = (\alpha, \lambda\alpha^\perp)$ denotes the set of coefficients on the baseline and $h_2^1 = (-\mu_2\lambda\alpha, \mu_2\alpha^\perp)$ denotes the coefficients on the line $m_2 = 2k$. The iteration produces a family of coefficient sets $\{h_l^\alpha = [h_l^0, \mu_l h_l^1] \mid l \in N\}$ by

$$h_l^0 = (h_{l-1}^0,\ h_{l-1}^1),$$

$$h_l^1 = \left(-\mu_{l-1}\mu_l h_{l-1}^0, \frac{\mu_l}{\mu_{l-1}} h_{l-1}^1\right).$$

It follows by induction that

$$\begin{aligned} \|h_l^1\|_2^2 &= \mu_{l-1}^2\mu_l^2\|h_{l-1}^0\|_2^2 + \frac{\mu_l^2}{\mu_{l-1}^2}\|h_{l-1}^1\|_2^2 \\ &= \mu_l^2\|h_{l-1}^1\|_2^2 + \mu_l^2\|h_{l-1}^0\|_2^2 && (32) \\ &= \mu_l^2\|h_l^0\|_2^2. && (33) \end{aligned}$$

The question remains of how to choose values for $\alpha = (\alpha_1, \alpha_2)$. To this end, we introduce the auxiliary values

$$n_l^{01}, n_l^{02}, n_l^{11}, n_l^{12},$$

recursively defined by

$$n_1^{01} = 1, \qquad n_1^{02} = 1, \qquad n_1^{11} = \lambda, \qquad n_1^{12} = -\lambda,$$

$$n_l^{01} = n_{l-1}^{01} + n_{l-1}^{11}, \qquad n_l^{02} = n_{l-1}^{02} + n_{l-1}^{12},$$

$$n_l^{11} = \mu_l\left(-\mu_{l-1}n_{l-1}^{01} + \frac{1}{\mu_{l-1}}n_{l-1}^{11}\right), \qquad n_l^{12} = \mu_l\left(-\mu_{l-1}n_{l-1}^{02} + \frac{1}{\mu_{l-1}}n_{l-1}^{12}\right).$$

Then we define the vectors n_l and α by

$$n_l = \begin{pmatrix} n_l^{01} + n_l^{11} \\ n_l^{02} + n_l^{12} \end{pmatrix}, \qquad \alpha = \frac{n_l}{\|\, n_l \,\|_2^2}.$$

The procedure for constructing h_l^α starts by choosing arbitrary values $\lambda = \mu_1, \mu_2, \ldots, \mu_l$; then n_l, α, h_l^0, and h_l^1 are computed.

THEOREM 4.6. *The set of coefficients h_l^α satisfies the 2D orthogonality condition.*

Proof. The proof requires several induction arguments.

First, we have to check whether

$$\sum_{m\in\mathbb{Z}} (h_l^\alpha)_m = 1, \qquad \sum_{m\in\mathbb{Z}} (h_l^\alpha)_m^2 = 1/2. \tag{34}$$

From the definitions of $h_l^0, h_l^1, n_l^{01}, n_l^{02}, n_l^{11}$, and n_l^{12}, we easily deduce by induction that for any choice of $\alpha = (\alpha_1, \alpha_2)$, we have

$$\sum (h_l^0)_m = n_l^{01}\alpha_1 + n_l^{02}\alpha_2,$$

$$\sum (h_l^1)_m = n_l^{11}\alpha_1 + n_l^{12}\alpha_2.$$

Therefore, we easily obtain

$$\sum_{m\in\mathbb{Z}} (h_l^\alpha)_m = \sum (h_l^0)_m + \sum (h_l^1)_m = \alpha \cdot n_l = 1.$$

With the help of (32), we see that

$$\begin{aligned}\sum_{m\in\mathbb{Z}} (h_l^\alpha)_m^2 &= \sum (h_l^0)_m^2 + \sum (h_l^1)_m^2 \\ &= (1+\mu_l^2\mu_{l-1}^2)\sum (h_{l-1}^0)_m^2 + (1+\mu_l^2/\mu_{l-1}^2)\sum (h_{l-1}^1)_m^2 \\ &= (1+\mu_l^2)\sum_{m\in\mathbb{Z}} (h_{l-1}^\alpha)_m^2 \\ &= (1+\mu_l^2)\cdot\dots\cdot(1+\mu_2^2)(1+\lambda^2)(\alpha_1^2+\alpha_2^2) \\ &= (1+\mu_l^2)\cdot\dots\cdot(1+\mu_2^2)(1+\lambda^2)/\parallel n_l \parallel_2^2.\end{aligned}$$

The auxiliary values have been chosen such that (proof by induction)

$$n_l^{11} = \mu_l n_l^{02}, \qquad n_l^{12} = -\mu_l n_l^{01},$$

$$\parallel n_l \parallel_2^2 = (1+\mu_l^2) \parallel n_{l-1} \parallel_2^2, \qquad \parallel n_1 \parallel_2^2 = 2(1+\lambda^2).$$

The geometric interpretation of the above says that the equations in (34) restrict the choice of $\alpha = (\alpha_1, \alpha_2)$ to the intersection of a straight line with a circle. Incidentially, the line is tangential to the circle as in Theorem 4.5.

Finally, we have to check that the expansion of $|H|^2$ contains no even cos terms. However, let q denote the square modulus of the Fourier series associated with h_l^α. The terms $\cos(m\cdot\omega)$ for an even $m \neq (0,0)$ vanish by the same convolution argument as in the proof of Theorem 4.5. The only critical term is the coefficient of $\cos(2k\omega_2)$. The coefficient of $\cos(2k\omega_2)$ is given by

$$\langle h_l^1, h_l^0\rangle = -\mu_l\mu_{l-1}\langle h_{l-1}^0, h_{l-1}^0\rangle + \frac{\mu_l}{\mu_{l-1}}\langle h_{l-1}^1, h_{l-1}^1\rangle = 0.$$

The last equality follows from (32). □

Acknowledgments. The author would like to thank the referees for their careful reading of the manuscript and their suggestions for improving this paper.

REFERENCES

[1] J. Bemelmans, *Free Boundary Problems for the Navier–Stokes Equations*, in Lecture Notes in Math. 1357, S. Hildebrandt and R. Leis, eds., Springer-Verlag, Berlin, 1988, pp. 96–116.

[2] C. Chui, ed., *Wavelets: A Tutorial in Theory and Applications*, Academic Press, London, 1992.

[3] A. Cohen, *Ondelettes, analyses multiresolutions et filtres miroirs en quadrature*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 439–459.

[4] A. Cohen and I. Daubechies, *Non-separable bidimensional wavelet bases*, Rev. Mat. Iberoamericana, 9 (1993), pp. 51–138.

[5] D. Colella and C. Heil, *Characterizations of scaling functions* I: *Continuous solutions*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 496–518.

[6] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1989), pp. 909–996.

[7] ———, *Ten Lectures on Wavelets*, CBMS–NSF Monograph Series 61, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[8] T. Eirola, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[9] K. Gröchenig and W. Madych, *Multiresolution analysis, Haar bases and self similar tilings of $\mathbb{R}^n$*, IEEE Trans. Inform. Theory, 38 (1992), pp. 556–568.

[10] A. Grossmann, J. Morlet, and T. Paul, *Transforms associated to square integrable group representations* II: *Examples*, Ann. Inst. H. Poincaré Phys. Théor. 45 (1986), pp. 294–309.

[11] J. Kovacevic and M. Vetterli, *Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for $\mathbb{R}^n$*, IEEE Trans. Inform. Theory, 38 (1992), pp. 533–555.

[12] S. Mallat, *A theory for muliresolution signal decomposition: The wavelet representation*, IEEE Trans. Pattern Anal. Machine Intelligence, 11 (1989), pp. 674–693.

[13] Y. Meyer, *Wavelets and Operators*, Cambridge University Press, Cambridge, UK, 1992.

[14] G. Polya and G. Szegö, *Aufgaben und Lehrsätze der Analysis*, vol. II, Springer-Verlag, Berlin, 1971.

[15] A. Rieder, *Semi-algebraic multilevel methods based on wavelet decomposition*, East–West J. Numer. Math., 2 (1994), pp. 313–330.

[16] M. B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael, eds., *Wavelets and Their Applications*, Jones and Bartlett, Cambridge, MA, 1992.

[17] O. Rioul and M. Vetterli, *Wavelets and signal processing*, IEEE Signal Process. Magazine, 8 (1991), pp. 14–38.

[18] L. F. Villemoes, *Continuity of nonseparable quincunx wavelets*, Appl. Comput. Harmonic Anal., 1 (1994), pp. 180–187.

SIAM J. MATH. ANAL.
Vol. 27, No. 5, pp. 1482–1501, September 1996

DIMENSION AND LOCAL BASES OF HOMOGENEOUS SPLINE SPACES*

PETER ALFELD†, MARIAN NEAMTU‡, AND LARRY L. SCHUMAKER§

Abstract. Recently, we have introduced spaces of splines defined on triangulations lying on the sphere or on sphere-like surfaces. These spaces arose out of a new kind of Bernstein–Bézier theory on such surfaces. The purpose of this paper is to contribute to the development of a constructive theory for such spline spaces analogous to the well-known theory of polynomial splines on planar triangulations. Rather than working with splines on sphere-like surfaces directly, we instead investigate more general spaces of homogeneous splines in $\mathbb{R}^3$. In particular, we present formulas for the dimensions of such spline spaces, and construct locally supported bases for them.

Key words. multivariate splines, piecewise polynomial functions, homogeneous spline spaces, dimensions, sphere-like surfaces, sphere, interpolation, approximation, data fitting

AMS subject classifications. 41A63, 41A15, 65D07

1. Introduction. Let $\Delta := \{T^{[i]}\}_1^N$ be a planar triangulation of a set Ω, and let $0 \leq r \leq d$ be integers. The classical *space of splines of degree d and smoothness r* is defined by

$$\mathcal{S}_d^r(\Delta) := \{s \in C^r(\Omega) \ : \ s|_{T^{[i]}} \in \mathcal{P}_d, \quad i = 1, \ldots, N\}, \tag{1}$$

where $\mathcal{P}_d$ is the space of bivariate polynomials of degree at most d. These spaces of spline functions have found numerous applications in interpolation, data fitting, finite element solutions of boundary-value problems, computer aided geometric design, image processing, and elsewhere.

There is a well-developed (albeit incomplete) constructive theory for the polynomial spline spaces $\mathcal{S}_d^r(\Delta)$ which includes

(1) dimension formulas,
(2) construction of local bases,
(3) estimates on the approximation power,
(4) algorithms for manipulating the splines,
(5) algorithms for interpolation, data fitting, etc.

Recently [4], we introduced analogous spaces of splines defined on a triangulation on the sphere or on a sphere-like surface. As suggested by our companion paper [6], we believe that such spaces have important applications, and hence it is important to develop the analogous constructive theory.

Following [4], we will analyze spherical splines by investigating a more general class of splines associated with a *trihedral decomposition* $\mathcal{T} := \{T^{[i]}\}_1^N$ of a set $\Omega \subseteq \mathbb{R}^3$

*Received by the editors October 28, 1994; accepted for publication (in revised form) April 19, 1995.

†Department of Mathematics, University of Utah, Salt Lake City, UT 84112 (alfeld@math.utah.edu). The research of this author was supported by National Science Foundation grant DMS-9203859.

‡Department of Mathematics, Vanderbilt University, Nashville, TN 37240 (neamtu@orpheus.cas.vanderbilt.edu).

§Department of Mathematics, Vanderbilt University, Nashville, TN 37240 (s@mars.cas.vanderbilt.edu). The research of this author was supported by National Science Foundation grant DMS-9208413.

(see §2 below). Given such a decomposition, the associated spaces of *homogeneous splines* are defined by

$$\mathcal{H}_d^r(\mathcal{T}) := \{s \in C^r(\Omega) : s|_{T^{[i]}} \in \mathcal{H}_d, \quad i = 1, \dots, N\}, \tag{2}$$

where $\mathcal{H}_d$ denotes the space of trivariate polynomials of degree d which are homogeneous of degree d (recall that a function f defined on $\mathbb{R}^3$ is *homogeneous of degree* d provided $f(\alpha v) = \alpha^d f(v)$ for all real numbers α and all $v \in \mathbb{R}^3$). Splines on the sphere or on a sphere-like surface S are then obtained by restricting $\mathcal{H}_d^r(\mathcal{T})$ to S.

The main purpose of this paper is to establish dimension formulas for spaces of homogeneous splines and to show how to construct bases of locally supported splines. Homogeneous splines can be stored and evaluated using the algorithms presented in [4] for homogeneous polynomials. The question of the approximation power of homogeneous and spherical splines will be dealt with elsewhere. Applications to the interpolation and fitting of scattered data on the sphere or on a sphere-like surface are discussed in [6]. Even though we are working in $\mathbb{R}^3$, because of the nature of homogeneous polynomials—which are essentially bivariate functions—the entire development is closely modelled after the analysis of the bivariate spaces of splines $\mathcal{S}_d^r(\Delta)$ carried out in [8, 15, 16, 17].

2. Homogeneous spline spaces. We begin by introducing some notation, closely following [4].

DEFINITION 1. *Let $\{v_1, v_2, v_3\}$ be a set of linearly independent unit vectors in $\mathbb{R}^3$. We call*

$$T = \{v \in \mathbb{R}^3 : v = b_1 v_1 + b_2 v_2 + b_3 v_3 \quad \textit{with} \quad b_i \geq 0\} \tag{3}$$

the trihedron generated by $\{v_1, v_2, v_3\}$. *As in* [4], *we call the real numbers b_1, b_2, b_3 the* trihedral coordinates *of v with respect to T. They are homogeneous linear functions in the coordinates of v.*

We call the set $\{v \in T : b_i = 0\}$ the (ith) *face* of T, and the set $\{\alpha v_i : \alpha \geq 0\}$ the (ith) *ray* of T (or the ray generated by v_i). To avoid awkward repetitions, we abuse our notation slightly: in addition to writing v for a unit vector, we also use v to denote the associated point in $\mathbb{R}^3$ and the associated ray generated by v.

DEFINITION 2. *Let $\mathcal{T} = \{T^{[i]}\}_{i=1}^N$ be a nonempty set of trihedra, and let $\Omega := \cup T^{[i]}$. Then we call $\mathcal{T}$ a* trihedral decomposition of Ω *provided that*

(1) *the interiors of the trihedra in $\mathcal{T}$ are pairwise disjoint*;

(2) *the set $\Omega \cap S$ is homeomorphic to a two-dimensional disk or equals S, where S is the unit sphere*;

(3) *each face of a trihedron in $\mathcal{T}$ is either on the boundary of Ω or it is a common face of precisely two trihedra in $\mathcal{T}$.*

Each of the $T^{[i]} \cap S$ is a spherical triangle, and $\Delta = \{T^{[i]} \cap S\}_{i=1}^N$ is a *spherical triangulation*; cf. [19]. We say a trihedral decomposition $\mathcal{T}$ is *total* if $\Omega = \mathbb{R}^3$. Otherwise, we say that it is *partial.*

It will be convenient to denote the set of unit vectors defining the rays of the trihedra in $\mathcal{T}$ by $\mathcal{V}$. If $\mathcal{T}$ is a partial trihedral decomposition, it is natural to define rays to be *boundary rays* of $\mathcal{T}$ provided they are associated with vectors $v \in \mathcal{V}$ which lie on the boundary of Ω. All other rays will be called *interior rays.* We denote the sets of boundary and interior rays in $\mathcal{T}$ by $\mathcal{V}_B$ and $\mathcal{V}_I$, respectively. Clearly, all rays of a total trihedral decomposition are interior rays. Following the notation used for planar triangulations, we denote the number of boundary and interior rays of $\mathcal{T}$ by

V_B and V_I, respectively. Similarly, we denote the number of boundary and interior faces of $\mathcal{T}$ by E_B and E_I. For a partial decomposition, the number of rays is given by $V := V_B + V_I$, and the number of faces is given by $E := E_B + E_I$. For a total decomposition, $V = V_I$ and $E = E_I$.

Let T be a trihedron generated by $\{v_1, v_2, v_3\}$, and let b_1, b_2, b_3 denote the corresponding trihedral coordinates as functions of $v \in \mathbb{R}^3$. The *homogeneous Bernstein basis polynomials of degree d associated with* T are the polynomials

$$B^d_{ijk}(v) := \frac{d!}{i!j!k!} b_1^i b_2^j b_3^k, \qquad i+j+k=d, \tag{4}$$

which closely resemble bivariate Bernstein basis polynomials [12, 13, 14].

The space $\mathcal{H}_d$ of trivariate homogeneous polynomials is a $\binom{d+2}{2}$-dimensional linear space, and, as observed in [4], it is spanned by the set of $\binom{d+2}{2}$ Bernstein basis polynomials defined in (4). Thus each $p \in \mathcal{H}_d$ can be written uniquely in the form

$$p(v) = \sum_{i+j+k=d} c_{ijk} B^d_{ijk}(v). \tag{5}$$

In [4], p is referred to as a *homogeneous Bernstein–Bézier (HBB) polynomial of degree d.*

It will be convenient to define the *domain points* associated with T to be the points

$$P_{ijk} := \frac{iv_1 + jv_2 + kv_3}{d}, \qquad i+j+k=d. \tag{6}$$

In contrast to the case of polynomial splines on planar triangles, this definition of P_{ijk} is not the only natural one (see Remark 24 in §9).

If we look at all of the domain points for all of the trihedra in a trihedral decomposition, it is clear that the domain points associated with a common face of two trihedra coincide. If we eliminate such repetitions, we see that for a given trihedral decomposition $\mathcal{T}$, there are one point associated with each ray, $d-1$ points associated with each face, and $\binom{d-1}{2}$ associated with the interior of each trihedron. Thus the set $\mathcal{G}$ of distinct domain points has cardinality

$$\#(\mathcal{G}) = V + (d-1)E + \binom{d-1}{2} N. \tag{7}$$

The importance of the HBB form of homogeneous polynomials is that it provides a simple way to describe when two such polynomials defined on adjoining trihedra join together smoothly. Indeed, suppose $T^{[1]}$ and $T^{[2]}$ are two trihedra generated by the sets $\{v_1, v_2, v_3\}$ and $\{v_1, v_3, v_4\}$, respectively. Then as shown in [4], the two associated homogeneous polynomials $p^{[1]}$ and $p^{[2]}$ of degree d agree on the face shared by $T^{[1]}$ and $T^{[2]}$ in value and all derivatives up to order r if and only if

$$c^{[2]}_{ijk} = \sum_{\mu+\nu+\kappa=k} c^{[1]}_{i+\mu,\nu,j+\kappa} B^k_{\mu\nu\kappa}(v_4) \quad \text{for all} \quad k \le r, \quad i+j+k=d, \tag{8}$$

where $B^k_{\mu\nu\kappa}$ are the Bernstein basis polynomials of degree k associated with $T^{[1]}$.

By (8), $p^{[1]}$ and $p^{[2]}$ join *continuously* across their common face if and only if

$$c^{[2]}_{ij0} = c^{[1]}_{i0j}, \qquad i + j = d. \tag{9}$$

We conclude that a spline $s \in \mathcal{H}^0_d(\mathcal{T})$ is uniquely defined by a set of $\#(\mathcal{G})$ coefficients, one associated with each point $P \in \mathcal{G}$. This implies that the space $\mathcal{H}^0_d(\mathcal{T})$ has dimension $\#(\mathcal{G})$.

For later use, for each $P \in \mathcal{G}$, it will be convenient to define a linear functional λ_P defined on $\mathcal{H}^0_d(\mathcal{T})$ with the property that for any $s \in \mathcal{H}^0_d(\mathcal{T})$,

$$\lambda_P s = c_P, \tag{10}$$

where c_P is the coefficient associated with the point P. We denote the set of all such linear functionals by Λ. Clearly, $\#(\Lambda) = \#(\mathcal{G})$.

For each $\lambda \in \Lambda$, there is a unique spline $s_\lambda \in \mathcal{H}^0_d(\mathcal{T})$ such that

$$\gamma s_\lambda = \delta_{\gamma,\lambda}, \quad \text{all } \gamma \in \Lambda. \tag{11}$$

The spline s_λ has all coefficients equal to 0 except for the coefficient λs_λ which has value 1. By construction, s_λ has one of the following supports:

(1) a single trihedron T if the coefficient λs_λ is associated with a domain point in the interior of T;

(2) a pair of adjoining trihedra if the coefficient λs_λ is associated with a domain point in the interior of a face separating two trihedra;

(3) the union of all trihedra which share the ray v if the coefficient λs_λ is associated with the domain point v.

In view of these properties, we say that such splines have *local support*. The duality property (11) assures that the splines s_λ for $\lambda \in \Lambda$ are linearly independent, and since there are precisely $\#(\mathcal{G})$ of them, they form a basis for $\mathcal{H}^0_d(\mathcal{T})$.

To obtain analogous results for $\mathcal{H}^r_d(\mathcal{T})$, we follow [8, 15, 17]. To get an upper bound on dimension, we construct a *determining set* $\Gamma \subset \Lambda$ such that if $s \in \mathcal{H}^r_d(\mathcal{T})$,

$$\gamma s = 0 \quad \text{for all} \quad \gamma \in \Gamma \qquad \text{implies} \qquad s \equiv 0. \tag{12}$$

Then as shown in [8], $\dim \mathcal{H}^r_d(\mathcal{T})$ is bounded above by the cardinality of Γ. We can get a lower bound for the dimension (and construct a basis at the same time) if Γ is chosen so that for each $\lambda \in \Gamma$, there exists a spline $s_\lambda \in \mathcal{H}^r_d(\mathcal{T})$ satisfying

$$\gamma s_\lambda = \delta_{\gamma,\lambda}, \quad \text{all } \gamma \in \Gamma. \tag{13}$$

This duality implies that the splines $\{s_\lambda\}$ are linearly independent, and it follows that the dimension of $\mathcal{H}^r_d(\mathcal{T})$ is equal to the cardinality of Γ and that these splines form a basis. Such a set Γ is called a *minimal determining set.*

We close this section by presenting the main result of the paper. Its proof will be developed in the following sections.

THEOREM 3. *Let* $r \geq 0$ *and* $d \geq 3r + 2$. *Suppose* $\mathcal{T}$ *is a trihedral decomposition of a set* $\Omega \subseteq \mathbb{R}^3$. *Let*

$$\sigma := \sum_{v \in \mathcal{V}_I} \sigma_v, \quad \textit{where} \quad \sigma_v := \sum_{m=1}^{d-r} (r + m + 1 - me_v)_+ \tag{14}$$

and e_v is the number of distinct planes containing the faces that meet at the ray v. Then

$$\dim \mathcal{H}_d^r(\mathcal{T}) = (d-r)(d-2r)V - 2d^2 + 6dr - 3r^2 + 3r + 2 + \sigma \tag{15}$$

if $\mathcal{T}$ is a total decomposition, and

$$\begin{aligned} \dim \mathcal{H}_d^r(\mathcal{T}) =& \frac{(d-r+1)(d-r)}{2} V_B + (d-r)(d-2r)V_I \\ & - \frac{2d^2 - 6dr + 3r^2 - 3r - 2}{2} + \sigma \end{aligned} \tag{16}$$

if $\mathcal{T}$ is a partial decomposition. In either case, there exists a basis for $\mathcal{H}_d^r(\mathcal{T})$ consisting of splines such that the support of each spline is either a single trihedron, an adjoining pair, or the set of trihedra containing a single ray.

3. Minimal determining sets for splines on oranges.

In [8, 15], the key to analyzing the dimension of bivariate spline spaces was first to examine the special case of a *cell* consisting of a set of triangles sharing one vertex. In this section, we construct minimal determining sets for spline spaces on the trihedral analog of cells. In the context of tetrahedral decompositions, these were called oranges in [10, 20]. Throughout this section, we assume only that $0 \le r < d$.

DEFINITION 4. *A trihedral decomposition $\mathcal{O}$ consisting of a set of trihedra sharing one ray v is called an* orange. *We call v the axis of the orange*; *see Fig.* 1.

Suppose the trihedra in $\mathcal{O}$ are labeled in counterclockwise order as $T^{[1]}, T^{[2]}, \dots, T^{[N]}$ as we move around the axis v, where the rays of $T^{[\ell]}$ are v, v_ℓ, and $v_{\ell+1}$. If v is an interior ray, we have $v_{N+1} = v_1$. We can label the domain points in these trihedra as

$$P_{ijk}^{[\ell]} := \frac{iv + jv_\ell + kv_{\ell+1}}{d}, \qquad i+j+k = d. \tag{17}$$

THEOREM 5. *If $\mathcal{O}$ is an orange associated with a boundary ray v, then*

$$\dim H_d^r(\mathcal{O}) = \binom{d+2}{2} + (N-1)\binom{d-r+1}{2}. \tag{18}$$

If $\mathcal{O}$ is an orange associated with an interior ray v, then

$$\dim \mathcal{H}_d^r(\mathcal{O}) = \binom{r+2}{2} + N\binom{d-r+1}{2} + \sum_{m=1}^{d-r} (r+m+1-me)_+, \tag{19}$$

where e denotes the number of distinct planes shared by trihedra in $\mathcal{O}$.

Proof. Let Π be a plane which intersects the axis of $\mathcal{O}$ at a point w which is not the origin, and so that Π is perpendicular to the axis. The intersections with Π of those faces of $\mathcal{O}$ which contain the axis are rays in Π emanating from w. If we replace them with unit line segments with one end at w and then connect their endpoints in order, we get a planar triangulation Δ consisting of a set of triangles sharing the vertex w. Clearly, the restriction of a spline in $\mathcal{H}_d^r(\mathcal{O})$ to Δ is a spline in $\mathcal{S}_d^r(\Delta)$. Conversely, by the homogeneity of the splines in $\mathcal{H}_d^r(\mathcal{O})$, a spline in $\mathcal{S}_d^r(\Delta)$ extends uniquely to a spline in $\mathcal{H}_d^r(\mathcal{O})$. The two spaces $\mathcal{S}_d^r(\Delta)$ and $\mathcal{H}_d^r(\mathcal{O})$ are therefore isomorphic, and the dimension assertion follows from Theorem 2.2 in [17]. □

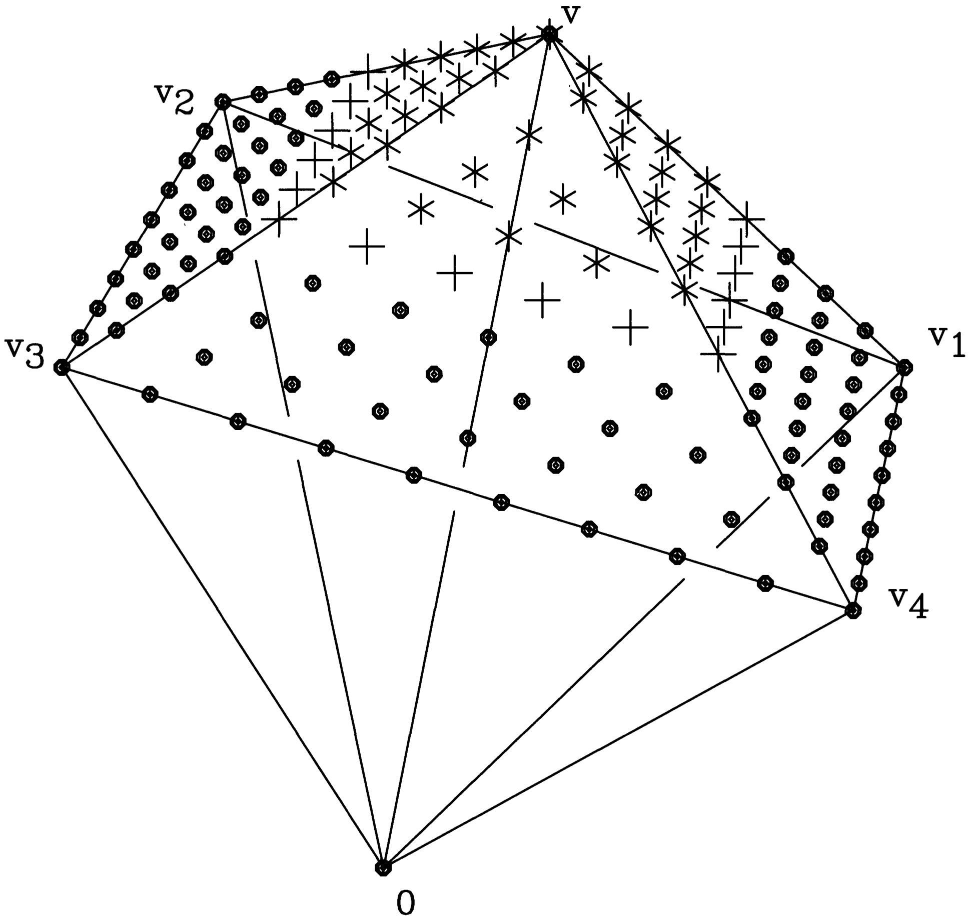


FIG. 1. *An orange with axis v.*

Following the proofs of Lemma 3.1 in [15] and Lemma 3.1 in [8], we now construct minimal determining sets for $\mathcal{H}_d^r(\mathcal{O})$ when $\mathcal{O}$ is an orange. We need the concept of a ring of domain points around a ray v.

DEFINITION 6. *Let $\mathcal{O}$ be an orange as above. Then given an integer d, the* mth ring *of $\mathcal{O}$ is the set of domain points*

$$\left\{P^{[\ell]}_{d-m,j,k} : j+k=m,\ \ell=1,2,\ldots,N\right\}. \tag{20}$$

*The m-*disk *in $\mathcal{O}$ is the union of the* 0*th through mth rings.*

The concepts of ring and disk are illustrated in Fig. 1. In particular, the domain points in the 5-ring around the vertex v in the figure are marked with + signs. The domain points in the 5-disk include all points marked with * or with +. To avoid cluttering the picture, the domain points in the far face (with vertices v, v_1, and v_2) have been omitted.

THEOREM 7. *Suppose $\mathcal{O}$ is an orange associated with a boundary ray v. Then the set*

$$\left\{P^{[1]}_{ijk} : i+j+k=d\right\} \cup \bigcup_{\ell=2}^{N} \left\{P^{[\ell]}_{ijk} : k \geq r+1\right\} \tag{21}$$

is a minimal determining set for $\mathcal{H}^r_d(\mathcal{T})$. Suppose $\mathcal{O}$ is an orange surrounding an interior ray v, and let e be as in Theorem 5. *Let*

$$\mu_{N-e+1} < \mu_{N-e+2} < \cdots < \mu_{N-1} = N, \quad \mu_N = N+1, \tag{22}$$

be such that the associated edges are pairwise noncollinear, and let

$$\mu_1 < \mu_2 < \cdots < \mu_{N-e} \tag{23}$$

be the complementary set so that

$$\{\mu_1, \mu_2, \ldots, \mu_N\} = \{2, 3, \ldots, N+1\}. \tag{24}$$

Let $\Gamma_0 \subseteq \Lambda$ be the set of functionals corresponding to domain points in the trihedron $T^{[1]}$. In addition, for each $m = 1, \ldots, d-r$, let Γ_m be the set of functionals corresponding to the first $Nm - (r+m+1) + (r+m+1-me)_+$ points in the ordered set

$$\{P^{[\mu_1]}_{d-m-r,m-1,r+1}, \ldots, P^{[\mu_1]}_{d-m-r,0,m+r}, \ldots, P^{[\mu_N]}_{d-m-r,m-1,r+1}, \ldots, P^{[\mu_N]}_{d-m-r,0,m+r}\}. \tag{25}$$

Then

$$\Gamma = \bigcup_{m=0}^{d-r} \Gamma_m \tag{26}$$

is a minimal determining set for $H^r_d(\mathcal{O})$.

Proof. We prove the result only for the case where the axis of the orange is an interior ray; the other case is similar. It is easy to check that the cardinality of Γ is given by the formula (19), and so we only need to show that Γ is a minimal determining set. To that end, consider the plane Π that is perpendicular to the vector v and passes through the point v. Explicitly,

$$\Pi = \left\{u \in \mathbb{R}^3 : (u-v)\cdot v = 0\right\}, \tag{27}$$

where $\cdot$ denotes the ordinary dot product. Let w_ℓ denote the orthogonal projection of v_ℓ onto Π, i.e.,

$$w_\ell = v_\ell + \left(1 - \frac{v_\ell \cdot v}{v \cdot v}\right) v, \quad \ell = 1, 2, \ldots, N. \tag{28}$$

The intersection of $\mathcal{O}$ with Π forms a two-dimensional cell Δ in the sense of [17]. Let Γ_Δ denote the functionals defined on $S^r_d(\Delta)$ corresponding to the projections of the points defining Γ. In view of the correspondence between bivariate polynomials and trivariate homogeneous polynomials, by Theorem 3.3 of [17], Γ_Δ is a (minimal) determining set of $S^r_d(\Delta)$. The fact that Γ is a determining set for $\mathcal{H}^r_d(\mathcal{O})$ now follows from a careful comparison of the smoothness conditions for $S^r_d(\Delta)$ and $\mathcal{H}^r_d(\mathcal{O})$. Any spline $s \in \mathcal{H}^0_d(\mathcal{O})$ can be expressed on the trihedron $T^{[\ell]}$ in the form (5) with $c_{ijk} = c^{[\ell]}_{ijk}$. To obtain the smoothness conditions for $\mathcal{H}^r_d(\mathcal{O})$, we write

$$v_{\ell+1} = r_\ell v + s_\ell v_{\ell-1} + t_\ell v_\ell, \tag{29}$$

where for convenience we treat all rays, domain points, and coefficients cyclically as we move around v (so that $v_0 = v_N$ for example). By (8), a spline in $H_d^0(\mathcal{O})$ belongs to $H_d^r(\mathcal{O})$ if and only if for all $\ell = 1, 2, \ldots, N$,

$$c_{ijk}^{[\ell+1]} = \sum_{\mu+\nu+\kappa=k} c_{i+\mu,\nu,j+\kappa}^{[\ell]} \frac{k!}{\mu!\nu!\kappa!} r_\ell^\mu s_\ell^\nu t_\ell^\kappa \tag{30}$$

for $k \leq r$ and $i + j + k = d$. Consider now the corresponding smoothness conditions for $S_d^r(\Delta)$. It can be checked that the projections of the v_ℓ satisfy

$$w_{\ell+1} = \tilde{r}_\ell v + s_\ell w_{\ell-1} + t_\ell w_\ell, \tag{31}$$

where

$$\tilde{r}_\ell := 1 - s_\ell - t_\ell. \tag{32}$$

Using a tilde to denote the coefficients of a spline in $S_d^r(\Delta)$, we obtain the conditions

$$\tilde{c}_{ijk}^{[\ell+1]} = \sum_{\mu+\nu+\kappa=k} \tilde{c}_{i+\mu,\nu,j+\kappa}^{[\ell]} \frac{k!}{\mu!\nu!\kappa!} \tilde{r}_\ell^\mu s_\ell^\nu t_\ell^\kappa, \qquad \ell = 1, \ldots, N. \tag{33}$$

We now show that Γ is a determining set for $\mathcal{H}_d^r(\mathcal{O})$. Consider a spline $s \in \mathcal{H}_d^r(\mathcal{O})$. We work our way through the rings of the orange. The 0th ring is v itself. It is in Γ and therefore the coefficient corresponding to it must be zero. Suppose now that the coefficients corresponding to the first m rings are all zero and consider the $(m+1)$th ring and the smoothness conditions (30) and (33) for $k = m$. In spite of r_ℓ and $\tilde{r}_\ell$ being different, these equations are equivalent since the terms where $r_\ell \neq 0$ and $\tilde{r}_\ell \neq 0$ contain coefficients which are zero by the induction hypothesis. Thus the coefficients of s must vanish on the $(m+1)$th ring and it follows that Γ is a determining set. Since it has cardinality equal to the dimension of $H_d^r(\mathcal{O})$, it follows that Γ is minimal. □

Remark 8. The argument used in the proof of Theorem 7 applies to all minimal determining sets which have $\#\Gamma_m$ points on the $(r+m)$th ring for $m = 1, \ldots, d-r$. However, it is not true in general that the analogue of a minimal determining set for a two-dimensional cell is also a minimal determining set for a corresponding orange as is shown in the following example.

Example 9. Let $\mathcal{O}$ be an orange with $N = 4$, $v_3 = -v_1$ and $v_4 = -v_2$.

Discussion. Figure 2 shows a minimal determining set for $S_2^1(\Delta)$ that is not determining for $H_2^1(\mathcal{O})$, where points corresponding to functionals not in the set are marked with a dot, and the functionals corresponding to all other points are in the set. Note that in particular the functional corresponding to the center point (which is at v) is *not* in the set. Clearly, in the two-dimensional cell, the coefficients at the points marked with a crosshair ($\oplus$) or a triangle ($\triangle$) determine the coefficient at v. In fact, we have

$$\bar{c}_v = \frac{\bar{c}_\oplus + c_\triangle}{2}. \tag{34}$$

On the other hand, the relevant smoothness condition for $H_2^1(\mathcal{O})$ is

$$c_\oplus = -c_\triangle, \tag{35}$$

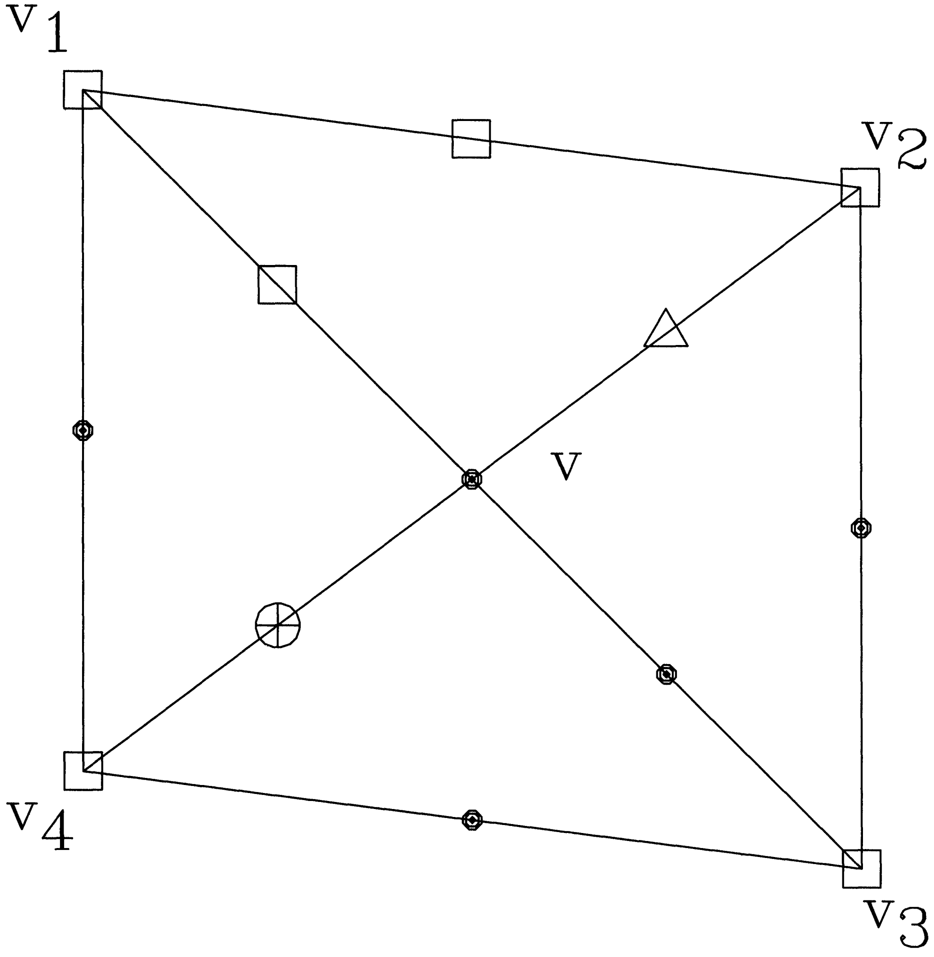


FIG. 2. *A nondetermining set.*

and so the two points marked with $\oplus$ or $\triangle$ cannot both be in the minimal determining set. It is of course easy to construct sets that are minimal determining for both $S_2^1(\Delta)$ and $H_2^1(\mathcal{O})$. An example (conforming to Theorem 7) can be obtained from Fig. 2 by replacing the point marked with $\oplus$ with v. □

4. A minimal determining set for $\mathcal{H}_d^r(\mathcal{T})$ when $d \geq 3r+2$. In this section, we construct a minimal determining set Γ for $\mathcal{H}_d^r(\mathcal{T})$ in the case where $d \geq 3r+2$. As in the bivariate case [8, 15], the key to the construction is to partition the Bézier coefficients into suitable subsets. Consider a trihedron T generated by the vectors v_1, v_2, v_3, and let $\mathcal{P} := \{P_{ijk}\}_{i+j+k=d}$ be the associated set of Bézier coefficients. To make the description of Γ easier, we recall the correspondence between coefficients, domain points, and the associated linear functionals,

$$c_{ijk} \sim P_{ijk} \sim \lambda_{P_{ijk}}, \tag{36}$$

and work only with domain points P_{ijk} here. We define the *distance of P_{ijk} from the* *ray* v to be

$$\text{dist}\,(P_{ijk}, v) := \begin{cases} d-i & \text{if } v = v_1, \\ d-j & \text{if } v = v_2, \\ d-k & \text{if } v = v_3. \end{cases} \tag{37}$$

For $i = 1, 2, 3$, let

$$\begin{aligned} \text{‘}\mathcal{D}_\mu(v_i) &:= \big\{P \in \mathcal{P} : \text{dist}(P, v_i) \le \mu\big\}, \\ \mathcal{A}(v_i) &:= \big\{P \in \mathcal{P} : \text{dist}(P, v_i) > \mu,\ \text{dist}(P, v_{i+1}) \ge d - r, \\ &\qquad\qquad \text{dist}(P, v_{i+2}) \ge d - r\big\}, \\ \mathcal{F}(v_i) &:= \big\{P \in \mathcal{P} : \text{dist}(P, v_i) \ge d - r\big\}, \\ \mathcal{E}(v_i) &:= \big\{P \in \mathcal{F}(v_i) : |\text{dist}(P, v_{i+1}) - \text{dist}(P, v_{i+2})| \le d - 3r - 2\big\}, \\ \mathcal{B}_L(v_i) &:= [\mathcal{F}(v_i) \cap \mathcal{D}_{2r}(v_{i+1})] \setminus [\mathcal{D}_\mu(v_{i+1}) \cup \mathcal{A}(v_{i+1}) \cup \mathcal{E}(v_i)], \\ \mathcal{B}_R(v_i) &:= [\mathcal{F}(v_i) \cap \mathcal{D}_{2r}(v_{i+2})] \setminus [\mathcal{D}_\mu(v_{i+2}) \cup \mathcal{A}(v_{i+2}) \cup \mathcal{E}(v_i)], \\ \mathcal{C} &:= \big\{P \in \mathcal{P} : \text{dist}(P, v_j) < d - r,\ \ j = i, i+1, i+2\big\}, \end{aligned} \tag{38}$$

where

$$\mu := r + \left\lfloor \frac{r+1}{2} \right\rfloor, \tag{39}$$

and we identify $v_4 = v_1$ and $v_5 = v_2$.

The set $\mathcal{D}_\mu(v_i)$ contains the points in a disk around v_i of radius μ. $\mathcal{A}(v_i)$ (called a *cap* in [15]) is the set of points not in $\mathcal{D}_\mu(v_i)$ but whose corresponding coefficients are involved in smoothness conditions of order up to r across the two faces sharing v_i. The sets $\mathcal{E}(v_i)$, $\mathcal{B}_L(v_i)$, and $\mathcal{B}_R(v_i)$ include only domain points whose corresponding coefficients are involved in smoothness conditions across the face opposite the ray v_i. Finally, $\mathcal{C}$ corresponds to coefficients which do not enter any smoothness conditions.

In Fig. 3, we have marked the domain points associated with one trihedron for the case $d = 23$ and $r = 6$ to show which of the above sets they belong to. Dots correspond to points in the sets $\mathcal{D}_\mu(v_i)$, circles to points in the sets $\mathcal{E}(v_i)$, asterisks to points in the caps $\mathcal{A}(v_i)$, plus signs to points in the sets $\mathcal{B}_L(v_i)$ and $\mathcal{B}_R(v_i)$, and ×'s to points in the set $\mathcal{C}$.

As in the bivariate case, in order to describe a minimal determining set for $\mathcal{H}_d^r(\mathcal{T})$, we have to take account of certain degenerate faces. In [15], an edge F of a planar triangulation is defined to be degenerate at one of its endpoints v if the edges preceding and succeeding F and connected to v are collinear. We require a similar concept for trihedral decompositions.

Definition 10. *Let F be an interior face of a trihedral decomposition $\mathcal{T}$, and let v be one of the two rays generating it. We say that F is* degenerate *at v if the faces other than F of the two trihedra sharing F and meeting in v are coplanar.*

We also need to adapt the familiar concept of a singular vertex.

Definition 11. *An interior ray v of a trihedral decomposition $\mathcal{T}$ is said to be* singular *if it has precisely four faces meeting at v which lie in two distinct planes.*

In contrast to the planar case where an edge can be degenerate at only one endpoint, for trihedral decompositions, it is possible for a face to be degenerate at both of the rays defining it, see Example 19 below. We are now ready to describe a minimal determining set Γ for $\mathcal{H}_d^r(\mathcal{T})$ in the case $d \ge 3r + 2$.

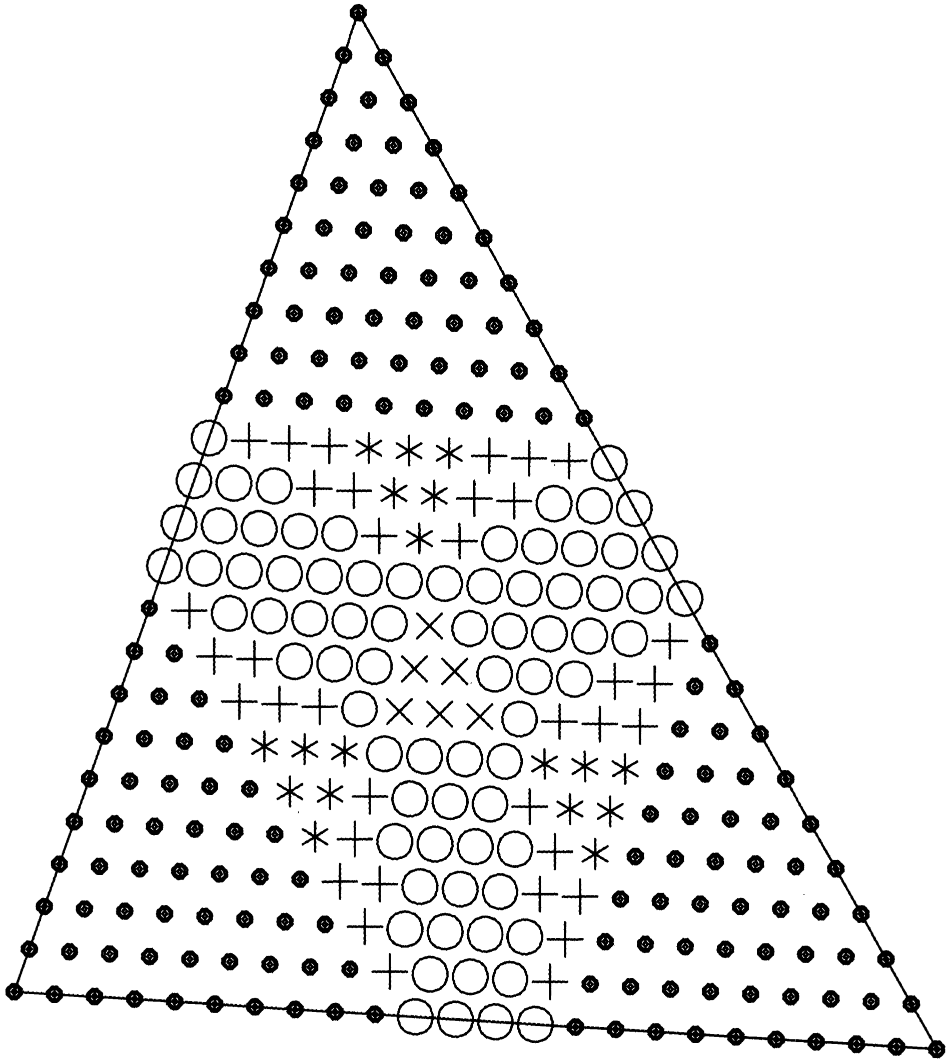


FIG. 3. *Division of domain points by Algorithm* 12; $d = 23$, $r = 6$, $\mu = 9$.

ALGORITHM 12. If $d \geq 3r + 2$, choose the set Γ as follows:

(1) For each *interior ray* v of $\mathcal{T}$, choose a minimal determining set as described in Theorem 7 for the space $\mathcal{H}_d^r(\mathcal{T})$ restricted to the μ-disk of $\mathcal{O}_v$, where $\mathcal{O}_v$ is the orange surrounding v.

(2) For each *boundary ray* v of $\mathcal{T}$, choose a minimal determining set as described in Theorem 7 for the space $\mathcal{H}_d^r(\mathcal{T})$ restricted to the μ-disk of $\mathcal{O}_v$, where $\mathcal{O}_v$ is the orange containing v.

(3) For each *trihedron* T in $\mathcal{T}$, choose the functionals corresponding to $\mathcal{C}$ and all three of the sets $\mathcal{A}(v_i)$ associated with T.

(4) For each *face* F in T, include the functionals corresponding to the set $\mathcal{E}(v)$ associated with a ray v in an adjoining trihedron and opposite to F. If F is a boundary face, there is only one such trihedron, while if it is an interior face, we can work with either of the two trihedra sharing it. If F is a boundary face, also include

the functionals associated with the two sets $\mathcal{B}_L(v)$ and $\mathcal{B}_R(v)$.

(5) Suppose that v is an interior vertex and that m of the faces attached to v are degenerate at v. Then for each such face F, remove the functionals corresponding to the cap nearest to v in the triangle preceding F (in counterclockwise order), and replace them by the functionals in the set $\mathcal{B}_L$ associated with F and lying in the same triangle. If F is degenerate at both of its ends, carry out this step at each end. It is easy to see that m can only be 1, 2, or 4. For an illustration of this step in the case $m = 1$, see Figs. 1 and 2 in [15]).

(6) If v is singular, add the functionals corresponding to one cap $\mathcal{A}(v)$ in one of the trihedra containing v.

THEOREM 13. *Let $\mathcal{T}$ be a trihedral decomposition and let $d \geq 3r+2$ and $r \geq 0$. Then the set Γ constructed in Algorithm* 12 *is a minimal determining set for $\mathcal{H}_d^r(\mathcal{T})$, and its cardinality is given by* (15) *if $\mathcal{T}$ is total and by* (16) *if $\mathcal{T}$ is partial. For each $\lambda \in \Gamma$, there exists a unique spline $s_\lambda \in \mathcal{H}_d^r(\mathcal{T})$ such that* (13) *holds. Then $\{s_\lambda\}_{\lambda\in\Gamma}$ forms a basis for $\mathcal{H}_d^r(\mathcal{T})$ such that the support of each spline is either a single trihedron, an adjoining pair, or an orange.*

Proof. We give the proof only in the case where $\mathcal{T}$ is total since the case where it is partial is very similar. First, we observe that the cardinalities of the sets defined in (38) are as follows:

$$\begin{aligned}
\#\mathcal{D}_\mu(v_i) &= \binom{\mu+2}{2},\\
\#\mathcal{A}(v_i) = \#\mathcal{B}_L(v_i) = \#\mathcal{B}_R(v_i) &= \binom{2r-\mu+1}{2},\\
\#\mathcal{E}(v_i) &= dr+d-12\mu r-3\mu-1+6r^2+4\mu^2,\\
\#\mathcal{C} &= \binom{d-3r-1}{2}.
\end{aligned} \tag{40}$$

Moreover, the sets are pairwise disjoint and their union is the set of all domain points in the trihedron $\mathcal{T}$.

Next, we show that the cardinality of the set Γ is given by (15) when $\mathcal{T}$ is total. It can be shown that in this case

$$N = 2(V-2) \quad\text{and}\quad E = 3(V-2), \tag{41}$$

where E is the number of faces of $\mathcal{T}$. Note that step 5 of Algorithm 12 does not change the cardinality of Γ, and that step 2 does not contribute since there are no boundary rays. With these observations, it follows from Algorithm 12 and Theorem 7 that

$$\begin{aligned}
\#\Gamma = {} & \sum_{v\in\mathcal{V}}\left[\binom{r+2}{2}+E_v\binom{\mu-r+1}{2}+\tilde{\sigma}_v\right] && \text{(step 1)}\\
+ {} & N\left[\binom{d-3r-1}{2}+3\binom{2r-\mu+1}{2}\right] && \text{(step 3)}\\
+ {} & E\left[dr+d-12\mu r-3\mu-1+6r^2+4\mu^2\right] && \text{(step 4)}\\
+ {} & K\binom{2r-\mu+1}{2}, && \text{(step 6)}
\end{aligned} \tag{42}$$

where E_v is the number of interior faces meeting at the ray v, K is the number of singular rays, and

$$\tilde{\sigma}_v := \sum_{m=1}^{\mu-r}(r+m+1-me_v)_+\,. \tag{43}$$

Using

$$\sum_{v\in\mathcal{T}} E_v = 2E \tag{44}$$

and (41), the equality of the right-hand sides of (42) and (15) follows after a straightforward manipulation. (Note that for singular rays, the $\tilde{\sigma}_v$ and the factor multiplying K combine to produce σ_v.)

We now show that Γ is a determining set for $\mathcal{H}_d^r(\mathcal{T})$. In the absence of degenerate faces, this follows as in [15]. For a degenerate face, note that the coefficients corresponding to points in the cap moved in step 5 of Algorithm 12 are implied to be zero by the smoothness conditions (8) across the degenerate face, independent of the possible relocation of other caps.

To complete the proof, we now construct a basis for $\mathcal{H}_d^r(\mathcal{T})$ satisfying (13). Clearly, for a given $\lambda \in \Gamma$, we can set the coefficient $\lambda s = 1$ and all other coefficients corresponding to $\gamma \in \Gamma$ with $\gamma \neq \lambda$ to zero, we can solve for the remaining coefficients using the smoothness conditions. If the domain point P corresponding to λ is contained in a set $\mathcal{C}$, then the resulting spline s_λ has support on the trihedron T containing P. If P is in a set of the form $\mathcal{E}(v_i)$, then s_λ has support on the union of the two trihedra containing the face opposite v_i. In all other cases, s_λ has support on an orange. □

Remark 14. Instead of constructing an explicit basis, it is also possible to prove the dimension statement in Theorem 13 by showing that the expressions in (15) provides a *lower bound* on $\dim \mathcal{H}_d^r(\mathcal{T})$ as was done in [1] in the planar case. This is done by thinking of $\mathcal{H}_d^r(\mathcal{T})$ as a subspace of $\mathcal{H}_d^0(\mathcal{T})$, enforcing the smoothness conditions in the μ-disks via Theorem 7, and then subtracting the number of appropriate smoothness conditions (8) needed to enforce smoothness across the interior faces of $\mathcal{T}$.

5. A minimal determining set for $\mathcal{H}_d^r(\mathcal{T})$ when $d \geq 4r+1$. As in the case of splines defined on a planar triangulation [8], the construction of a minimal determining set can be greatly simplified if $d \geq 4r+1$. In this case, the disks of radius $2r$ around rays of T do not overlap, and the remaining smoothness conditions across faces of $\mathcal{T}$ decouple. In that case, the following much simpler algorithm can be used:

ALGORITHM 15. If $d \geq 4r+1$, choose the set Γ as follows:

(1) For each *interior ray* v of $\mathcal{T}$, choose a minimal determining set for $\mathcal{H}_{2r}^r(\mathcal{O}_v)$ as described in Theorem 7, where $\mathcal{O}_v$ is the orange surrounding v.
(2) For each *boundary ray* v of $\mathcal{T}$, choose a minimal determining set for $\mathcal{H}_{2r}^r(\mathcal{O}_v)$ as described in Theorem 7. where $\mathcal{O}_v$ is the orange containing v.
(3) For each *trihedron* T in $\mathcal{T}$, choose $\mathcal{C}$.
(4) For each *face* in $\mathcal{T}$, choose

$$\tilde{\mathcal{E}}(v_1) = \mathcal{F}(v_1) \setminus [\mathcal{D}_{2r}(v_2) \cup \mathcal{D}_{2r}(v_3)], \tag{45}$$

where v_1, v_2, v_3 define a trihedron such that v_2 and v_3 span the face.

For the case $d = 23$ and $r = 5$, Fig. 4 shows the choice of the domain points for a single trihedron T using Algorithm 15. As in Fig. 3, dots correspond to points in sets of the form $\mathcal{D}_{2r}(v_i)$ and circles correspond to points in $\tilde{\mathcal{E}}(u_i)$, while ×'s mark the points in $\mathcal{C}$.

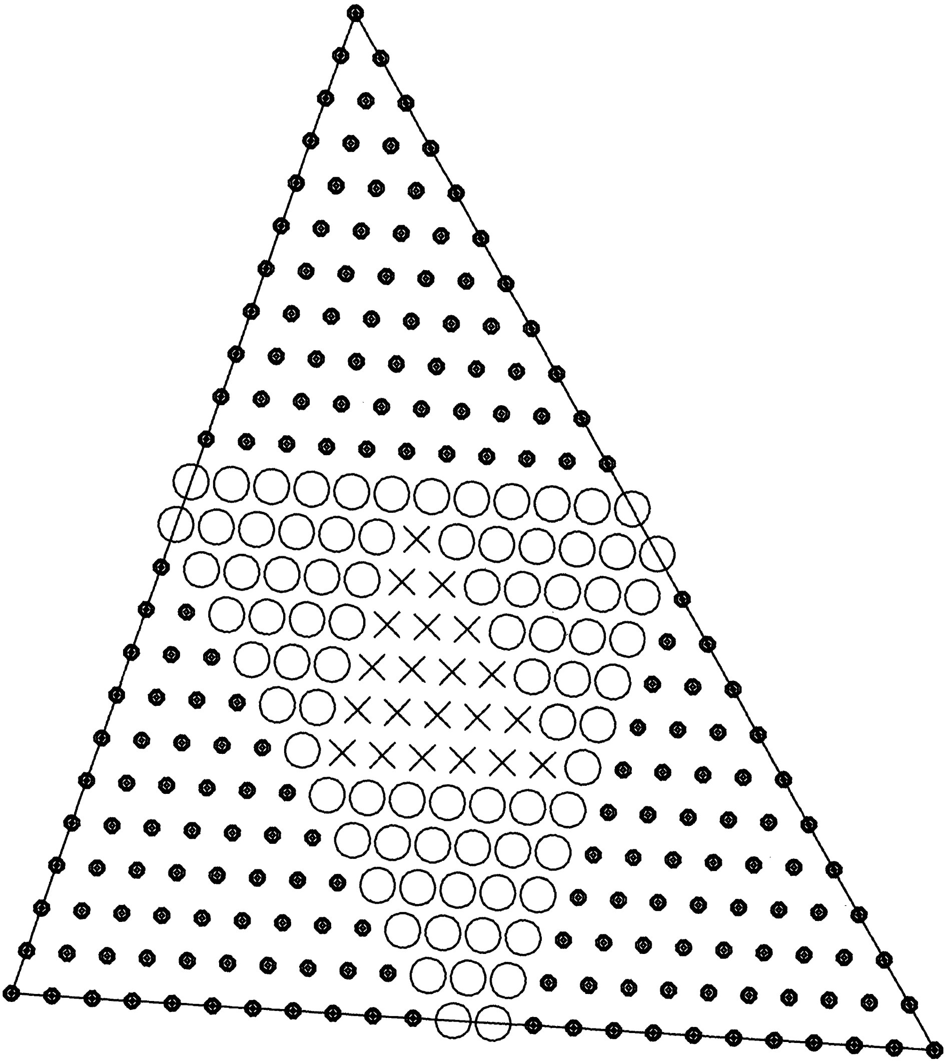


FIG. 4. *Division of domain points by Algorithm* 15; $d = 23$, $r = 5$.

6. The case $d \leq 3r+1$. As in the planar case, it is also possible to treat spline spaces for $d \leq 3r+1$ provided we restrict the class of trihedral decompositions somewhat.

THEOREM 16. *Let $d = 3r+1$, and suppose that the trihedral decomposition $\mathcal{T}$ does not possess any degenerate faces. Then the dimension of $H_{3r+1}^r(\mathcal{T})$ is given by* (15) *or* (16), *depending on whether $\mathcal{T}$ is total or partial. Moreover, there exists a basis with local supports as in Theorem* 13.

Proof. A minimal determining set can be constructed by an obvious adaptation of the prescription given in [9] for the planar case. □

It is also of interest to consider certain generic decompositions; see [11] for the planar case.

DEFINITION 17. *A trihedral decomposition $\mathcal{T}$ is said to be* generic with respect to r and d *provided that for all sufficiently small perturbations of the rays of $\mathcal{T}$, the* *resulting trihedral decomposition* $\tilde{\mathcal{T}}$ *satisfies*

$$\dim H_d^r(\tilde{\mathcal{T}}) = \dim H_d^r(\mathcal{T}). \tag{46}$$

THEOREM 18. *Fix* $d \in \{2,3,4\}$, *and suppose* $\mathcal{T}$ *is a generic trihedral decomposition with respect to* $r = 1$ *and* d. *Then the dimension of* $H_d^1(\mathcal{T})$ *is given by* (15) *or* (16), *depending on whether* $\mathcal{T}$ *is total or not.*

Proof. The space $H_d^1(\mathcal{T})$ is isomorphic to the space $\mathcal{S}_d^1(\Delta)$, where Δ is the generalized triangulation (see [11]) obtained by projecting the points in $\mathcal{V}$ through the origin onto a plane that does not contain the origin and is not parallel to any of the rays in $\mathcal{T}$. The result then follows from Theorems 27 and 33 in [11]. □

The proof of Theorem 18 does not involve finding a minimal determining set. For $d = 4$, it may be possible to construct one using the techniques in [7]. However, in the case $d \in \{2,3\}$, no general procedure for finding a minimal determining set is known even in the (generic) planar case.

7. Doubly degenerate faces. While the structure of bivariate splines on planar triangulations and homogeneous splines on trihedral decompositions in $\mathbb{R}^3$ are very similar, there is a situation which can occur in the homogeneous case but cannot occur in the planar case: it is possible for a face to be degenerate at both rays. We illustrate this in the following example.

Example 19. Let

$$v_i = -v_{i+3} = e_i, \quad i = 1,2,3, \tag{47}$$

where e_i denote the standard unit vectors, and let $\mathcal{T}^*$ be the set of trihedra generated by the sets

$$\begin{array}{llll} \{v_1,v_2,v_3\}, & \{v_1,v_2,v_6\}, & \{v_1,v_3,v_5\}, & \{v_1,v_5,v_6\}, \\ \{v_2,v_3,v_4\}, & \{v_2,v_4,v_6\}, & \{v_3,v_4,v_5\}, & \{v_4,v_5,v_6\}. \end{array} \tag{48}$$

The convex hull of these points forms a regular octahedron; see Fig. 5. In the resulting trihedral decomposition, each face is degenerate at each of its two rays, and at each ray each face sharing the ray is contained in one of only two planes. Thus all rays of $\mathcal{T}^*$ are singular.

As a check on our formulas and to provide actual numbers for comparison purposes, we have computed the dimensions of $H_d^r(\mathcal{T}^*)$ in Example 19 for $1 \le r \le 5$ and $1 \le d \le 15$ by setting up the smoothness conditions and numerically computing the rank of the matrix describing the smoothness conditions using the Goliath package [2, 3] and other special purpose software. For the trihedral decomposition $\mathcal{T}^*$, there are six singular rays. Thus for $d \ge 2r$, the expression (15) becomes

$$\begin{aligned} \phi_d^r &:= 4d^2 - 12dr + 9r^2 + 3r + 2 + 6\sum_{m=1}^{d-r}(r+1-m)_+ \\ &= 2\,(2d^2 - 6dr + 6r^2 + 3r + 1). \end{aligned} \tag{49}$$

This gives

$$\phi_d^r = \begin{cases} 4d^2 - 12d + 20 & \text{if } d \ge 2 \text{ and } r = 1, \\ 4d^2 - 24d + 62 & \text{if } d \ge 4 \text{ and } r = 2, \\ 4d^2 - 36d + 128 & \text{if } d \ge 6 \text{ and } r = 3, \\ 4d^2 - 48d + 218 & \text{if } d \ge 8 \text{ and } r = 4, \\ 4d^2 - 60d + 252 & \text{if } d \ge 10 \text{ and } r = 5. \end{cases} \tag{50}$$

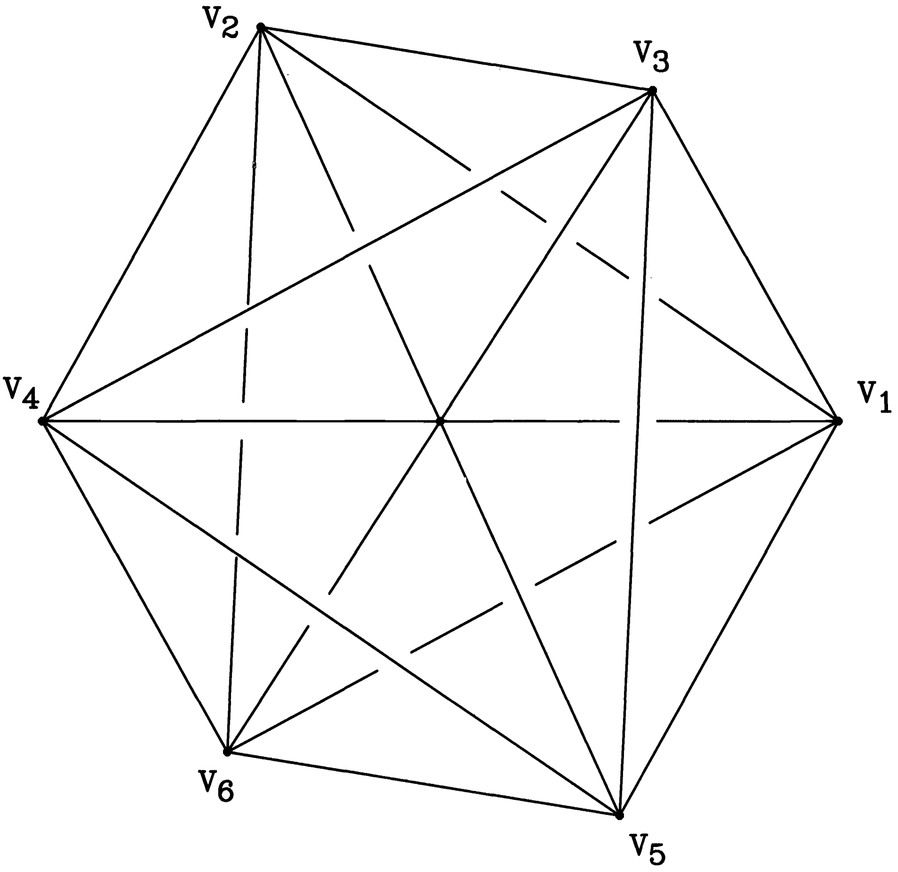


FIG. 5. *The regular octahedron.*

In Table 1, we have used an asterisk to mark those cases where the computed dimensions of $H_d^r(\mathcal{T}^*)$ differ from the values of ϕ_d^r. As a curiosity, we note that for the trihedral decomposition $\mathcal{T}^*$, the formulas are in fact correct for $d = 3r + 1$ (and of course all larger values) but not for $d \le 3r$, even though $\mathcal{T}^*$ is not generic and all faces are degenerate.

TABLE 1

Dimensions of $H_d^r(\mathcal{T}^*)$ *on the regular octahedron.*

d:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$r = 1$:	3*	9*	19*	36	60	92	132	180	236	300	372	452	540	636	740
$r = 2$:	3*	6*	13*	24*	39*	61*	90	126	170	222	282	350	426	510	602
$r = 3$:	3*	6*	10*	18*	30*	46*	66*	93*	127*	168	216	272	336	408	488
$r = 4$:	3*	6*	10*	15*	24*	37*	54*	75*	100*	132*	171*	217*	270	330	398
$r = 5$:	3*	6*	10*	15*	21*	31*	45*	63*	85*	111*	141*	178*	222*	273*	331*

8. Super splines. As in the planar case [15], the methods above can also be used to compute the dimension and to construct locally supported bases for spaces of homogeneous *super splines*:

$$\mathcal{H}_d^{r,\theta}(\mathcal{T}) := \{s \in \mathcal{H}_d^r(\mathcal{T}) : s \in C^{\rho_v}(v), \quad v \in \mathcal{V}\}, \tag{51}$$

where $\theta := \{\rho_v\}_{v \in \mathcal{V}}$ and $r \le \rho_v < d$ for all v. Here $s \in C^{\rho_v}(v)$ means that all of the derivatives up to order ρ_v of the pieces of s which join at v have a common value at v. We assume throughout that the μ- and ρ_v-disks around neighboring rays do not overlap, i.e.,

$$\max\{\mu, \rho_u\} + \max\{\mu, \rho_v\} < d \tag{52}$$

for all pairs of vertices u and v which generate a face of $\mathcal{T}$, where μ is defined in (39).

THEOREM 20. *Let $\mathcal{T}$ be a partial trihedral decomposition and suppose that $d \ge 3r + 2$ and that* (52) *holds. Then*

$$\begin{aligned}\dim \mathcal{H}_d^{r,\theta}(\mathcal{T}) = {} & \left[(d-r)(d-2r) - \binom{r+2}{2}\right] V + \sum_{v \in \mathcal{V}} \binom{\rho_v + 2}{2} \\ & - \sum_{v \in \mathcal{V}} \left[E_v \binom{\rho_v - r + 1}{2}\right] - 2d^2 + 6dr - 3r^2 + 3r + 2 \\ & + \sum_{v \in \mathcal{V}} \sum_{m=\rho_v - r+1}^{d-r} (r + m + 1 - m e_v)_+ \end{aligned} \tag{53}$$

if $\mathcal{T}$ is a total trihedral partition, and

$$\begin{aligned}\dim \mathcal{H}_d^{r,\theta}(\mathcal{T}) = {} & \frac{(d-r+1)(d-r)}{2} V_B + \left[(d-r)(d-2r) - \binom{r+2}{2}\right] V_I \\ & - \frac{2d^2 - 6dr + 3r^2 - 3r - 2}{2} \\ & - \sum_{v \in \mathcal{V}} \left[E_v \binom{\rho_v - r + 1}{2}\right] + \sum_{v \in \mathcal{V}_I} \binom{\rho_v + 2}{2} \\ & + \sum_{v \in \mathcal{V}_I} \sum_{m=\rho_v - r+1}^{d-r} (r + m + 1 - m e_v)_+ \end{aligned} \tag{54}$$

if $\mathcal{T}$ is a partial trihedral decomposition. Here E_v is the number of interior faces attached to the vertex v for each $v \in \mathcal{V}$. Moreover, there exists a basis of splines for $\mathcal{H}_d^{r,\theta}(\mathcal{T})$ such that the support of each spline is either a single trihedron, an adjoining pair, or an orange.

Proof. nWe give the proof for the case of a partial trihedral decomposition. The proof when the decomposition is total is similar (and simpler). The key observation is that the set of points chosen by Algorithm 12 and lying inside the disk $\mathcal{D}_{\rho_v}(v)$ is a minimal determining set for $\mathcal{H}_{\rho_v}^r(\mathcal{O}_v)$, where $\mathcal{O}_v$ is the orange with axis v. Thus, if we now impose C^{ρ_v} continuity at v, then we can replace those $\dim \mathcal{H}_{\rho_v}^r(\mathcal{O}_v)$ points by $\binom{\rho_v+2}{2}$ points lying in one trihedron in $\mathcal{O}_v$. This shows that for each ray v, the change in the number of points in the minimal determining set Γ constructed by Algorithm 12 is given by

$$\dim H_{\rho_v}^r(\mathcal{O}_v) - \dim H_{\rho_v}^{\rho_v}(\mathcal{O}_v) = \dim H_{\rho_v}^r(\mathcal{O}_v) - \binom{\rho_v + 2}{2}. \tag{55}$$

Thus

$$\begin{aligned}\dim\mathcal{H}_d^{r,\theta}(\mathcal{T}) &= \dim\mathcal{H}_d^r(\mathcal{T}) - \sum_{v\in\mathcal{V}}\left[\dim H^r_{\rho_v}(\mathcal{O}_v) - \binom{\rho_v+2}{2}\right]\\ &= \frac{(d-r+1)(d-r)}{2}V_B + (d-r)(d-2r)V_I\\ &\quad - \frac{2d^2-6dr+3r^2-3r-2}{2} + \sum_{v\in\mathcal{V}}\sum_{m=1}^{d-r}(r+m+1-me_v)_+\\ &\quad - \sum_{v\in\mathcal{V}_I}\left[\binom{r+2}{2} + E_v\binom{\rho_v-r+1}{2} + \sum_{m=1}^{\rho_v-r}(r+m+1-me_v)_+\right]\\ &\quad - \sum_{v\in\mathcal{V}_B}\left[\binom{\rho_v+2}{2} + E_v\binom{\rho_v-r+1}{2}\right] + \sum_{v\in\mathcal{V}}\binom{\rho_v+2}{2}.\end{aligned}\tag{56}$$

Now, combining terms, we get (54).

Our new minimal determining set can now be used to construct a basis of locally supported splines as was done in the proof of Theorem 3. □

The case where all ρ_v are equal is of particular interest.

COROLLARY 21. *Suppose that*

$$\rho_v = \rho \geq r, \quad v \in \mathcal{V},\tag{57}$$

and that $2\rho < d$. *Then*

$$\begin{aligned}\dim\mathcal{H}_d^{r,\theta}(\mathcal{T}) &= \frac{(2d^2-6dr-3r^2+12r\rho+3r-5\rho^2-3\rho)}{2}V\\ &\quad + (-2d^2+6rd+3r^2-3r+6\rho^2-12r\rho+6\rho+2)\\ &\quad + \sum_{v\in\mathcal{V}}\sum_{m=\rho-r+1}^{d-r}(r+m+1-me_v)_+\end{aligned}\tag{58}$$

if $\mathcal{T}$ *is a total trihedral partition, and*

$$\begin{aligned}\dim\mathcal{H}_d^{r,\theta}(\mathcal{T}) &= \frac{(d^2-2rd-r^2+d+r-2\rho^2+4\rho r-2\rho)}{2}V_B\\ &\quad + \frac{(2d^2-6rd-3r^2+12\rho r+3r-5\rho^2-3\rho)}{2}V_I\\ &\quad + \frac{(-2d^2+6rd+3r^2-3r+6\rho^2-12r\rho+6\rho+2)}{2}\\ &\quad + \sum_{v\in\mathcal{V}_I}\sum_{m=\rho-r+1}^{d-r}(r+m+1-me_v)_+\end{aligned}\tag{59}$$

if $\mathcal{T}$ *is a partial trihedral decomposition. Moreover, there exists a basis of splines for* $\mathcal{H}_d^{r,\theta}(\mathcal{T})$ *such that the support of each spline is either a single trihedron, an adjoining pair, or an orange.*

Proof. Substituting (57) in (53) and using (41) and (44) leads to (58). For a partial decomposition, the classical Euler relations for a triangulation imply

$$\sum_{v\in\mathcal{V}} E_v = 2E_I = 2(V_B+3V_I-3).\tag{60}$$

Now, substituting (57) in (54) and using (60) leads to (59). □

In both Theorem 20 and Corollary 21, the formula for a total trihedral decomposition can be obtained from the formula for a partial one by dropping the term with V_B and doubling the constant term. Moreover, if we set $\rho = r$ in the corollary, of course, we recover the formulas in Theorem 3.

9. Remarks.

Remark 22. The proof of Theorem 7 is based on the proof of Theorem 3.3 of [17] for polynomial splines on planar triangulations. The description of the minimal determining set for a cell given there is not quite correct in that it allows $\mu_{N-1} < N$, which could lead to the same point being included in Γ twice. This is easily fixed by requiring that $\mu_{N-1} = N$, as we have done here.

Remark 23. As in our paper [5], it is possible to develop a theory of homogeneous splines defined on a (total or partial) decomposition of $\mathbb{R}^2$ by wedges (the two-dimensional analogs of trihedra). Such splines can be restricted to a circle or a similar curve to obtain univariate functions along the curve. The corresponding dimensions and minimal determining sets can be obtained in a straightforward manner by considering a single ring in Theorem 7.

Remark 24. In the bivariate polynomial spline case, there is no question that the right way to define domain points P_{ijk} associated with the Bernstein–Bézier coefficients of a polynomial is by the formula in (6). In that case, the set of pairs $\{P, c_P\}_{P\in\mathcal{G}}$ is called the *Bézier net* of s and has an important geometric interpretation. However, in the trihedral setting, it is not so clear what the best way is to define the analogous points. As discussed in [4], there are reasonable alternatives, although it appears that there is no definition which carries the full geometric significance of the domain points in the planar case. Our choice here is a useful way to label control coefficients.

Remark 25. For polynomial spline spaces on planar triangulations, there are well-known lower and upper bounds on the dimension of $\mathcal{S}_d^r(\Delta)$ which are of interest for $d < 3r + 2$; see, e.g., [18] and references therein. Similar bounds can be derived for our homogeneous spline spaces and will be treated elsewhere.

Remark 26. The formula (54) given in Theorem 20 for a partial trihedral decomposition is much simpler than the corresponding formula in Theorem 2.4 of [15]. Since our proof of Theorem 20 can also be used in the bivariate case, the simpler formula (54) is also valid there.

REFERENCES

[1] P. Alfeld, *On the dimension of multivariate piecewise polynomials*, in Numerical Analysis, D. F. Griffiths and G. A. Watson, eds., Longman Scientific and Technical, Harlow, UK, 1986, pp. 1–23.

[2] P. Alfeld and D. Eyre, *The exact analysis of sparse rectangular linear systems*, ACM Trans. Math. Software, 17 (1991), pp. 502–518.

[3] ———, *Algorithm* 701, *Goliath: A software system for the exact analysis of rectangular rank-deficient sparse rational linear systems*, ACM Trans. Math. Software, 17 (1991), pp. 519–532.

[4] P. Alfeld, M. Neamtu, and L. L. Schumaker, *Bernstein–Bézier polynomials on spheres and sphere-like surfaces*, Comput. Aided Geom. Design, to appear.

[5] ———, *Circular Bernstein–Bézier polynomials*, in Mathematical Methods for Curves and Surfaces, M. Daehlen, T. Lyche, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1995, pp. 11–20.

[6] ———, *Fitting scattered data on sphere-like surfaces using spherical splines*, J. Comput. Appl. Math., 74 (1996), to appear.

[7] P. ALFELD, B. PIPER, AND L. L. SCHUMAKER, *An explicit basis for C^1 quartic bivariate splines*, SIAM J. Numer. Anal., 24 (1987), pp. 891–911.

[8] P. ALFELD AND L. L. SCHUMAKER, *The dimension of bivariate spline spaces of smoothness r for degree $d \geq 4r + 1$*, Constr. Approx., 3 (1987), pp. 189–197.

[9] ———, *On the dimension of bivariate splines spaces of smoothness r and degree $d = 3r + 1$*, Numer. Math., 3 (1990), pp. 651–661.

[10] P. ALFELD, L. L. SCHUMAKER, AND M. SIRVENT, *On dimension and existence of local bases for multivariate spline spaces*, J. Approx. Theory, 70 (1992), pp. 243–264.

[11] P. ALFELD, L. L. SCHUMAKER, AND W. WHITELEY, *The generic dimension of the space of C^1 splines of degree $d \geq 8$ on tetrahedral decompositions*, SIAM J. Numer. Anal., 30 (1993), pp. 889–920.

[12] C. DE BOOR, *B-form basics*, in Geometric Modeling: Algorithms and New Trends, G. E. Farin, ed., Society for Industrial and Applied Mathematics, Philadelphia, 1987, pp. 131–148.

[13] ———, *Multivariate piecewise polynomials*, in Acta Numerica, A. Iserless, ed., Cambridge University Press, Cambridge, UK, 1993, pp. 65–109.

[14] G. FARIN, *Curves and Surfaces for Computer Aided Geometric Design, A Practical Guide*, 2nd ed., Academic Press, New York, 1990.

[15] A. IBRAHIM AND L. L. SCHUMAKER, *Superspline spaces of smoothness r and degree $d \geq 3r + 2$*, Constr. Approx., 7 (1991), pp. 401–423.

[16] J. MORGAN AND R. SCOTT, *A nodal basis for C^1 piecewise polynomials in two variables*, Math. Comp., 29 (1975), pp. 736–740.

[17] L. L. SCHUMAKER, *Dual bases for spline spaces on cells*, Comput. Aided Geom. Design, 5 (1988), pp. 277–284.

[18] ———, *Recent progress on multivariate splines*, in The Mathematics of Finite Elements and Applications VII, J. Whiteman, ed., Academic Press, London, 1991, pp. 535–562.

[19] ———, *Triangulation methods in CAGD*, IEEE Comput. Graph. Appl., 13 (1993), pp. 47–52.

[20] M. SIRVENT, *The dimension of multivariate spline spaces*, Ph.D. thesis, Department of Mathematics, University of Utah, Salt Lake City, UT, 1990.

SIAM J. MATH. ANAL.
Vol. 27, No. 6, pp. 1503–1519, November 1996

ON A REPRESENTATION FORMULA FOR B. TEMPLE SYSTEMS*

S. BENZONI-GAVAGE†

Abstract. The author gives an inf-sup representation formula associated with genuinely nonlinear characteristic fields of B. Temple systems of conservation laws which reduces to the Lax formula in the convex scalar case. The proof is derived by means of geometrical arguments together with a method of characteristics. It holds for piecewise-smooth entropy solutions to the Cauchy problem for a large variety of initial data including Riemann data.

Key words. systems of conservation laws, B. Temple systems, weak entropy solutions

AMS subject classifications. 35L45, 35L60, 35L65, 35L67

1. Introduction.

In the theory of hyperbolic systems of conservation laws, the class of B. Temple systems [7] is known to generalize the properties of scalar conservation laws to a certain extent. Indeed, let $f : \Omega$ convex domain $\subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be a $\mathcal{C}^2$ strictly hyperbolic flux function, i.e., such that for all $u \in \Omega$, the Jacobian matrix $df(u)$ has n distinct real eigenvalues [2].

DEFINITION 1. *We say that a characteristic field associated with a given eigenvalue belongs to the B. Temple class if the associated left eigenvector field is orthogonal to a foliation by hyperplanes of Ω. By B. Temple system, we mean that every characteristic field belongs to the B. Temple class.*

This is a very strong requirement which implies two "scalar-like properties" for B. Temple systems. First, the elementary wave curves are straight lines. This result was part of the original paper by Temple [7]. Second, Serre [5] has shown that such systems are *rich*, that is, one can construct an infinite number of entropies. In [6], Serre proved several integrability properties of such systems. More precisely, he showed that if $v_x = u$ and $v_t = -f(u)$ with f a B. Temple flux function according to Definition 1, then the graph $\mathcal{G} = \{x, t, v(x,t)\}$ of v lies in a surface $\Sigma \subset \mathbb{R}^{n+2}$ given by the intersection of n envelopes of hyperplanes. This result is not complete, however. It does not enable us to distinguish entropy solutions from other weak solutions. This means that Σ may contain loops—corresponding to the occurrence of shocks (in the genuinely nonlinear case)—that should be eliminated from $\mathcal{G}$ (which in principle is the graph of a Lipschitz-continuous function). The inf-sup representation formula presented in this paper does involve some entropy criteria. Since it reduces to the Lax formula [2] in the convex scalar case, it may be a way to erase the irrelevant parts of the graph in all cases.

It is well known that shocks are closely related to genuinely nonlinear fields.

DEFINITION 2 (Lax). *A characteristic field associated with an eigenvalue λ and a right eigenvector r is genuinely nonlinear if $d\lambda \cdot r$ never vanishes on Ω.*

The aim of this paper is to produce as many representation formulas as there are genuinely nonlinear fields in a given B. Temple system. Thus for any λ-shock in an entropy solution of $u_t + f(u)_x = 0$, the ambiguity regarding the location of v in the envelope may be eliminated through the scalar representation formula associated with the λ-field.

In order to prove those formulas, we shall first show that for each genuinely nonlinear field, certain kinds of generalized convexity inequalities are satisfied. Then it

* Received by the editors June 1, 1994; accepted for publication (in revised form) August 1, 1995.

† Unité de Mathématiques Pures et Appliquées (CNRS UMR 128), École Normale Supérieure de Lyon, 46 Allée d'Italie, F-69364 Lyon cedex 07, France (benzoni@umpa.ens-lyon.fr).

follows rather easily from a method of characteristics that the involved "inf-suprema" are nonpositive. It is a somewhat awkward task to see they are actually equal to zero. We shall prove them to be so for a large variety of initial data, including "monotone" (in a sense that we will specify) and especially Riemann data. The result remains a conjecture for general bounded-variation (BV) initial data.

2. Statement of the problem. As mentioned above, we consider the system of conservation laws

$$(1) \qquad u_t + f(u)_x = 0, \qquad u(x,t) \in \mathbb{R}^n, \quad x \in \mathbb{R},\ t > 0,$$

which is assumed to be strictly hyperbolic, the eigenvalues being denoted by $\lambda_1(u), \dots, \lambda_n(u)$ for each state $u \in \Omega \subset \mathbb{R}^n$. This is assumed to be a B. Temple system. It is now classical that this implies the existence of a set $\{w_1, \dots, w_n\}$ of n independent strong Riemann invariants (i.e, for all u, the forms $dw_1(u), \dots, dw_n(u)$ are left eigenvectors of $df(u)$ associated, respectively, with $\lambda_1(u), \dots, \lambda_n(u)$). And this ensures the existence [3], [4] and the uniqueness [1] of a BV entropy solution to the Cauchy problem

$$(1) \qquad u_t + f(u)_x = 0,$$

$$(2) \qquad u(x,0) = u_0(x) \in \Omega, \quad x \in \mathbb{R}$$

provided that $u_0 \in \mathrm{BV}(\mathbb{R})$. Initially (through §3), these results are sufficient for our analysis. We do not need to assume that *each* characteristic field belongs to the B. Temple class. However, this will be used explicitly in Lemma 4.1 which suggests that certain strong interactions between all the fields are involved in formula (5). This point is not completely clear since this assumption is not used in the results of §§4.1 and 4.2.

Henceforth, we suppose that (λ, r) (r being a right eigenvector associated with λ) is a genuinely nonlinear B. Temple field of system (1). (Note that in the absence of genuine nonlinearity, it is impossible to derive any kind of Lax formula.) Since $d\lambda \cdot r$ does not vanish on Ω, we may choose, for instance, r such that

$$\text{(i)} \qquad d\lambda(u) \cdot r(u) \ > \ 0, \quad \forall u \in \Omega.$$

Let w be a strong Riemann invariant associated with this field. By the strict hyperbolicity of system (1), we know that $dw \cdot r$ cannot vanish. Let us assume, possibly changing w into $-w$, that

$$\text{(ii)} \qquad dw(u) \cdot r(u) \ > \ 0, \quad \forall u \in \Omega.$$

By Definition 1, we know that the family of characteristic submanifolds $\{u \in \Omega; w(u) = a\}$ consists of a foliation by hyperplanes of Ω. Thus for all $a \in w(\Omega)$, there exists a hyperplane $\Pi_a \subset \mathbb{R}^n$ such that

$$\{u \in \Omega; w(u) = a\} = \Pi_a \cap \Omega.$$

The one-parameter family of hyperplanes $(\Pi_a)_{a \in w(\Omega)}$ may be described by means of their affine equations

$$\Pi_a = \{u \in \mathbb{R}^n; e_a \cdot u = m_a\},$$

where the e_a's are linear forms and the m_a's are scalar quantities. Clearly,

$$e_a \parallel dw(u) \quad \forall u \in \Pi_a \cap \Omega$$

so that we can choose e_a pointing into the same half-space as $dw(u)$ for $u \in \Pi_a \cap \Omega$. This means we also have

$$\text{(iii)} \qquad e_a \cdot r(u) \ > \ 0 \quad \forall u \in \Pi_a \cap \Omega.$$

It is not difficult to see that $e_a \cdot f(u)$ is also a constant on $\Pi_a \cap \Omega$, which we shall denote by q_a:

$$e_a \cdot f(u) = q_a \quad \forall u \in \Pi_a \cap \Omega.$$

DEFINITION 3. *We refer to a genuinely nonlinear B. Temple field* (λ, r) *with its related objects* w, e_a, m_a, *and* q_a *as a* GNLBT field.

We now assume the entropy solution u of the Cauchy problem (1)–(2) to be piecewise smooth. Thus it can be expressed as the space derivative of v such that

$$v_t + f(v_x) = 0, \tag{3}$$

$$v(x,0) = v_0(x) = \int_0^x u_0(\eta)d\eta. \tag{4}$$

With these assumptions, the conjectured representation formula reads as follows:

$$\inf_y \sup_a \{e_a \cdot v_0(y) - e_a \cdot v(x,t) + (x-y)\, m_a - t\, q_a\} = 0. \tag{5}$$

This formula clearly reduces to a Lax formula in the strictly convex scalar case. Indeed, in that case, there is just one field which is obviously of the B. Temple class and is genuinely nonlinear by the strict convexity. Inequalities (i)–(iii) are satisfied with $\lambda(u) = f'(u)$, $r \equiv 1$, $w(u) = u$, $e_a = 1$, $m_a = a$, and $q_a = f(a)$. Substituting into (5), we obtain

$$\inf_y \sup_a \{v_0(y) - v(x,t) + (x-y)a - tf(a)\} = 0$$

or, equivalently,

$$\inf_y \left\{ v_0(y) + \sup_a [(x-y)a - tf(a)] \right\} = v(x,t),$$

i.e., for $t > 0$,

$$v(x,t) = \inf_y \left\{ v_0(y) + tf^* \left(\frac{x-y}{t} \right) \right\}.$$

This is the well-known Lax formula [2].

Formula (5) was motivated by a result of Serre, who proved [6] that $\mathcal{L}(y,a;x,t) := e_a(v_0(y)) - e_a(v(x,t)) + (x-y)m_a - tq_a$ vanishes when y is given by the bottom of the λ-characteristic passing through (x,t) and a is the constant value $w \circ u(x,t) = w \circ u_0(y)$ of $w \circ u$ along this characteristic. Actually, (5) is meant to clarify the point concerning

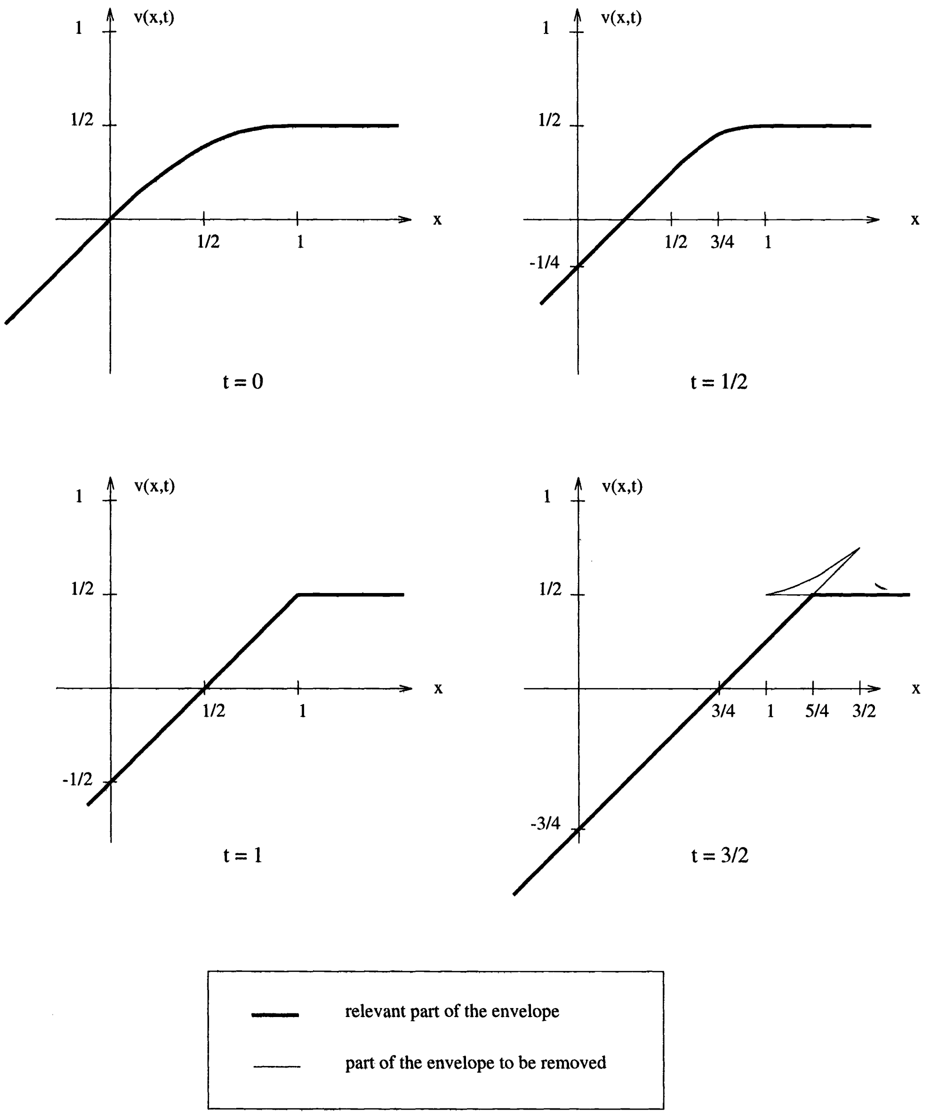


FIG. 1.

shocks in Serre's work. Indeed, his result says that the graph $\mathcal{G} \subset \mathbb{R}^{n+2}$ of v is included in the envelope of the hyperplanes

$$P_y = \{X, T, V; e_{w\circ u_0(y)} \cdot V - X\, m_{w\circ u_0(y)} + T\, q_{w\circ u_0(y)} = e_{w\circ u_0(y)} \cdot v_0(y) - y\, m_{w\circ u_0(y)}\}.$$

If there are shocks, some parts of this envelope should be removed because they should not be attained by $\mathcal{G}$.

Example 1. We can examine the Burgers equation with initial data generating a shock in finite time (at time $t = 1$ on Figure 1). We see that the envelope does contain a loop which is eliminated by the Lax formula.

The next section is devoted to proving the first "half-result":

$$\inf_y \sup_a \{e_a \cdot v_0(y) - e_a \cdot v(x,t) + (x-y)m_a - tq_a\} \leq 0.$$

This will be achieved through a very similar method to that of the scalar case using a sort of "generalized convexity inequality."

3. A first result. Let (λ, r) be a GNLBT field. Then we have the following.

THEOREM 3.1. *If* (i)–(iii) *hold, then for all $a \in w(\Omega)$ and for all $u \in \Omega$,*

$$e_a \cdot f(u) - q_a - \lambda(u)(e_a \cdot u - m_a) \leq 0. \tag{6}$$

Note that in the scalar case, this is merely the classical convexity inequality

$$f(u) - f(a) - f'(u)(u-a) \leq 0,$$

which implies that the graph of f lies above its tangents.

We will use the following lemma in the proof of Theorem 3.1.

LEMMA 3.2. *Let (λ, r) be a GNLBT field such that* (i)–(iii) *hold. Let $s \in I \mapsto u(s)$ be the integral curve of r passing through a given state b at $s = 0$ (with $0 \in I$). Then the real-valued function $s \in I \mapsto \mathcal{F}(u(s); b) := dw(b) \cdot (f(u(s)) - f(b) - \lambda(u(s))(u(s) - b))$ admits a strict global maximum at $s = 0$.*

Proof. Using the identity $df \cdot r \equiv \lambda r$, we immediately compute the first derivative

$$\frac{d}{ds}\mathcal{F}(u(s); b) = -(d\lambda(u(s)) \cdot r(u(s)))\, dw(b) \cdot (u(s) - b)$$

since $du/ds = r(u(s))$. Now by (i), we see that $(d/ds)\mathcal{F}(u(s); b)$ and $dw(b) \cdot (u(s) - b)$ vanish simultaneously at $s = 0$ and take opposite signs elsewhere. Moreover, the integral curve of r passing through b encounters the hyperplane $\Pi_{w(b)}$ only at the point b (corresponding to $s = 0$). Therefore, by (ii), we get that $dw(b) \cdot (u(s) - b) > 0$ for $s > 0$ and $dw(b) \cdot (u(s) - b) < 0$ for $s < 0$. □

Proof of Theorem 3.1. Relation (6) is equivalent to

$$dw(b) \cdot (f(u) - f(b) - \lambda(u)(u - b)) \leq 0 \tag{7}$$

for all states $u, b \in \Omega$. However, $\mathcal{F}(u; b) = dw(b) \cdot (f(u) - f(b) - \lambda(u)(u - b))$ has the same sign for all $b \in \Pi_a \cap \Omega$ so it is sufficient to show (7) for just one $b \in \Pi_a \cap \Omega$ (depending on u). We can choose b to lie on the integral curve of r passing through u. This curve is transverse to the hyperplanes Π_a (since $dw \cdot r \neq 0$) and thus encounters all of them. Then the trick is in considering the opposite point of view, which will make our computations much easier. We take $b \in \Pi_a \cap \Omega$ and examine $\mathcal{F}(u(s); b)$, where $s \mapsto u(s)$ is the integral curve of r passing through b (such that $u(0) = b$). Now Lemma 3.2 and the fact that $\mathcal{F}(u; u) = 0$ imply that $\mathcal{F}(u(s); b) \leq 0$ for all s. This completes the proof. □

Now it is quite easy to prove the following.

THEOREM 3.3. *Let (λ, r) be a GNLBT field such that* (i)–(iii) *hold. If v satisfies equations* (3) *and* (4) *and is such that $u = v_x$ is piecewise smooth and is the entropy solution to equations* (1) *and* (2), *then for all $(x,t) \in \mathbb{R} \times \mathbb{R}^+$,*

$$\inf_y \sup_a \{e_a \cdot v_0(y) - e_a \cdot v(x,t) + (x-y)\, m_a - t\, q_a\} \leq 0. \tag{8}$$

Proof. Let us denote by $\mathcal{L}$ the function

$$\mathcal{L}(y, a; x, t) = e_a \cdot v_0(y) - e_a \cdot v(x, t) + (x - y)m_a - tq_a.$$

This is a Lipschitz-continuous function with respect to x and t. Moreover, $\mathcal{L}(y, a;\ .\,,.)$ is a nonincreasing function of time along any λ-characteristic for every y and a. This is a straightforward consequence of Theorem 3.1 since the relations

$$\mathcal{L}_t(y, a; x, t) = e_a \cdot f(u(x, t)) - q_a, \tag{9}$$

$$\mathcal{L}_x(y, a; x, t) = -\, e_a \cdot u(x, t) + m_a \tag{10}$$

imply that

$$\mathcal{L}_t(y, a; x, t) + \lambda(u(x, t))\mathcal{L}_x(y, a; x, t)$$

$$= e_a \cdot f(u(x, t)) - q_a - \lambda(u(x, t))(e_a \cdot u(x, t) - m_a) \leq 0$$
$$\forall a \in w(\Omega), \quad \forall y \in \mathbb{R}.$$

Serre pointed out in [6] that the Lax entropy inequalities enable us to go *backward* along any λ-characteristic. Therefore, for any point $(x, t) \in \mathbb{R} \times \mathbb{R}^+$, there is at least one point of intersection between the λ-characteristic passing through (x, t) and the space axis, which will be denoted by $(y(x, t), 0)$ and referred to as the "(x, t)-foot." Note that if (x, t) lies on a λ-shock curve, there may be two "feet," which we will then denote by $y_l(x, t)$ on the left and $y_r(x, t)$ on the right. In this case, $y(x, t)$ refers to either one.

DEFINITION 4. *For any point $(x, t) \in \mathbb{R} \times \mathbb{R}^+$, we define* (x, t)-foot *to be a point $(y(x, t), 0)$ of intersection between the λ-characteristic passing through (x, t) and the space axis. If there is a λ-shock passing through (x, t), we denote by $(y_l(x, t), 0)$ (respectively, $(y_r(x, t), 0)$) the intersection point lying on the left (respectively, on the right) of the shock.*

Here we have

$$\mathcal{L}(y(x, t), a; x, t) \leq \mathcal{L}(y(x, t), a; y(x, t), 0) = 0 \quad \forall a \in w(\Omega),$$

which proves (8). □

Remark 1. Let us note that for all $a \neq w \circ u(x, t)$, $\mathcal{L}(y, a; x, t)$ is a *strictly* decreasing function of time along the λ-characteristics while $\mathcal{L}(y, w \circ u(x, t);\, .\,,.)$ is constant along the λ-characteristic passing through (x, t). (We recall that $w \circ u(x, t)$ is also a constant along such a characteristic.) Indeed, the inequality in Theorem 3.1 is strict unless $w(u) = a$ (cf. Lemma 3.2). This implies

$$\sup_a \mathcal{L}(y(x, t), a; x, t) = 0,$$

where the supremum is actually a strict maximum attained at $a = w \circ u(x, t) = w \circ u_0(y(x, t))$.

Therefore, if the conjectured formula is true, it amounts to saying that the "worst" situation appears at the foot $y = y(x, t)$ in the sense that the infimum

$$\inf_y [\sup_a \mathcal{L}(y, a; x, t)]$$

is, in fact, a minimum attained at $y = y(x, t)$.

4. The reverse inequality. In this section, we discuss the situation in which

$$\sup_a \mathcal{L}(y,a;x,t) \geq 0 \quad \forall y \in \mathbb{R}. \tag{11}$$

In the scalar case, inequality (11) is true due to another monotonic property of $\mathcal{L}$ which states that $\mathcal{L}(y,a;y+f'(a)t,t)$ is a nondecreasing function in time. It is not difficult to show that this is a consequence of the "reverse" convexity inequality

$$f(u)-f(a)-f'(a)(u-a) \geq 0.$$

Thus we have

$$\mathcal{L}(y,a;x,t) \geq \mathcal{L}(y,a;y,0) = 0$$

for a such that $f'(a)=(x-y)/t$. (If $(x-y)/t$ does not belong to the range of f', then y does not contribute to the infimum since $\sup_a[((x-y)/t)a - f(a)] = +\infty$.) In fact, $a = f^*((x-y)/t)$, f^* denoting the convex conjugate function of f.

In the case of a B. Temple system, we have a kind of "reverse" inequality to (7).

LEMMA 4.1. *We assume that all of system* (1) *is of the B. Temple class. Let* (λ, r) *be a GNLBT field such that* (i)–(iii) *hold. Let* $s \in I \mapsto u(s)$ *be the integral curve of* r *passing through a given state* b *at* $s=0$ (*with* $0 \in I$). *Then the real-valued function* $s \in I \mapsto \mathcal{G}(u(s);b) := dw(b)\cdot(f(u(s)) - f(b) - \lambda(b)(u(s)-b))$ *admits a strict global minimum at the point* $s=0$. *In other words, whenever the states* u *and* b *lie on the same integral curve of* r, *they satisfy*

$$dw(b)\cdot(f(u)-f(b)-\lambda(b)(u-b)) \geq 0. \tag{12}$$

Proof. Our computations are even easier than in Lemma 3.2. We have

$$\frac{d}{ds}\mathcal{G}(u(s);b) = (\lambda(u(s)) - \lambda(b))dw(b)\cdot r(u(s)),$$

which clearly vanishes at $s=0$. Moreover, from (i), we know that $\lambda(u(s)) > \lambda(b)$ for $s>0$ and $\lambda(u(s)) < \lambda(b)$ for $s<0$. However, integral curves of r are, in fact, straight lines. (This is where we use the property of the whole system belonging to the B. Temple class; see [7].) Therefore, $dw(b)\cdot r(u(s))$ cannot vanish and remains positive since $dw(b)\cdot r(b) > 0$ (ii). □

Unlike (7), which is equivalent to (6), inequality (12) cannot in general be written just in terms of a. There is one particular case in which Lemma 4.1 enables us to prove equation (5); that is when $\lambda(b)$ depends only on $w(b)$ (which, of course, includes the scalar case).

PROPOSITION 4.2. *Assume that all of system* (1) *is of the B. Temple class,* (λ, r) *is a GNLBT field such that* (i)–(iii) *hold, and*

$$d\lambda \wedge dw = 0 \quad \textit{in } \Omega.$$

If v *satisfies equations* (3) *and* (4) *and is such that* $u = v_x$ *is piecewise smooth and is the entropy solution to equations* (1) *and* (2), *then for all* $(x,t) \in \mathbb{R}\times\mathbb{R}^+$,

$$\inf_y \sup_a \{e_a\cdot v_0(y) - e_a\cdot v(x,t) + (x-y)\,m_a - t\,q_a\} = 0.$$

Proof. Indeed, λ is then constant on $\Pi_a \cap \Omega$. Let us denote its value by λ_a. From Lemma 4.1 (with b the intersection of Π_a and the integral curve of r passing through u), we have for all states $u \in \Omega$ and for all values of $a \in w(\Omega)$,

$$e_a \cdot f(u) - q_a - \lambda_a(e_a \cdot u - m_a) \geq 0. \tag{13}$$

This enables us to work out the formula in the same way as in the scalar case.

Let y be any point in $\mathbb{R}$. From (13), it is easy to see that $\mathcal{L}(y, a; y + \lambda_a t, t)$ is a nondecreasing function of time. If $c = (x - y)/t_0$ is in the range of λ_a, let a such that $\lambda_a = c$. We get

$$\mathcal{L}(y, a; x, t_0) \geq \mathcal{L}(y, a; y, 0) = 0.$$

Now, if $c > \sup\limits_a \lambda_a$, let a_M be an upper bound for $w \circ u$. Then $c > \lambda_{a_M}$ and we have

$$\frac{d}{dt}\mathcal{L}(y, a_M; y + ct, t) = e_{a_M} \cdot f(u(y + ct, t)) - q_{a_M} - c(e_{a_M} \cdot u(y + ct, t) - m_{a_M})$$

with $e_{a_M} \cdot u(y + ct, t) - m_{a_M} \leq 0$. Thus $(d/dt)\mathcal{L}(y, a_M; y + ct, t) \geq 0$ and therefore

$$\mathcal{L}(y, a_M; x, t_0) \geq \mathcal{L}(y, a_M; y, 0) = 0.$$

Of course, if $c < \inf_a \lambda_a$, the same holds with a lower bound a_m. Thus we have (11), which with Theorem 3.3 completes the proof. □

Note that this case is very similar to the scalar case since λ-characteristics are straight lines.

In the general case, the eigenvalue λ is not a constant on the hyperplanes Π_a. Lemma 4.1 states that $\mathcal{L}(y, a; X_a(t), t)$ is nondecreasing along curves that are no longer straight lines but are defined as

$$\frac{dX_a}{dt} = \lambda(b_a(\, u(X_a(t), t))),$$

where $b_a(u)$ denotes the intersection of the integral curve of r passing through u (that is, $u + \mathbb{R}r(u)$) with Π_a. These curves are not easy to handle. Unlike the λ-characteristics, they do not have to intercept the x-axis. Indeed, in order to use the Lax entropy inequalities $\lambda(u_r) < \sigma < \lambda(u_l)$, we need some information on b_a, such as

$$\lambda(b_a(u_r)) \leq \lambda(u_r)$$

and

$$\lambda(b_a(u_l)) \geq \lambda(u_l).$$

There is no obvious way to derive such inequalities unless $w(u_r) \leq a \leq w(u_l)$ (which is undoubtedly related to Lemma 4.9 in §4.2). In any case, the existence of such points of intersection would not even solve the problem. For a given a and a possible corresponding point $y_a(x, t)$, the quantity $\mathcal{L}(y, a; y_a(x, t), 0)$ does not have to be nonnegative.

These remarks are meant to point out the kind of difficulties encountered when we try to check whether or not inequality (11) holds for all y (and especially for $y \neq y(x, t)$). Nevertheless, we conjecture this inequality—and thus also the representation formula (5)—to be true. The proof is now given for several particular cases when the initial data are "nice enough" so that it is possible to produce at least one a (depending on $(x,t;y)$) such that $\mathcal{L}(y,a;x,t) \geq 0$.

Before going further, let us show that (11) holds at $t = 0$ for any initial data of bounded variation. Indeed,

$$\mathcal{L}(y,a;x,0) = e_a \cdot (v_0(y) - v_0(x)) + (x-y)m_a,$$

which is equal to zero if $y = x$ (for all $a \in w(\Omega)$) or, for $y \neq x$,

$$\mathcal{L}(y,a;x,0) = (y-x)\left[e_a \cdot \left(\frac{1}{y-x}\int_x^y u_0\right) - m_a\right],$$

which is equal to zero for $\boldsymbol{a} = \boldsymbol{w}\ ((1/(y-x))\int_x^y \boldsymbol{u}_o)$. (The mean value $(1/(y-x))\int_x^y u_0$ still lies in the convex set Ω.)

4.1. Proof of equation (5) for special initial data. For $t > 0$, let us begin with a very simple case.

THEOREM 4.3. *Let* (λ, r) *be a GNLBT field such that* (i)–(iii) *hold. Assume that* $w \circ u_0$ *is* nondecreasing. *If* v *satisfies equations* (3) *and* (4) *and is such that* $u = v_x$ *is piecewise smooth and is the entropy solution to equations* (1) *and* (2), *then for all* $(x,t) \in \mathbb{R} \times \mathbb{R}^+$,

$$\inf_y \sup_a \{e_a \cdot v_0(y) - e_a \cdot v(x,t) + (x-y)\, m_a - t\, q_a\} = 0.$$

Proof. This is a simple case in the sense that there is no λ-shock in the solution with such initial data. In addition, the proof works out almost immediately. Indeed, let $\boldsymbol{a}_{(x,t)} = \boldsymbol{w} \circ \boldsymbol{u}(\boldsymbol{x},\boldsymbol{t}) = \boldsymbol{w} \circ \boldsymbol{u}_o(\boldsymbol{y}(\boldsymbol{x},\boldsymbol{t}))$. Then $\mathcal{L}(y, a_{(x,t)}; \cdot, \cdot)$ is constant along the λ-characteristic passing through (x,t) (see Remark 1) and thus

$$\mathcal{L}(y, a_{(x,t)}; x, t) = \mathcal{L}(y, a_{(x,t)}; y(x,t), 0) \quad \forall y \in \mathbb{R}.$$

Now let us compare $\mathcal{L}(y, a_{(x,t)}; y(x,t), 0)$ to $\mathcal{L}(y, a_{(x,t)}; y, 0) = 0$ for any $y \in \mathbb{R}$. From equation (10), (ii), and (iii), we know that

$$\operatorname{sgn} \mathcal{L}_x(y,a;x,t) = \operatorname{sgn}(a - w \circ u(x,t)). \tag{14}$$

Therefore,

- if $y \leq y(x,t)$, then $w \circ u_0(z) \leq a_{(x,t)}$ for all $z \in [y, y(x,t)]$ and $\mathcal{L}(y, a_{(x,t)}; z, 0)$ is nondecreasing for $z \in [y, y(x,t)]$;
- if $y \geq y(x,t)$, then $w \circ u_0(z) \geq a_{(x,t)}$ for all $z \in [y(x,t), y]$ and $\mathcal{L}(y, a_{(x,t)}; z, 0)$ is nonincreasing for $z \in [y(x,t), y]$.

In both situations, we see that

$$0 = \mathcal{L}(y, a_{(x,t)}; y, 0) \leq \mathcal{L}(y, a_{(x,t)}; y(x,t), 0) = \mathcal{L}(y, a_{(x,t)}; x, t). \qquad \square$$

Let us now take some rather simple initial data which may give rise to a λ-shock. Let $X_0 \in \mathbb{R}$.

THEOREM 4.4. *Let* (λ, r) *be a GNLBT field such that* (i)–(iii) *hold. Assume that* $w \circ u_0$ *reads*

$$w \circ u_0(z) = \begin{cases} w_l & \textit{if } z < X_0, \\ w_r & \textit{if } z > X_0. \end{cases}$$

If v satisfies equations (3) *and* (4) *and is such that $u = v_x$ is piecewise smooth and is the entropy solution to equations* (1) *and* (2), *then for all $(x,t) \in \mathbb{R} \times \mathbb{R}^+$,*

$$\inf_y \sup_a \{e_a \cdot v_0(y) - e_a \cdot v(x,t) + (x-y)\, m_a - t\, q_a\} = 0.$$

Note that this case contains the Riemann problems but is more general in the sense that u_0 is only required to stay in a given hyperplane (Π_{w_l} or Π_{w_r}) on each interval instead of being a constant. However, there is no additional difficulty in the proof.

Proof. If $w_l \leq w_r$ (i.e., a λ-rarefaction wave appears for $t > 0$), then, in fact, Theorem 4.3 holds. Henceforth, we assume that $w_l > w_r$. Then a λ-shock takes place for $t > 0$. However, due to the Lax entropy inequalities, the λ-characteristic passing through (x,t) cannot cross this shock at any time $s \leq t$. For example, let us assume that the λ-characteristic passing through (x,t) lies to the right (of the λ-shock curve initiated at X_0) until time t. Let us split the study of $\sup_a \mathcal{L}(y,a;x,t)$ into two cases depending on whether the point $(y,0)$ lies on the same side as the λ-shock curve or the opposite one.

1. Let us take $y \geq X_0$. Then the proof does not depend on $w \circ u_0(Y)$ for $Y < X_0$. Indeed, let $\boldsymbol{a = w \circ u(x,t) = w \circ u_0(y(x,t)) = w \circ u_0(y) = w_r}$. Then we have

$$0 = \mathcal{L}(y,a;y,0) = \mathcal{L}(y,a;y(x,t),0) = \mathcal{L}(y,a;x,t),$$

which proves that $\sup_a \mathcal{L}(y,a;x,t) \geq 0$.

2. Now let us take $y < X_0$. We will see that $\boldsymbol{a = w \circ u_0(y)}$ still gives the desired estimate. We denote by $X(t)$ the position of the λ-shock at time t. Then going backwards along the λ-characteristics, we see that $w \circ u(\xi,t) = w_r < w_l = w \circ u_0(y) = a$ for all $\xi \in]X(t), x]$. Therefore, using (14), we have

$$\mathcal{L}(y,a;x,t) \geq \mathcal{L}(y,a;X(t),t).$$

Now we consider $y_l(X(t),t)$ the "$(X(t),t)$-left-foot" as in Definition 4. Then, applying case 1 to $(X(t),t)$ instead of (x,t) and w_l instead of w_r, we get

$$0 = \mathcal{L}(y,a;y,0) = \mathcal{L}(y,a;y_l(X(t),t),0) = \mathcal{L}(y,a;X(t),t)$$

since $a = w_l = w \circ u(X(t)-0,t) = w \circ u_0(y_l(X(t),t))$. □

Remark 2. In case 1 of the proof, we may weaken the assumption on $w \circ u_0$. Indeed, let us assume that $w \circ u_0(z)$ is only nondecreasing for any $z > X_0$. Let $\boldsymbol{a = w \circ u(x,t) = w \circ u_0(y(x,t))}$ (which is possibly no longer equal to $w \circ u_0(y)$). Then we have (see the proof of Theorem 4.3)

$$0 = \mathcal{L}(y,a;y,0) \leq \mathcal{L}(y,a;y(x,t),0) = \mathcal{L}(y,a;x,t),$$

which still proves $\sup_a \mathcal{L}(y,a;x,t) \geq 0$.

Remark 3. We have seen in case 2 that the proof does not always work with the "simple" value $a = w \circ u(x,t)$. This is not surprising. For the scalar case, the "working value" is $a = f^*((x-y)/t)$. For systems, we do not have a tool similar to f^*, which explains the difficulties related to the derivation of more general results.

Remark 2 prompts us to try the following generalization of Theorems 4.3 and 4.4. Let us take $X_0 \in \mathbb{R}$.

THEOREM 4.5. *Let* (λ, r) *be a GNLBT field such that* (i)–(iii) *hold. Assume that* $w \circ u_0$ *is a nondecreasing function on each of the intervals* $]-\infty, X_0)$ *and* $(X_0, +\infty[$*. If* v *satisfies equations* (3) *and* (4) *and is such that* $u = v_x$ *is piecewise smooth and is the entropy solution to equations* (1) *and* (2)*, then for all* $(x,t) \in \mathbb{R} \times \mathbb{R}^+$,

$$\inf_y \sup_a \{e_a \cdot v_0(y) - e_a \cdot v(x,t) + (x-y)m_a - tq_a\} = 0.$$

We shall only consider the case when

$$w \circ u_0(X_0 - 0) > w \circ u_0(X_0 + 0).$$

(Otherwise Theorem 4.3 holds and gives the result.) This generates a single λ-shock wave. More precisely, we need the following lemma.

LEMMA 4.6. *Let* u_0 *be of bounded variation such that* $w \circ u_0$ *is a nondecreasing function on* $]-\infty, X_0)$ *and* $(X_0, +\infty[$ *and* $w \circ u_0(X_0 - 0) > w \circ u_0(X_0 + 0)$*. Then the entropy solution to equations* (1) *and* (2) *consists of a* λ*-shock wave* $t \mapsto X(t)$ *and* $w \circ u(x,t)$ *is a nondecreasing function of* x *for* $x < X(t)$ *and for* $x > X(t)$*. Moreover, let*

$$w_l(t) = w \circ u(X(t) - 0, t)$$

and

$$w_r(t) = w \circ u(X(t) + 0, t).$$

Then $w_l(t)$ (*respectively,* $w_r(t)$) *is a continuous nonincreasing* (*respectively, nondecreasing*) *function of* $t > 0$.

Proof. The monotonic properties are derived by going backward along the λ-characteristics and using the properties of $w \circ u_0$. If $w \circ u_0$ has some points of discontinuity apart from X_0, then they generate rarefaction waves; hence we have the continuity of $w_{l,r}(t)$. □

In particular, Lemma 4.6 implies the existence of $w_{l,r}^\infty = \lim_{t \to +\infty} w_{l,r}(t)$ such that

$$w_l(0) \geq w_l^\infty \geq w_r^\infty \geq w_r(0). \tag{15}$$

Proof of Theorem 4.5. If $(y, 0)$ lies on the same side of the λ-shock as (x,t), the proof of equation (5) is the same as the proof of Theorem 4.3 (see Remark 2).

The remaining problem is when (x,t) and $(y,0)$ lie on opposite sides. Let us assume, for instance, that (x,t) lies on the left, i.e., $x < X(t)$ and thus $y(x,t) < X_0$. Let us take $y > X_0$. We shall discuss different cases concerning the location of $y(x,t)$. Let us define

$$\begin{aligned} Y_l^\infty &= \sup\{Y \leq X_0; w \circ u_0(Y) \leq w_l^\infty\}, \\ Y_r^\infty &= \sup\{Y \leq X_0; w \circ u_0(Y) \leq w_r^\infty\}, \\ Y_r^0 &= \sup\{Y \leq X_0; w \circ u_0(Y) \leq w_r(0)\}. \end{aligned}$$

Clearly, from (15), these points (possibly equal to $-\infty$) are in the following order:

$$Y_r^0 \leq Y_r^\infty \leq Y_l^\infty \leq X_0.$$

1. If $y(x,t) \leq Y_r^0$, then $w \circ u(x,t) = w \circ u_0(y(x,t)) \leq w_r(0)$. In this case, the proof is the same as if $w \circ u_0$ were nondecreasing on the whole interval $[y(x,t), y]$ (whereas there is a "bump" between Y_r^0 and X_0). Indeed, let $\boldsymbol{a = w \circ u(x, t)}$. Then $w \circ u_0(z) \geq a$ for any $z \in [y(x,t), y]$. Therefore, due to (14), $\mathcal{L}(y, a; z, 0)$ is a nonincreasing function of z for any $z \in [y(x,t), y]$, which gives

$$0 = \mathcal{L}(y, a; y, 0) \leq \mathcal{L}(y, a; y(x,t), 0) = \mathcal{L}(y, a; x, t).$$

2. If $Y_r^0 < y(x,t) \leq Y_r^\infty$, then $w_r(0) \leq w \circ u(x,t) \leq w_r^\infty$. Now if $w \circ u(x,t) < w_r^\infty$, there exists a $T \geq 0$ such that $w \circ u(x,t) = w_r(T)$. Let $\boldsymbol{a = w \circ u(x, t)}$. Let us consider $y_r(X(T), T)$ to be the $(X(T), T)$-right-foot. There are two possibilities:

- If the shock has already canceled at time T, then $w_r(T) = w_l(T)$ and we have

$$0 = \mathcal{L}(y, a; y, 0) \leq \mathcal{L}(y, a; y_r(X(T), T), 0); = \mathcal{L}(y, a; X(T), T) = \mathcal{L}(y, a; x, t).$$

- If we have $w_r(T) < w_l(T)$, then the λ-characteristic passing through (x,t) cannot have encountered the λ-shock at time T. Indeed, from Lemma 4.6, we have $w \circ u(x,t) = w_r(T) < w_l(T) \leq w_l(\tau)$ for all $\tau \leq T$. Therefore, let χ be the position of this characteristic at time T.

On one hand, $a = w \circ u(x,t)$ forces $\mathcal{L}(y, a; ., .)$ to be a constant along the λ-characteristic passing through (x,t) while $a = w_r(T)$ forces $\mathcal{L}(y, a; ., .)$ to be a constant along the λ-characteristic(s) passing through $(X(T), T)$. On the other hand, since $w \circ u_0(z) \geq w \circ u_0(y(x,t))$ for $z \in [y(x,t), X_0]$ and due to (14), $\mathcal{L}(y, a; ., T)$ is nonincreasing on $[\chi, X(T)]$. Thus we can write

$$\begin{aligned} 0 = \mathcal{L}(y, a; y, 0) &\leq \mathcal{L}(y, a; y_r(X(T), T), 0) = \mathcal{L}(y, a; X(T), T) \\ &\leq \mathcal{L}(y, a; \chi, T) = \mathcal{L}(y, a; x, t). \end{aligned}$$

The limit case $w \circ u(x,t) = w_r^\infty$ follows from the continuity of $\mathcal{L}$, which gives

$$0 \leq \mathcal{L}(y, w_r^\infty; x, t).$$

3. If $Y_l^\infty \leq y(x,t) < X_0$, then $w_l^\infty \leq w \circ u(x,t) \leq w_l(0) = w \circ u_0(X_0 - 0)$. Now, if $w \circ u(x,t) > w_l^\infty$, there exists a $T \geq 0$ such that $w \circ u(x,t) = w_l(T)$. Then the λ-characteristic passing through (x,t) encounters the shock curve at time T and, of course, $t \leq T$. Let $\boldsymbol{a = w_r(T)}$. Let us consider $y_r(X(T), T)$ to be the $(X(T), T)$-right-foot. Since $\mathcal{L}(y, a; ., .)$ is a nonincreasing function of time along λ-characteristics and is constant iff a is the constant value of $w \circ u$ along such characteristics, we have

$$\begin{aligned} 0 = \mathcal{L}(y, a; y, 0) &\leq \mathcal{L}(y, a; y_r(X(T), T), 0) = \mathcal{L}(y, a; X(T), T) \\ &\leq \mathcal{L}(y, a; x, t). \end{aligned}$$

The limit case $w \circ u(x,t) = w_l^\infty$ follows from the continuity of $\mathcal{L}$, which again gives

$$0 \leq \mathcal{L}(y, w_r^\infty; x, t).$$

4. If $Y_r^\infty < y(x,t) < Y_l^\infty$, then $w_r^\infty \leq w \circ u(x,t) \leq w_l^\infty$. The λ-characteristics issued from any point in $[Y_r^\infty, Y_l^\infty]$ are necessarily defined for any time. Let $\xi_{l,r}$ be the location at time t of the λ-characteristic issued from $Y_{l,r}^\infty$. We have $\xi_r < x < \xi_l$ and $w \circ u(., t) \geq w_r^\infty$ on $]x, \xi_l]$. Therefore, let $\boldsymbol{a = w_r^\infty}$. The function $\mathcal{L}(y, a; ., t)$ is nonincreasing on $[x, \xi_l]$ and we have

$$0 \leq \mathcal{L}(y, a; \xi_l, t) \leq \mathcal{L}(y, a; x, t).$$

This ends the proof of Theorem 4.5. $\square$

Before going further, it may be interesting to notice that there is a kind of paradox in this proof, which actually throws light on the asymptotic behavior of the solution. This is summed up in the following statement.

COROLLARY 4.7. *Let u_0 be of bounded variation such that $w \circ u_0$ is a nondecreasing function on $]-\infty, X_0)$ and $(X_0, +\infty[$ and $w \circ u_0(X_0 - 0) > w \circ u_0(X_0 + 0)$. We denote by $t \mapsto X(t)$ the generated λ-shock wave in the entropy solution to equations* (1) *and* (2) *and*

$$w_l(t) = w \circ u(X(t) - 0, t),$$

$$w_r(t) = w \circ u(X(t) + 0, t).$$

If $w \circ u_0(-\infty) < w \circ u_0(+\infty)$, then

$$\lim_{t \to +\infty} w_l(t) = \lim_{t \to +\infty} w_r(t).$$

If $w \circ u_0(-\infty) \geq w \circ u_0(+\infty)$, then

$$\lim_{t \to +\infty} w_l(t) = w \circ u_0(-\infty) \quad \text{and} \quad \lim_{t \to +\infty} w_r(t) = w \circ u_0(+\infty).$$

Proof. We are going to show that case 4 in the proof of Theorem 4.5 is void. In other words, there are no points (x, t) such that

$$w_r^\infty < w \circ u(x, t) < w_l^\infty. \tag{16}$$

Indeed, if (x, t) were such a point and were lying on the left of the shock, we would have from case 4 that

$$\mathcal{L}(y, w_r^\infty; x, t) \geq 0.$$

Since $w \circ u(x,t) < w_l^\infty$, the λ-characteristic passing through (x, t) could not encounter the shock in finite time, and because $w \circ u(x, t) > w_r^\infty$, the function $\mathcal{L}(y, w_r^\infty; ., .)$ would be strictly decreasing along this characteristic. Thus this function would have a finite nonnegative limit as $t \to +\infty$ and its derivative would tend to zero. However, this derivative is

$$e_{w_r^\infty} \cdot f(u(x, t)) - q_{w_r^\infty} - \lambda(u(x, t))(e_{w_r^\infty} \cdot u(x, t) - m_{w_r^\infty})$$

and cannot tend to zero unless $w \circ u(x, t)$ tends to w_r^∞ (see Remark 2). This is impossible since $w \circ u(x, t)$ is a constant along the λ-characteristic and is not equal to w_r^∞ by assumption.

If (x, t) were a point lying on the right of the shock, we could show in exactly the same way (taking $(y, 0)$ on the left) that (16) cannot hold.

Thus we have either $]w_r^\infty, w_l^\infty[= \emptyset$, which means

$$w_r^\infty = w_l^\infty, \tag{17}$$

or

$$w_l^\infty = w \circ u_0(-\infty) \quad \text{and} \quad w_r^\infty = w \circ u_0(+\infty). \tag{18}$$

If $w \circ u_0(-\infty) < w \circ u_0(+\infty)$, equation (18) is impossible since we always have $w_r^\infty \leq w_l^\infty$, so equation (17) must hold. □

Therefore, the asymptotic behavior of the solution consists of either the cancellation of the shock or its adaptation to the profile given by $w \circ u_0(-\infty)$ and $w \circ u_0(+\infty)$. This is a rather nontrivial consequence of our analysis.

Using Theorems 4.3, 4.4, and 4.5, we can easily obtain the following results concerning "piecewise-constant" initial data (which may be interesting in relation to finite difference schemes).

THEOREM 4.8. *Let (λ, r) be a GNLBT field such that* (i)–(iii) *hold. Assume that $w \circ u_0$ is* piecewise constant *and is such that*

1. *$w \circ u_0$ is nondecreasing,*
2. *$w \circ u_0$ is nonincreasing, or*
3. *the "mesh" contains at most three intervals.*

If v satisfies equations (3) *and* (4) *and is such that $u = v_x$ is piecewise smooth and is the entropy solution to equations* (1) *and* (2), *then for all $(x,t) \in \mathbb{R} \times \mathbb{R}^+$,*

$$\inf_y \sup_a \{e_a \cdot v_0(y) - e_a \cdot v(x,t) + (x-y)\, m_a - t\, q_a\} = 0.$$

Proof. Case 1 is contained in Theorem 4.3. Case 2 can be proved similarly to Theorem 4.4 with $a = w \circ u_0(y)$. Actually, it will follow from the next section that $a = w \circ u_0(y)$ gives the desired estimate for *any* nonincreasing initial data. As for case 3, the initial data takes at most three values. In any case, we can apply either cases 1 or 2 or Theorem 4.5. Unfortunately, with more than three intervals, our method would become too complicated. □

4.2. Attempt to extend the proof to any initial data. One attempt to extend the proof to any kind of initial data proceeds as follows. Drawing inspiration from Theorems 4.3 and 4.4, we may consider the following situation. Let $(x,t) \in \mathbb{R} \times \mathbb{R}^{+*}$ and $(y(x,t),0)$ be its foot. Let $y < y(x,t)$. Consider a point $z(y;x,t) \in [y, y(x,t)]$ such that $w \circ u_0(z) \leq w \circ u_0(z(y;x,t))$ for all $z \in [y, y(x,t)]$.

Assume that the λ-characteristic issued from $z(y;x,t)$ attains time t. We denote by $Z(y;x,t)$ its location at that time. Let $\boldsymbol{a = w \circ u_0(z(y;x,t))}$. By the definition of a, we have that $w \circ u_0(z) \leq a$, for all $z \in [y, z(y;x,t)]$ and also that

$$w \circ u(Z,t) \leq w \circ u(Z(y;x,t),t) = w \circ u_0(z(y;x,t)) = a$$

for all $Z \in [Z(y;x,t), x]$. Therefore, we get from equation (14) that

$$\mathcal{L}(y,a;x,t) \geq \mathcal{L}(y,a;Z(y;x,t),t) = \mathcal{L}(y,a;z(y;x,t),0) \geq \mathcal{L}(y,a;y,0) = 0.$$

In the case where the λ-characteristic issued from $z(y;x,t)$ encounters a λ-shock *before* time t, the preceding argument fails in general. However, it does work in the following situation. If $w \circ u_0$ is nonincreasing, then the proof is based on an additional monotonic property of $\mathcal{L}$ which we state in the following lemma.

LEMMA 4.9. *Let (λ, r) be a GNLBT field such that* (i)–(iii) *hold. Let $t \mapsto X(t)$ be a λ-shock wave. Moreover, let us denote by*

$$w \circ u(X(t)-0,t) = w_l(t), \qquad w \circ u(X(t)+0,t) = w_r(t).$$

Then $t \mapsto \mathcal{L}(y,a;X(t),t)$ is a nondecreasing function of t on any interval $[t_0,t_1]$ provided that

$$\sup_{[t_0,t_1]} \boldsymbol{w_r} \leq \boldsymbol{a} \leq \inf_{[t_0,t_1]} \boldsymbol{w_l}.$$

Proof. We look at the first derivative of the Lipschitz-continuous function $t \mapsto \mathcal{L}(y, a; X(t), t)$. Equations (9) and (10) imply that

$$\begin{aligned}\frac{d}{dt}\mathcal{L}(y, a; X(t), t) &= \mathcal{L}_t(y, a; X(t), t) + \sigma(t)\mathcal{L}_x(y, a; X(t), t)\\ &= e_a \cdot f(u_{l,r}(t)) - q_a - \sigma(t)(e_a \cdot u_{l,r}(t) - m_a).\end{aligned}$$

(Both values give the same result owing to the Rankine–Hugoniot condition.) Now we appeal to the classical convex entropy of (1),

$$u \mapsto |e_a \cdot u - m_a|,$$

whose entropy flux is $u \mapsto \operatorname{sgn}(e_a \cdot u - m_a)(e_a \cdot f(u) - q_a)$. The associated entropy inequality along the shock curve reads

$$\sigma(|e_a \cdot u_r - m_a| - |e_a \cdot u_l - m_a|)$$

$$\geq \operatorname{sgn}(e_a \cdot u_r - m_a)(e_a \cdot f(u_r) - q_a) - \operatorname{sgn}(e_a \cdot u_l - m_a)(e_a \cdot f(u_l) - q_a).$$

Note that due to the Rankine–Hugoniot condition, this inequality would be trivial if we had $(e_a \cdot u_r - m_a)(e_a \cdot u_l - m_a) > 0$. However, the assumption $(w(u_r) \leq a \leq w(u_l))$, (ii), and (iii) imply that

$$e_a \cdot u_r - m_a \leq 0 \leq e_a \cdot u_l - m_a.$$

These quantities cannot vanish simultaneously. (Otherwise, there would be no shock at all.) Therefore the entropy inequality together with the Rankine–Hugoniot condition imply that

$$-2\sigma(e_a \cdot u_l - m_a) \geq -2(e_a \cdot f(u_l) - q_a)$$

or, symmetrically,

$$-2\sigma(e_a \cdot u_r - m_a) \geq -2(e_a \cdot f(u_r) - q_a).$$

This proves that $(d/dt)\mathcal{L}(y, a; X(t), t)$ is nonnegative. □

THEOREM 4.10. *Let* (λ, r) *be a GNLBT field such that* (i)–(iii) *hold. Assume that* $w \circ u_0$ *is nonincreasing. If* v *satisfies equations* (3) *and* (4) *and is such that* $u = v_x$ *is piecewise smooth and is the entropy solution to equations* (1) *and* (2), *then for all* $(x, t) \in \mathbb{R} \times \mathbb{R}^+$,

$$\inf_y \sup_a \{e_a \cdot v_0(y) - e_a \cdot v(x, t) + (x - y)\, m_a - t\, q_a\} = 0.$$

Proof. Let $(x, t) \in \mathbb{R} \times \mathbb{R}^{+*}$ and $(y(x, t), 0)$ be its foot. Let, for example, $y < y(x, t)$. We again consider a point $z(y; x, t) \in [y, y(x, t)]$ such that $w \circ u_0(z) \leq w \circ u_0(z(y; x, t))$ for all $z \in [y, y(x, t)]$. The assumption on $w \circ u_0$ enables us to take $z(y; x, t) = y$. If the λ-characteristic issued from $z(y; x, t) = y$ attains time t, we can stick to the preceding argument. If the λ-characteristic issued from $z(y; x, t) = y$ encounters a λ-shock at some time $t_0 < t$, we denote by $s \mapsto X(s)$ this λ-shock curve and

$$w \circ u(X(t) - 0, t) = w_l(t), \qquad w \circ u(X(t) + 0, t) = w_r(t).$$

By definition, we have either $w \circ u_0(y) = w_l(t_0)$ or $w \circ u_0(y) = w_r(t_0)$. Moreover, by a lemma similar to Lemma 4.6, we have that w_l is a nondecreasing function of time whereas w_r is nonincreasing. Let $\boldsymbol{a} = \boldsymbol{w} \circ \boldsymbol{u}_0(\boldsymbol{y})$. In any case, we get

$$\sup_{[t_0,t]} w_r \leq a \leq \inf_{[t_0,t]} w_l.$$

Applying Lemma 4.9 and equation (14), we get

$$\mathcal{L}(y,a;x,t) \geq \mathcal{L}(y,a;X(t),t) \geq \mathcal{L}(y,a;X(t_0),t_0) = \mathcal{L}(y,a;y,0) = 0. \qquad \square$$

Eventually, we will have proved the formula in both monotonic cases for $w \circ u_0$. That is one of the reasons why we make the following conjecture.

CONJECTURE 1. *Let* (λ, r) *be a GNLBT field such that* (i)–(iii) *hold. Let* u_0 *belong to* BV($\mathbb{R}$). *If* v *satisfies equations* (3) *and* (4) *and is such that* $u = v_x$ *is piecewise smooth and is the entropy solution to equations* (1) *and* (2), *then for all* $(x,t) \in \mathbb{R} \times \mathbb{R}^+$,

$$\inf_y \sup_a \{e_a \cdot v_0(y) - e_a \cdot v(x,t) + (x-y)m_a - tq_a\} = 0. \tag{5}$$

Remark 4. We can prove that $\sup_a \mathcal{L}(y,a;x,t) \geq 0$ for points (x,t) lying outside a cone with vertex y. Indeed, let $a_{m,M}$ be such that

$$a_m \leq w \circ u_0 \leq a_M.$$

It is sufficient to prove that

$$\sup_{a \in [a_m,a_M]} \mathcal{L}(y,a;x,t) \geq 0.$$

Since $\mathcal{L}$ is continuous with respect to a (in fact, it is at least $\mathcal{C}^1$) and is Lipschitz continuous with respect to $(y;x,t)$, the upper envelope

$$L_0(y;x,t) = \sup_{a \in [a_m,a_M]} \mathcal{L}(y,a;x,t) = \max_{a \in [a_m,a_M]} \mathcal{L}(y,a;x,t)$$

is Lipschitz continuous. We shall easily prove that there exists some nonnegative constant C such that $L_0(y;x,t) \geq 0$ if $|(x-y)/t| \geq C$.

Indeed, suppose that we have some compact and convex *invariant domain* $K \subset \Omega$ that contains u_0. For each $a \in [a_m, a_M]$, let

$$\delta_a(u) = \begin{cases} \dfrac{e_a \cdot f(u) - q_a}{e_a \cdot u - m_a} & \text{if } u \notin \Pi_a \cap \Omega, \\ \lambda(u) & \text{if } u \in \Pi_a \cap \Omega. \end{cases}$$

This gives a continuous function of the two arguments a and u. Then let $C = \min_{[a_m,a_M]\times K} |\delta_a(u)|$. We have—among other things—that

$$\mathcal{L}_t(y,a;x,t) = e_a \cdot f(u(x,t)) - q_a \geq -C|\, m_a - e_a \cdot u(x,t)\,| = -C|\,\mathcal{L}_x(y,a;x,t)\,|.$$

In particular, this gives, for $x < y$,

$$\mathcal{L}_t(y,a_m;x,t) \geq +C\mathcal{L}_x(y,a_m;x,t)$$

and, for $x > y$,

$$\mathcal{L}_t(y, a_M; x, t) \geq -C\mathcal{L}_x(y, a_M; x, t).$$

After integrating these inequalities along the lines with "slopes" $-C$ and $+C$, respectively, we obtain that $\mathcal{L}$ is nondecreasing along those lines. Since equation (14) implies that

$$\mathcal{L}(y, a_m; Y, 0) \geq 0, \quad Y < y,$$

and that

$$\mathcal{L}(y, a_M; Y, 0) \geq 0, \quad Y > y,$$

we must have that $L_0(y; x, t) \geq 0$ for all (x, t) such that $|(x - y)/t| \geq C$.

5. Conclusion. Formula (5) has been proved for a large variety of initial data, including Riemann data and more generally monotone initial data (in the coordinates of Riemann invariants). For general initial data of bounded variation, its proof should rely on

$$L_0(y; x, t) \geq 0 \quad \forall y \in \mathbb{R} \tag{11}$$

for all $(x, t) \in \mathbb{R} \times \mathbb{R}^+$. It was noted that inequality (11) holds at $t = 0$ and also for $(y; x, t)$ such that $|(x - y)/t| \geq C > 0$. However, there is still a great deal of analysis to do in order to remove all the restrictions.

Acknowledgments. This study was suggested to me by D. Serre. I would especially like to thank him for his constant encouragement. A. Heibig also took an interest in my work for which I thank him sincerely.

REFERENCES

[1] A. Heibig, *Existence and uniqueness of solutions for some hyperbolic systems of conservation laws*, Arch. Rational Mech. Anal., 126 (1994), pp. 79–101.

[2] P. D. Lax, *Hyperbolic systems of conservation laws* I, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[3] R. Leveque and B. Temple, *Stability of Godunov's method for a class of* 2×2 *systems of conservation laws*, Trans. Amer. Math. Soc., 288 (1985), pp. 115–123.

[4] D. Serre, *Solutions à variation bornée pour certains systèmes hyperboliques de lois de conservation*, J. Differential Equations, 68 (1987), pp. 137–169.

[5] ———, *Richness and the classification of quasilinear hyperbolic systems*, IMA Vol. Math. Appl., 29 (1991), pp. 315–333.

[6] ———, *Temple's fields and integrability of hyperbolic systems of conservation laws*, in Proc. International Conference on Nonlinear PDEs, Guangchang Dong and Fanghua Lin, eds., International Academic Publishers, 1993, pp. 233–251.

[7] B. Temple, *Systems of conservation laws with invariant submanifolds*, Trans. Amer. Math. Soc., 280 (1983), pp. 781–795.

SIAM J. MATH. ANAL.
Vol. 27, No. 6, pp. 1520–1543, November 1996

002

CONVERGENCE OF THE HOMOGENIZATION PROCESS FOR A DOUBLE-POROSITY MODEL OF IMMISCIBLE TWO-PHASE FLOW*

ALAIN BOURGEAT†, STEPHAN LUCKHAUS‡, AND ANDRO MIKELIĆ§

Abstract. In this paper, we justify by periodic homogenization the double-porosity model for immiscible incompressible two-phase flow. The volume fraction of the fissured part and the nonfissured part are kept positive constants and of the same order. The scaling is such that, in the final homogenized equations, the less permeable part of the matrix contributes as a nonlinear memory term. To prove the convergence of the total velocity and of the "reduced" pressure, we use the two-scale convergence since it seems to be appropriate for the problem, even though it would be possible to work with periodic modulation. However, in the final step, the degenerate ellipticity prevents the use of the two-scale convergence method and leads us to use periodic modulation.

Key words. flow in porous medium, double porosity, fractured reservoir, homogenization, dilation

AMS subject classifications. 35B27, 35B45, 35K55, 35K65, 76S05

1. Introduction.

Naturally fractured reservoirs may contain many fractures that permeate different regions of the reservoir and are characterized by the existence of a system of high-conductivity fissures together with a large number of matrix blocks containing most of the oil. The fractures serve as highly conductive flow paths for the reservoir's fluid, increasing the reservoir's effective permeability significantly over the permeability corresponding only to the rock matrix. The reservoir mechanism of fractured systems is significantly different from that of a so-called single-porosity system; see, for instance, [15, 17, 18, 23, 26]. To describe the flow of the fluid in such a fractured reservoir, several authors in the engineering literature [8, 16, 19, 31] showed that if there are many well-connected fractures, the network of fractures behaves as an equivalent porous medium, described by the so-called "dual-porosity" model.

In Barenblatt's dual-porosity model of the fractures, the width is considerably greater than the characteristic dimensions of the pores and the permeability K^* of the fissure system considerably exceeds the permeability k of the individual blocks of porous media. At the same time, the fissures occupy a smaller volume than the pores, so the ratio of the volume of the fissures to the total volume is smaller than the porosity of any individual block of porous media.

To obtain the dual-porosity model, the fracture system's local properties are averaged over a volume containing both the fractures and a matrix. The so-called dual-porosity model for a porous medium consists of an equivalent coarse-grained porous medium in which the fissures play the role of "pores" and the blocks of porous media play the role of "grains."

While no flow is allowed between blocks, only matrix–fractures flow is possible,

*Received by the editors November 2, 1994; accepted for publication (in revised form) July 18, 1995.

†Equipe d'Analyse Numérique, Université de Saint-Etienne, 23 Rue Dr. P. Michelon, 42023 Saint-Etienne cedex 2, France (bourgeat@anumsun1.univ-st-etienne.fr).

‡Institut für Angewardte Mathematik, Universität Bonn, Wegeler Straße 6, D-53115 Bonn, Germany.

§Laboratoire d'Analyse Numérique, Université Lyon 1, Bâtiment 101, 43 Boulevard du 11 Novembre, 69622 Villeurbanne cedex, France (andro@lan1.univ-lyon1.fr).

and the porous-rock matrix system plays the role of a global source term macroscopically distributed over the entire equivalent coarse-grained porous medium.

Since flow in the fractures is much more rapid than inside the matrix, the fluid does not flow directly from one matrix block to another. Rather it first flows into the fractures system and then can pass into a block or remain in the fractures.

Denote by ε the ratio between the size of one block of porous media Ω_m^ε to the size of the whole domain of calculation Ω; then the characteristic time scale for any parabolic evolution in one block Ω_m^ε will be of order ε^{-2}.

In our case, the parabolic evolution is driven by the system of continuity (2.1)–(2.6) below. On the other hand, to have a permeability ratio of order ε^2 between the system of equations in the blocks of porous media and the system of equations in the fissures means that the ratio of the characteristic time for the rescaled flow (by $y = x/\varepsilon$) in a single block and the characteristic time for the flow through the entire system of fractures is of order ε^{-2}. Roughly speaking, we may say that a time ratio of ε^2 between the fractures and the porous block will give a time scaling allowing at the global level (i.e., $\varepsilon \to 0$) the matrix–fracture interaction phenomena described before and will lead to the dual-porosity model. At a time $t \ll 1$, a large fraction of the reserves is extracted from the fractures; then at time $t \sim 0(1)$, the exchange between porous blocks and fissures as described in [25] begins. This effect has already been observed in the case of a diffusion equation coupled with Darcy flow in the paper of Vogt [30]. It should also be noticed that this ε^2 time scaling is done in the engineering literature, as, for instance, in [27, 28], but is motivated by introducing a geometrical factor of transmissibility. If one takes the ratio of the two permeabilities of order one, then by the usual theory of homogenization the limit model will be, as, for instance, in [11, 12], a single-porosity model. If the ratio is smaller than of order ε^2, then there is no contribution from the blocks to the global continuity system of equations in the limit model, which then corresponds to the homogenization of only the system of fissures.

The precise physical assumptions made before averaging for the system of porous blocks and fissures are as follows. The medium is of "dual-porosity" type; i.e., the two parts of this medium, the fissures and the blocks, both behave like porous media and obey the generalized Darcy law. They differ only by their porosity and absolute permeability. Moreover, the two phases—the wetting and the nonwetting—are assumed to be incompressible and in capillary equilibrium at the pore level. This last assumption means that the faster time scale (or order ε^2 as discussed before) is sufficiently large to be bigger than the required time to establish pore equilibrium in both parts of the porous medium; moreover, we assume that there are no capillary hysteresis effects, which means this capillary equilibrium is unique. These last two assumptions lead in both parts of the medium to capillary pressure and relative permeability curves depending only on saturation and space and, finally, to the two-phase immiscible-flow parabolic-elliptic system of equations with a degeneracy. In this we follow the existing engineering literature.

It should be noticed that in some experimental situations as described in [9] or [10], the physical assumptions above are not valid; some other models of fissured porous media must be used. Some of these models in which the capillary pressure in the fissured part is neglected lead to a system with a purely convective equation instead of the degenerate one. This type of purely convective equation is outside the scope of this paper and of the homogenization methods presented herein.

The main objective of this paper is to derive rigorously (from the mathematical point of view) the dual-porosity model for incompressible two-phase flow. Rigorous mathematical proof of this dual-porosity model has been obtained before only for the single-phase flow in [6], i.e., when the equations are linear. For the case of two-phase flow, where the leading equations are much more complicated (a nonlinear system of coupled equations with parabolic degeneracy), the authors of [6] were able only to derive this model by formal asymptotic expansion in [7].

To prove our result, we use the two-scale convergence method defined in [3] and [24] to prove convergence of the total velocity and of the "reduced" pressure, but the difficulty added by the degenerate ellipticity prevents the use of two-scale convergence of the saturation and leads to the use of periodic modulation as defined in [6] or [30].

For the sake of simplicity, as usual in the engineering literature (see, for instance [25, 28, 31]), we assume periodic distribution of the cells; each cell contains only one matrix. However, the two-scale method could also be used with randomly distributed cells as in [3], and although the periodic modulation is used herein under periodic assumptions, we think that it could certainly be extended to some type of randomly distributed media.

The remainder of the paper is as follows.

In the next section, we present the equations that describe the microscopic nature of the two-phase flow in a naturally fractured reservoir. These equations are spacially scaled by ε, the ratio between the size of the blocks and the size of the domain Ω, and are time scaled as explained above by ε^2. After extension of the solution and the pressure from the fissured part to the whole domain Ω, we obtain from the a priori estimates in §3 the convergence to the homogenized macroscopic model in §4.

2. The microscopic model. The system of small fractures surrounding the block of porous medium is considered itself as a porous medium with dimensionless absolute permeability K^* and porosity ϕ^*. For this medium, the relative permeability curves are denoted $K_{ri}(S)$, $i = o, w$, and the capillary pressure curve is denoted $P_c(S)$.

In each individual cell Y, the matrix block of porous medium Y_m has dimensionless absolute permeability k and porosity φ with relative permeability $k_{ri}(s)$, $i = o, w$, and capillary pressure curve $p_c(s)$.

Denoting by Ω_m^ε the blocks of porous media, by Ω_f^ε the fissures surrounding Ω_m^ε, and by Γ^ε the boundary between Ω_m^ε and Ω_f^ε, with $(0, T)$ a time interval, we write the conservation of mass in each phase, combined with the generalized Darcy law, as

$$\phi^*(x)\frac{\partial S_o^\varepsilon}{\partial t} - \operatorname{div}\left\{\frac{K^*(x)K_{ro}(S_o^\varepsilon)}{\mu_o}[\nabla P_o^\varepsilon - \rho_o g]\right\} = f_o(x,t) \tag{2.1}$$

and

$$\phi^*(x)\frac{\partial S_w^\varepsilon}{\partial t} - \operatorname{div}\left\{\frac{K^*(x)K_{rw}(S_w^\varepsilon)}{\mu_w}[\nabla P_w^\varepsilon - \rho_w g]\right\} = f_w(x,t) \quad \text{in} \quad \Omega_f^\varepsilon \times (0,T), \tag{2.2}$$

$$\varphi^\varepsilon(x)\frac{\partial s_o^\varepsilon}{\partial t} - \varepsilon^2\operatorname{div}\left\{\frac{k^\varepsilon(x)k_{ro}(s_o^\varepsilon)}{\mu_o}[\nabla p_o^\varepsilon - \rho_o g]\right\} = f_o(x,t) \tag{2.3}$$

and

$$\varphi^\varepsilon(x)\frac{\partial s_w^\varepsilon}{\partial t} - \varepsilon^2\operatorname{div}\left\{\frac{k^\varepsilon(x)k_{rw}(s_w^\varepsilon)}{\mu_w}[\nabla p_w^\varepsilon - \rho_w g]\right\} = f_w(x,t) \quad \text{in} \quad \Omega_m^\varepsilon \times (0,T), \tag{2.4}$$

$$\frac{K^* K_{ro}(S_o^\varepsilon)}{\mu_o} \, [\nabla P_o^\varepsilon - \rho_o g] \cdot \nu = \varepsilon^2 \frac{k^\varepsilon k_{ro}(s_w^\varepsilon)}{\mu_o} \, [\nabla p_o^\varepsilon - \rho_o g] \cdot \nu, \tag{2.5}$$

$$\frac{K^* K_{rw}(S_w^\varepsilon)}{\mu_w} \, [\nabla P_w^\varepsilon - \rho_w g] \cdot \nu = \varepsilon^2 \frac{k^\varepsilon k_{rw}(s_w^\varepsilon)}{\mu_w} \, [\nabla p_w^\varepsilon - \rho_w g] \cdot \nu \quad \text{on} \quad \Gamma^\varepsilon \times (0,T); \tag{2.6}$$

$$P_o^\varepsilon = p_o^\varepsilon \quad \text{on} \quad \Gamma^\varepsilon \times (0,T), \tag{2.7}$$

$$P_w^\varepsilon = p_w^\varepsilon \quad \text{on} \quad \Gamma^\varepsilon \times (0,T); \tag{2.8}$$

$$S_o^\varepsilon(x,0) = S_o^0(x) \quad \text{in} \quad \Omega_f^\varepsilon, \qquad s_o^\varepsilon(x,0) = s_o^0(x) \quad \text{in} \quad \Omega_m^\varepsilon; \tag{2.9}$$

$$S_w^\varepsilon(x,0) = S_w^0(x) \quad \text{in} \quad \Omega_f^\varepsilon, \qquad s_w^\varepsilon(x,0) = s_w^0(x) \quad \text{in} \quad \Omega_m^\varepsilon; \tag{2.10}$$

$$S_o^\varepsilon + S_w^\varepsilon = 1 \quad \text{in} \quad \Omega_f^\varepsilon, \qquad s_o^\varepsilon + s_w^\varepsilon = 1 \quad \text{in} \quad \Omega_m^\varepsilon \times (0,T); \tag{2.11}$$

$$\begin{aligned} P_w^\varepsilon - P_o^\varepsilon &= P_c(S_w^\varepsilon) \quad \text{in} \quad \Omega_f^\varepsilon \times (0,T), \\ p_w^\varepsilon - p_o^\varepsilon &= p_c(s_w^\varepsilon) \quad \text{in} \quad \Omega_m^\varepsilon \times (0,T), \end{aligned} \tag{2.12}$$

where ν denotes the outward normal to Γ^ε.

Also, we define the boundary conditions

$$P_o^\varepsilon = P_o^D \quad \text{on} \quad \Gamma^1 \times (0,T), \qquad P_w^\varepsilon = P_w^D \quad \text{on} \quad \Gamma^1 \times (0,T) \tag{2.13}$$

(i.e., Γ^1 represents the part of the boundary $\partial\Omega$ that is in contact with a liquid continuum).

$$\frac{K^* K_{ro}(S_o^\varepsilon)}{\mu_o} \, [\nabla P_o^\varepsilon - \rho_o g] \cdot \nu = g_o \quad \text{on} \quad \Gamma^2 \times (0,T), \tag{2.14}$$

$$\frac{K^* K_{rw}(S_w^\varepsilon)}{\mu_w} \, [\nabla P_w^\varepsilon - \rho_w g] \cdot \nu = g_w \quad \text{on} \quad \Gamma^2 \times (0,T) \tag{2.15}$$

(i.e., a prescribed flow rate is assumed through the pervious boundary Γ^2).

At this point, we state a number of assumptions on data.

(A1) To avoid too much notation, let $\Omega = \Omega_f^\varepsilon \cup \Omega_m^\varepsilon \cup \Gamma^\varepsilon$ be a cube and $Y =]0,1[^n$ be a cell. Let $\partial\Omega = \overline{\Gamma^1} \cup \overline{\Gamma^2}$, $\Gamma^1 \cap \Gamma^2 = \emptyset$, where each Γ^1 and Γ^2 is an $(n-1)$-dimensional manifold (with boundary).

(A2) Let $\phi^* \in L^\infty(\Omega)$ and $\phi^*(x) \geq \phi_* > 0$. Furthermore, let $\varepsilon^{-1} \in I\!N$ and $\varphi^\varepsilon(x) = \varphi^0\left(\frac{x}{\varepsilon}\right) \in L^\infty(\Omega)$, where φ^0 is a Y-periodic function and $\varphi^\varepsilon(x) \geq \varphi_* > 0$.

(A3) Let $k^\varepsilon(x) = k\left(\frac{x}{\varepsilon}\right)$, where k is a Y-periodic tensor, $k^\varepsilon \in L^\infty(\Omega)^{n^2}$, and

$$0 < k_*|\xi|^2 \leq k^\varepsilon \xi \xi \leq k^*|\xi|^2 \quad \forall \xi \in I\!R^n, \ \xi \neq 0;$$

furthermore, let $K^* \in L^\infty(\Omega)$ and $K^*(x) \geq k_* > 0$.

(A4) The capillary pressures P_c and p_c are strictly monotone increasing and locally Lipschitz continuous in $(0,1)$.

(A5) The saturations $S_i(x,Z)$ and $s_i(x,Z)$, considered as functions of the capillary pressure Z, are measurable in $x \in \Omega$ and continuous in $Z \in [0,1]$. Furthermore, $S_w(x,Z) = 0 = s_w(x,z)$ for $Z \leq P_{\min}$ (for $z \leq p_{\min}$, respectively) and $S_o(x,Z) = 0 = s_o(x,z)$ for $Z \geq P_{\max}$ (for $z \geq p_{\max}$, respectively); $-\infty \leq p_{\min} < 0 < p_{\max} < +\infty$ and $-\infty \leq P_{\min} < 0 < P_{\max} < +\infty$. Finally, S_o and s_o are monotone decreasing and S_w and s_w are monotone increasing in $Z \in [P_{\min}, P_{\max}]$ and in $z \in [p_{\min}, p_{\max}]$, respectively.

Now following [4] and [15], we transform our system (2.1)–(2.15) to a system formulated as an elliptic-parabolic problem for the "reduced" pressure, the total velocity, and the saturation of the w-phase.

Adding (2.1) and (2.2) gives

$$\operatorname{div}\{\varphi_o^\varepsilon + \varphi_w^\varepsilon\} = f_o + f_w \quad \text{in} \quad \Omega_f^\varepsilon \times (0,T), \tag{2.16}$$

where

$$\varphi_o^\varepsilon = -\frac{K^* K_{ro}(S_w^\varepsilon)}{\mu_o}\,[\nabla P_o^\varepsilon - \rho_o g],$$
$$\varphi_w^\varepsilon = -\frac{K^* K_{rw}(S_w^\varepsilon)}{\mu_w}\,[\nabla P_w^\varepsilon - \rho_w g].$$

Now we denote by K_{ro} and K_{rw} the relative permeability curves as functions of S_w^ε, and we define the "reduced" pressure as in [15]:

$$P^\varepsilon = \frac{1}{2}(P_o^\varepsilon + P_w^\varepsilon) + \int_0^{S_w^\varepsilon} \left(\frac{K_{rw}/\mu_w}{K_{ro}/\mu_0 + K_{rw}/\mu_w} - \frac{1}{2}\right)\frac{\partial P_c}{dS}dS. \tag{2.17}$$

We then have with definition (2.17) that

$$\nabla P^\varepsilon = \frac{K_{ro}(S_w^\varepsilon)/\mu_o}{K_{ro}(S_w^\varepsilon)/\mu_o + K_{rw}(S_w^\varepsilon)/\mu_w}\nabla P_o^\varepsilon + \frac{K_{rw}(S_w^\varepsilon)/\mu_w}{K_{ro}(S_w^\varepsilon)/\mu_o + K_{rw}(S_w^\varepsilon)/\mu_w}\nabla P_w^\varepsilon$$

and

$$\begin{aligned}&\varphi_o^\varepsilon + \varphi_w^\varepsilon\\ &= -K^*\left[\frac{K_{ro}(S_w^\varepsilon)}{\mu_o} + \frac{K_{rw}(S_w^\varepsilon)}{\mu_w}\right]\left\{\nabla P^\varepsilon - \frac{(K_{ro}(S_w^\varepsilon)/\mu_o)\rho_o g + (K_{rw}(S_w^\varepsilon)/\mu_w)\rho_w g}{K_{ro}(S_w^\varepsilon)/\mu_o + K_{rw}(S_w^\varepsilon)/\mu_w}\right\}.\end{aligned}$$

For simplicity, we suppose $f_o = 0$ and $f_w = 0$. Then (2.16) and (2.17) imply

$$\begin{aligned}&\operatorname{div}\left\{K^*\left[\frac{K_{ro}(S_w^\varepsilon)}{\mu_o} + \frac{K_{rw}(S_w^\varepsilon)}{\mu_w}\right]\right.\\ &\qquad\left.\cdot\left[\nabla P^\varepsilon - \frac{(K_{ro}(S_w^\varepsilon)/\mu_o)\rho_o + (K_{rw}(S_w^\varepsilon)/\mu_w)\rho_w}{K_{ro}(S_w^\varepsilon)/\mu_o + K_{rw}(S_w^\varepsilon)/\mu_w}g\right]\right\} = 0.\end{aligned} \tag{2.18}$$

Furthermore, we use the identity

$$\left[\frac{K_{ro}(S_w^\varepsilon)}{\mu_o} + \frac{K_{rw}(S_w^\varepsilon)}{\mu_w}\right]\varphi_w^\varepsilon = \frac{-K_{rw}(S_w^\varepsilon)}{\mu_w}\varphi_o^\varepsilon + \frac{K_{ro}(S_w^\varepsilon)}{\mu_o}\varphi_w^\varepsilon + \frac{K_{rw}(S_w^\varepsilon)}{\mu_w}(\varphi_o^\varepsilon + \varphi_w^\varepsilon),$$

which gives

$$\begin{aligned}\varphi_w^\varepsilon = &-\frac{K_{rw}(S_w^\varepsilon)/\mu_w}{K_{ro}(S_w^\varepsilon)/\mu_o + K_{rw}(S_w^\varepsilon)/\mu_w}\,\varphi_o^\varepsilon + \frac{K_{ro}(S_w^\varepsilon)/\mu_o}{K_{ro}(S_w^\varepsilon)/\mu_o + K_{rw}(S_w^\varepsilon)/\mu_w}\,\varphi_w^\varepsilon\\ &+\frac{K_{rw}(S_w^\varepsilon)/\mu_w}{K_{ro}(S_w^\varepsilon)/\mu_o + K_{rw}(S_w^\varepsilon)/\mu_w}\,(\varphi_o^\varepsilon + \varphi_w^\varepsilon).\end{aligned}$$

Now combining the last equality with (2.2) gives

$$\begin{aligned}\phi^*\frac{\partial S_w^\varepsilon}{\partial t} - \operatorname{div}\left\{K^*\frac{K_{rw}K_{ro}}{\mu_w\mu_o(K_{rw}/\mu_w + K_{ro}/\mu_o)}\left[\frac{\partial P_c}{dS}\nabla S_w^\varepsilon + (\rho_o - \rho_w)g\right]\right.\\ \left. - \frac{K_{rw}/\mu_w}{K_{ro}/\mu_o + K_{rw}/\mu_w}(\varphi_o^\varepsilon + \varphi_w^\varepsilon)\right\} = 0.\end{aligned} \tag{2.19}$$

Following Alt and DiBenedetto [4], we introduce the saturations of the wetting fluid as

$$S^\varepsilon(x,t) = S^\varepsilon_w\,(x, P^\varepsilon_w - P^\varepsilon_o) = S^\varepsilon_w\,(x, P_c(x,t)),$$
$$s^\varepsilon(x,t) = s^\varepsilon_w\,(x, p_c\,(x,t));$$

the "reduced" pressures as

$$P^\varepsilon(x,t) = \frac{1}{2}\,(P^\varepsilon_o + P^\varepsilon_w) + \int_0^{S^\varepsilon_w} \left(\frac{K_{rw}/\mu_w}{K_{ro}/\mu_o + K_{rw}/\mu_w} - \frac{1}{2}\right)\frac{\partial P_c}{\partial S}dS,$$
$$p^\varepsilon(x,t) = \frac{1}{2}\,(p^\varepsilon_o + p^\varepsilon_w) + \int_0^{s^\varepsilon_w} \left(\frac{k_{rw}/\mu_w}{k_{ro}/\mu_o + k_{rw}/\mu_w} - \frac{1}{2}\right)\frac{\partial p_c}{\partial s}ds;$$

and the total velocities as

$$V^\varepsilon(x,t) = -K^*(x)\,\Lambda\,(S^\varepsilon)[\nabla P^\varepsilon - E\,(S^\varepsilon)\,g],$$
$$v^\varepsilon(x,t) = -\varepsilon^2 k^\varepsilon(x)\,\lambda\,(s^\varepsilon)[\nabla p^\varepsilon - e\,(s^\varepsilon)\,g],$$

where

$$\Lambda(S) = \frac{K_{ro}(S)}{\mu_o} + \frac{K_{rw}(S)}{\mu_w},$$
$$\lambda(S) = \frac{k_{ro}(S)}{\mu_o} + \frac{k_{rw}(S)}{\mu_w},$$
$$E(S) = \frac{1}{\Lambda(S)}\left[\frac{K_{ro}(S)}{\mu_o}\rho_o + \frac{K_{rw}(S)}{\mu_w}\rho_w\right],$$
$$e(S) = \frac{1}{\lambda(S)}\left[\frac{k_{ro}(S)}{\mu_o}\rho_o + \frac{k_{rw}(S)}{\mu_w}\rho_w\right].$$

To simplify the notation inside (2.19), we introduce the notation

$$A(S) = \frac{K_{rw}(S)K_{ro}(S)}{\mu_o\mu_w\Lambda(S)}\frac{dP_c(S)}{dS},$$
$$a(S) = \frac{k_{rw}(S)k_{ro}(S)}{\mu_o\mu_w\lambda(S)}\frac{dp_c(S)}{dS},$$
$$B(S) = \frac{K_{rw}(S)K_{ro}(S)}{\mu_o\mu_w\Lambda(S)}(\rho_o - \rho_w),$$
$$b(S) = \frac{k_{rw}(S)k_{ro}(S)}{\mu_o\mu_w\lambda(S)}(\rho_o - \rho_w),$$
$$D(S) = \frac{K_{rw}(S)}{\mu_w\Lambda(S)},$$
$$d(S) = \frac{k_{rw}(S)}{\mu_w\lambda(S)}.$$

With this new notation, we have transformed the system (2.1)–(2.6), (2.11), (2.12) into

$$\phi^*\frac{\partial S^\varepsilon}{\partial t} - \operatorname{div}\{K^*A(S^\varepsilon)\nabla S^\varepsilon + K^*B(S^\varepsilon)g - D(S^\varepsilon)V^\varepsilon\} = 0 \quad \text{in } \Omega^\varepsilon_f \times (0,T), \tag{2.20}$$

$$V^\varepsilon = -K^*\Lambda(S^\varepsilon)[\nabla P^\varepsilon - E(S^\varepsilon)g] \quad \text{in } \Omega^\varepsilon_f \times (0,T), \tag{2.21}$$

$$\text{div} V^\varepsilon = 0 \quad \text{in} \quad \Omega_f^\varepsilon \times (0,T), \tag{2.22}$$

$$\varphi^\varepsilon(x)\frac{\partial s^\varepsilon}{\partial t} - \text{div}\,\{\varepsilon^2 k^\varepsilon a(s^\varepsilon)\nabla s^\varepsilon + k^\varepsilon \varepsilon^2 b(s^\varepsilon)g - d(s^\varepsilon)v^\varepsilon\} = 0 \quad \text{in} \quad \Omega_m^\varepsilon \times (0,T), \tag{2.23}$$

$$v^\varepsilon = -\varepsilon^2 k^\varepsilon \lambda(s^\varepsilon)\,[\nabla p^\varepsilon - e(s^\varepsilon)g] \quad \text{in} \quad \Omega_m^\varepsilon \times (0,T), \tag{2.24}$$

$$\text{div}\, v^\varepsilon = 0 \quad \text{in} \quad \Omega_m^\varepsilon \times (0,T), \tag{2.25}$$

where the following assumptions have been made.

(A6) The functions of S_w^ε, K_{ro}, K_{rw}, k_{rw}, and k_{ro} are continuous in $[0,1]$ and strictly positive in $(0,1)$, $K_{ro}(0) > 0$, $K_{rw}(1) > 0$, $k_{rw}(1) > 0$, $k_{ro}(0) > 0$, and $K_{rw}(0) = k_{rw}(0) = K_{ro}(1) = k_{ro}(1) = 0$. Furthermore, the functions Λ and λ satisfy

$$0 < \lambda_* \leq \min\{\Lambda(S),\ \lambda(S)\} < \max\{\Lambda(S),\ \lambda(S)\} \leq \lambda^* < +\infty.$$

(A7) A and a are continuous functions defined on $[0,1]$,

$$A(0) = A(1) = a(0) = a(1) = 0, \qquad A,\ a > 0 \quad \text{on} \quad (0,1).$$

(A8) B and b are continuous functions defined on $[0,1]$,

$$B(0) = B(1) = b(0) = b(1) = 0.$$

(A9) $D, d \in C[0,1]$ and $D(0) = d(0) = 0$.

(A10) $E, e \in C\,[0,1]$ and E and e are strictly positive.

After having derived in (2.20)–(2.25) the equations for P^ε, p^ε, S^ε, s^ε, V^ε, and v^ε, we turn to the boundary conditions (2.7) and (2.8) together with (2.12). The assumptions (A4)–(A5) imply

$$S^\varepsilon = s^\varepsilon \quad \text{on} \quad \Gamma^\varepsilon \times (0,T), \tag{2.26}$$

$$P^\varepsilon = p^\varepsilon \quad \text{on} \quad \Gamma^\varepsilon \times (0,T). \tag{2.27}$$

Finally, (2.5)–(2.6) and (2.22)–(2.24) imply

$$V^\varepsilon \cdot \nu = v^\varepsilon \cdot \nu \quad \text{on} \quad \Gamma^\varepsilon \times (0,T), \tag{2.28}$$

$$K^*[A(S^\varepsilon)\nabla S^\varepsilon + B(S^\varepsilon)g] \cdot \nu = k^\varepsilon[\varepsilon^2 a(s^\varepsilon)\nabla s^\varepsilon + \varepsilon^2 b(s^\varepsilon)g] \cdot \nu \quad \text{on} \quad \Gamma^\varepsilon \times (0,T). \tag{2.29}$$

It remains to write the boundary conditions on the outer boundary $\partial\Omega$ as

$$\left\{\begin{array}{l} V^\varepsilon \cdot \nu = g_o + g_w = 0 \quad \text{on} \quad \Gamma^2 \times (0,T), \\ K^*[A(S^\varepsilon)\nabla S^\varepsilon + B(S^\varepsilon)g] \cdot \nu = g_w \quad \text{on} \quad \Gamma^2 \times (0,T), \\ S^\varepsilon = S_w^\varepsilon(x, P_w^D - P_o^D) \quad \text{on} \quad \Gamma^1 \times (0,T). \end{array}\right. \tag{2.30}$$

To simplify the manipulations, we suppose that $S_w^\varepsilon(x, P_w^D - P_o^D) = 0$, i.e.,

$$S^\varepsilon = 0, P^\varepsilon = \frac{P_o^D + P_w^D}{2} = P_T^D \quad \text{on} \quad \Gamma^1 \times (0,T). \tag{2.31}$$

The initial conditions are

$$(2.32)\qquad S^\varepsilon(x,0)=S_w^0(x)\quad\text{in}\quad \Omega_f^\varepsilon,\qquad s^\varepsilon(x,0)=s_w^0(x)\quad\text{in}\quad \Omega_m^\varepsilon.$$

Now we define a weak solution to the problem (2.20)–(2.32).

DEFINITION (see Antontsev, Kazhikhov, and Monakhov [5]). *The measurable functions P^ε, p^ε, S^ε, s^ε, V^ε, and v^ε are a weak solution to the problem* (2.20)–(2.32) *if the following hold.*

(a) $0\le S^\varepsilon(x,t)\le 1$ *(a.e.) in* $\Omega_f^\varepsilon\times(0,T)$; $0\le s^\varepsilon(x,t)\le 1$ *(a.e.) in* $\Omega_m^\varepsilon\times(0,T)$.

(b) $\nabla P^\varepsilon\in L^\infty((0,T)\ ;\ L^2(\Omega_f^\varepsilon)^n)$, $\nabla p^\varepsilon\in L^\infty((0,T)\ ;\ L^2(\Omega_m^\varepsilon)^n)$,

$$A(S^\varepsilon)\nabla S^\varepsilon\in L^2(\Omega_f^\varepsilon\times(0,T))^n,\qquad \varepsilon a(s^\varepsilon)\nabla s^\varepsilon\in L^2(\Omega_m^\varepsilon\times(0,T))^n.$$

(c) *The boundary conditions* (2.31) *hold on* $\Gamma^1\times(0,T)$.

(d) *Conditions* (2.26) *and* (2.27) *are satisfied.*

(e) *For every* $\varphi\in H^1(\Omega\times(0,T))$ *and every* $\Psi\in H^1(\Omega)$ *satisfying* $\varphi=0$ *on* $\Gamma^1\times(0,T)$ *and* $\Psi=0$ *on* Γ^1, *we have*

$$(2.33)\qquad \int_0^t\int_{\Omega_f^\varepsilon}\phi^*S^\varepsilon\frac{\partial\varphi}{\partial t}+\int_0^t\int_{\Omega_m^\varepsilon}\varphi^\varepsilon s^\varepsilon\frac{\partial\varphi}{\partial t}-\int_0^t\int_{\Omega_f^\varepsilon}A(S^\varepsilon)K^*\nabla S^\varepsilon\nabla\varphi$$
$$-\int_0^t\int_{\Omega_m^\varepsilon}\varepsilon^2k^\varepsilon a(s^\varepsilon)\nabla s^\varepsilon\nabla\varphi-\int_0^t\int_{\Omega_f^\varepsilon}B(S^\varepsilon)K^*g\nabla\varphi-\int_0^t\int_{\Omega_m^\varepsilon}\varepsilon^2b(s^\varepsilon)k^\varepsilon g\nabla\varphi$$
$$+\int_0^t\int_{\Omega_f^\varepsilon}D(S^\varepsilon)V^\varepsilon\nabla\varphi+\int_0^t\int_{\Omega_m^\varepsilon}d(s^\varepsilon)v^\varepsilon\nabla\varphi=\int_{\Omega_f^\varepsilon}\phi^*S^\varepsilon(\cdot,t)\varphi(\cdot,t)dx$$
$$+\int_{\Omega_m^\varepsilon}\varphi^\varepsilon s^\varepsilon(\cdot,t)\varphi(\cdot,t)dx-\int_{\Omega_f^\varepsilon}\phi^*S_w^0\varphi(\cdot,0)-\int_{\Omega_f^\varepsilon}\varphi^\varepsilon s_w^0\varphi(\cdot,0)-\int_0^t\int_{\Gamma^2}g_w\varphi$$

for (a.e.) $t\in(0,T)$ *and*

$$(2.34)\qquad \int_{\Omega_f^\varepsilon}V^\varepsilon\nabla\Psi dx+\int_{\Omega_m^\varepsilon}v^\varepsilon\nabla\Psi dx=0\quad \textit{for (a.e.)}\quad t\in(0,T),$$

where V^ε *and* v^ε *are given by* (2.22) *and* (2.24).

Following Antontsev, Kazhikhov, and Monakhov [5, pp. 203–204], we make the following assumptions:

$$(2.35)\qquad 0\le S_w^0\le 1\quad\text{in}\quad\Omega\quad\text{and}\quad 0\le s_w^0\le 1\quad\text{in}\quad\Omega;$$
$$(2.36)\qquad P_T^D\in L^\infty(0,T\ ;\ H^1(\Omega)),\qquad g_w\in L^\infty((0,T)\times\Gamma^2).$$

Then the existence theory (see Alt and DiBenedetto [4], Antontsev, Kazhikhov, and Monakhov [5], or Kröner and Luckhaus [21]) gives the existence of at least one weak solution for all $\varepsilon>0$.

3. A priori estimates and extensions.

As a simple consequence of the construction of the weak solution, we get the estimates

$$(3.1)\qquad \int_0^T\int_{\Omega_f^\varepsilon}A(S^\varepsilon)|\nabla S^\varepsilon|^2\le C,$$

$$\int_0^T \int_{\Omega_m^\varepsilon} a(s^\varepsilon)|\nabla s^\varepsilon|^2 \le \frac{C}{\varepsilon^2}, \tag{3.2}$$

$$0 \le S^\varepsilon \le 1 \quad \text{(a.e.) in} \quad \Omega_f^\varepsilon \times (0,T), \tag{3.3}$$

$$0 \le s^\varepsilon \le 1 \quad \text{(a.e.) in} \quad \Omega_m^\varepsilon \times (0,T), \tag{3.4}$$

$$\int_0^T \int_{\Omega_f^\varepsilon} |\nabla P^\varepsilon|^2 \le C, \tag{3.5}$$

$$\int_0^T \int_{\Omega_m^\varepsilon} |\nabla p^\varepsilon|^2 \le \frac{C}{\varepsilon^2}, \tag{3.6}$$

$$\int_0^T \int_{\Omega_f^\varepsilon} |V^\varepsilon|^2 \le C, \tag{3.7}$$

$$\int_0^T \int_{\Omega_m^\varepsilon} |v^\varepsilon|^2 \le C\varepsilon^2. \tag{3.8}$$

To get some additional a priori estimates and to homogenize the ε-problem we need to extend S^ε and P^ε to the whole domain $\Omega = \Omega_m^\varepsilon \cup \Omega_f^\varepsilon \cup \Gamma^\varepsilon$.

We assume then the geometry of the cell Y defined as follows.

(A11) Y_m, the matrix part, is a connected open subset of $I\!R^n$; $\overline{Y_m} \subset Y^0$, with Lipschitz boundary; $Y_f = Y \setminus \overline{Y_m}$, the fissure part, is connected.

Now defining $\Omega_f^\varepsilon = \Omega \cap \bigcup_{k\in \mathbb{Z}^n}(Y_f + k)$ from Acerbi et al. [1], there exist three constants $k_i = k_i(Y_f, n, q) > 0$, $i = 0,1,2$, and a linear and continuous extension operator

$$\Pi_\varepsilon \ : W^{1,q}(\Omega_f^\varepsilon) \to W^{1,q}_{\text{loc}}(\Omega) \tag{3.9}$$

such that

$$\Pi_\varepsilon u = u \quad \text{(a.e.) in} \quad \Omega_f^\varepsilon, \tag{3.10}$$

$$\int_{\Omega(\varepsilon k_0)} |\Pi_\varepsilon u|^q dx \le k_1 \int_{\Omega_f^\varepsilon} |u|^q dx, \tag{3.11}$$

$$\int_{\Omega(\varepsilon k_0)} |\nabla(\Pi_\varepsilon u)|^q dx \le k_2 \int_{\Omega_f^\varepsilon} |\nabla u|^q dx \tag{3.12}$$

for all $u \in W^{1,q}(\Omega_f^\varepsilon)$ and where $\Omega(\varepsilon k_0) = \{x \in \Omega \ : \text{dist}(x, \partial\Omega) > \varepsilon k_0\}$.

To avoid dealing with boundary layers, we make the following additional assumption on the structure of the whole domain Ω.

(A12) $\Omega_m^\varepsilon = \Omega(\varepsilon k_0) \cap (\bigcup_{k\in\mathbb{Z}^n} \varepsilon(Y_m + k))$ and $\Omega_f^\varepsilon = \Omega \setminus \overline{\Omega_m^\varepsilon}$.

Now let us derive L^2-estimates for P^ε and p^ε. Supposing that the blocks are removed in an εk_0-neighborhood of $\partial\Omega$, (3.11), (3.12), and (3.5) imply

$$\int_0^T \int_\Omega |\nabla(\Pi_\varepsilon P^\varepsilon)|^2 \le C. \tag{3.13}$$

Using the boundary condition on Γ^1, we get

$$\int_0^T \int_\Omega |\Pi_\varepsilon P^\varepsilon|^2 \le C. \tag{3.14}$$

Furthermore, by the continuity of pressures, we have

$$\int_0^T \int_{\Omega_m^\varepsilon} |p^\varepsilon - \Pi_\varepsilon P^\varepsilon|^2 dxdt \leq C\varepsilon^2 \int_0^T \int_{\Omega_m^\varepsilon} |\nabla(p^\varepsilon - \Pi_\varepsilon P^\varepsilon)|^2 dxdt,$$

which implies

$$\| p^\varepsilon - \Pi_\varepsilon P^\varepsilon \|_{L^2(\Omega_m^\varepsilon \times (0,T))} \leq C\varepsilon \| \nabla p^\varepsilon \|_{L^2(\Omega_m^\varepsilon \times (0,T))^n} + C\varepsilon \| \nabla P^\varepsilon \|_{L^2(\Omega_f^\varepsilon \times (0,T))^n} \leq C$$

and

$$\int_0^T \int_{\Omega_m^\varepsilon} |p^\varepsilon|^2 dxdt \leq C. \tag{3.15}$$

Now we plug a test function $\psi^\varepsilon = p^\varepsilon - \Pi_\varepsilon P^\varepsilon$ in Ω_m^ε and 0 in Ω_f^ε into the variational equation (2.34).

Equation (2.34) and the a priori estimates (3.13)–(3.14) give

$$\| \nabla p^\varepsilon \|_{L^2(\Omega_m^\varepsilon \times (0,T))^n} \leq C. \tag{3.16}$$

We define $\widetilde{p}^\varepsilon$ by

$$\widetilde{p}^\varepsilon = \begin{cases} p^\varepsilon(x,t) & \text{if} \quad x \in \overline{\Omega}_m^\varepsilon, \\ P^\varepsilon(x,t) & \text{if} \quad x \in \Omega_f^\varepsilon \end{cases}$$

and get

$$\begin{cases} \int_0^T \int_\Omega |\widetilde{p}^\varepsilon|^2 dxdt \leq C, \\ \int_0^T \int_\Omega |\nabla \widetilde{p}^\varepsilon|^2 dxdt \leq C. \end{cases} \tag{3.17}$$

Now we define $Z(S) = \int_0^S \sqrt{A(\eta)} d\eta$. Obviously, Z is a monotone function of S. Furthermore, we set $Z^\varepsilon = Z(S^\varepsilon)$.

Then

$$0 \leq Z^\varepsilon \leq M_Z = \max_{\eta \in [0,1]} \sqrt{A(\eta)} \quad \text{(a.e.) in} \quad \Omega_f^\varepsilon \times (0,T),$$

$$\nabla Z^\varepsilon = \sqrt{A(S^\varepsilon)} \nabla S^\varepsilon \in L^2(\Omega_f^\varepsilon \times (0,T)).$$

Hence

$$\int_0^T \int_\Omega |\nabla(\Pi_\varepsilon Z^\varepsilon)|^2 dxdt \leq C, \tag{3.18}$$

$$0 \leq \Pi_\varepsilon Z^\varepsilon \leq M_Z \quad \text{(a.e.) in} \quad \Omega \times (0,T). \tag{3.19}$$

Introducing $\overline{S}^\varepsilon$ as $\overline{S}^\varepsilon = (Z)^{-1}(\Pi_\varepsilon Z^\varepsilon)$ leads to

$$\nabla \overline{S}^\varepsilon = \frac{1}{dZ/dS} \nabla(\Pi_\varepsilon Z^\varepsilon) = \frac{1}{\sqrt{a(\overline{S}^\varepsilon)}} \nabla(\Pi_\varepsilon Z^\varepsilon)$$

and to the estimate

$$\int_0^T \int_\Omega a(\overline{S}^\varepsilon) |\nabla \overline{S}^\varepsilon|^2 dxdt \leq C, \tag{3.20}$$

$$0 \le \overline{S}^\varepsilon \le 1 \quad \text{(a.e.) in} \quad \Omega \times (0,T). \tag{3.21}$$

To extend s^ε, we introduce as before $z = \int_0^s \sqrt{a(\eta)}d\eta$.
Since $dz/ds = \sqrt{a(s)} \ge 0$, z is monotone.
Once again, we have

$$0 \le z^\varepsilon \le M_z = \max_{\eta\in[0,1]} \sqrt{a(\eta)} \quad \text{(a.e.) in} \quad \Omega_m^\varepsilon \times (0,T),$$

$$\nabla z^\varepsilon = \sqrt{a(s^\varepsilon)}\nabla s^\varepsilon \in L^2(\Omega_m^\varepsilon \times (0,T)).$$

Hence for

$$\widetilde{z}^\varepsilon = \begin{cases} z^\varepsilon(x,t) & \text{for} \quad x \in \overline{\Omega}_m^\varepsilon, \\ Z^\varepsilon(x,t) & \text{for} \quad x \in \overline{\Omega}_f^\varepsilon, \end{cases}$$

we have

$$\int_0^T \int_\Omega |\nabla \widetilde{z}^\varepsilon|^2 dxdt \le C/\varepsilon^2, \tag{3.22}$$

$$0 \le \widetilde{z}^\varepsilon \le \max\{M_Z, M_z\} \quad \text{(a.e.) on} \quad \Omega \times (0,T). \tag{3.23}$$

Finally, we set

$$\widetilde{s}^\varepsilon = \begin{cases} s^\varepsilon(x,t) & \text{for} \quad x \in \overline{\Omega}_m^\varepsilon, \\ S^\varepsilon(x,t) & \text{for} \quad x \in \Omega_f^\varepsilon. \end{cases}$$

Then

$$0 \le \widetilde{s}^\varepsilon \le 1 \quad \text{(a.e.) on} \quad \Omega \times (0,T). \tag{3.24}$$

After deriving the estimates for the spatial derivatives, we turn to the behavior in time. In the case of nondegenerate nonlinear problems, one method is to formulate an estimate for the time derivative of a fractional order (see Aganović and Mikelić [2]). For this degenerate parabolic-elliptic system, we use a similar but more direct approach.

We start with the variational formulation (2.33) and define

$$B_1^\varepsilon(x,t) = A(S^\varepsilon)K^*\nabla S^\varepsilon + B(S^\varepsilon)K^*g - D(S^\varepsilon)V^\varepsilon \quad \text{in} \quad \Omega_f^\varepsilon \times (0,T),$$

$$B_2^\varepsilon(x,t) = a(s^\varepsilon)k^*\nabla s^\varepsilon + b(s^\varepsilon)k^*g - \varepsilon^{-2}v^\varepsilon d(s^\varepsilon) \quad \text{in} \quad \Omega_m^\varepsilon \times (0,T).$$

It is natural to set $B_i^\varepsilon = 0$ for $t \notin (0,T)$. Then

$$\int_{I\!R} \int_{\Omega_f^\varepsilon} |B_1^\varepsilon|^2 < +\infty \quad \text{and} \quad \int_{I\!R} \int_{\Omega_f^\varepsilon} \varepsilon^2 |B_2^\varepsilon|^2 < +\infty.$$

Let us take $\varphi\ (x,t) = \eta(t)\xi(x)$, $\eta \in C_0^\infty((0,T))$, $\xi \in H_0^1(\Omega)$ as a test function in (2.33). We get

$$\int_0^T \left\{ \frac{\partial \eta}{\partial t} \left[\int_{\Omega_f^\varepsilon} \phi^* S^\varepsilon \xi + \int_{\Omega_w^\varepsilon} \varphi^\varepsilon s^\varepsilon \xi \right] - \eta \left[\int_{\Omega_m^\varepsilon} B_1^\varepsilon \nabla \xi + \int_{\Omega_m^\varepsilon} \varepsilon^2 B_2^\varepsilon \nabla \xi \right] \right\} dt = 0,$$

which implies

$$-\frac{d}{dt}\left[\int_{\Omega_f^\varepsilon} \phi^* S^\varepsilon \xi + \int_{\Omega_m^\varepsilon} \varphi^\varepsilon s^\varepsilon \xi\right] = \int_{\Omega_f^\varepsilon} B_1^\varepsilon \nabla \xi + \int_{\Omega_m^\varepsilon} \varepsilon^2 B_2^\varepsilon \nabla \xi. \tag{3.25}$$

Now we choose an interval $[\tau, \tau + \Delta\tau] \subset (0, T)$, $\tau > 0, \Delta\tau > 0$, and after integrating (3.25) over that interval we find the equality

$$-\int_{\Omega_f^\varepsilon} \phi^* \xi \Delta_\tau S^\varepsilon - \int_{\Omega_m^\varepsilon} \varphi^\varepsilon \xi \Delta_\tau s^\varepsilon = \int_\tau^{\tau+\Delta\tau} \int_{\Omega_f^\varepsilon} B_1^\varepsilon \nabla \xi + \int_\tau^{\tau+\Delta\tau} \int_{\Omega_m^\varepsilon} \varepsilon^2 B_2^\varepsilon \nabla \xi, \tag{3.26}$$

where $\Delta_\tau u = u(\cdot, \tau + \Delta\tau) - u(\cdot, \tau)$. Noting that τ is a parameter in (3.26), we choose $\xi = \Delta_\tau \widetilde{z}^\varepsilon$ as a test function in (3.26). We get

$$\begin{aligned}
&\int_{\Omega_f^\varepsilon} \phi^* \Delta_\tau Z^\varepsilon \Delta_\tau S^\varepsilon + \int_{\Omega_m^\varepsilon} \varphi^\varepsilon \Delta_\tau z^\varepsilon \Delta_\tau s^\varepsilon \\
&\leq \left| \int_\tau^{\tau+\Delta\tau} \int_{\Omega_f^\varepsilon} B_1^\varepsilon(x,t) \cdot \nabla \Delta_\tau Z^\varepsilon dx dt \right| + \left| \int_\tau^{\tau+\Delta\tau} \int_{\Omega_m^\varepsilon} \varepsilon^2 B_2^\varepsilon(x,t) \nabla \Delta_\tau z^\varepsilon dx dt \right| \\
&\leq \left| \int_0^1 \int_{\Omega_f^\varepsilon} B_1^\varepsilon(x, \tau + \sigma\Delta\tau) \nabla \Delta_\tau Z^\varepsilon d\sigma dx \right| |\Delta\tau| + |\Delta\tau| \\
&\quad \cdot \left| \int_0^1 \int_{\Omega_m^\varepsilon} \varepsilon^2 B_2^\varepsilon(x, \tau + \sigma\Delta\tau) \nabla \Delta_\tau z^\varepsilon d\sigma dx \right|.
\end{aligned}$$

Integration in time over $[0, T - \Delta\tau]$ and Hölder's inequality give

$$\begin{aligned}
&\int_0^{T-\Delta\tau} \int_{\Omega_f^\varepsilon} \phi^* \Delta_\tau Z^\varepsilon \Delta_\tau S^\varepsilon + \int_0^{T-\Delta\tau} \int_{\Omega_m^\varepsilon} \varphi^\varepsilon \Delta_\tau z^\varepsilon \Delta_\tau s^\varepsilon \leq |\Delta\tau|, \\
&\{\|B_1^\varepsilon \|_{L^2(\Omega_f^\varepsilon \times (0,T))^n} \cdot \| \nabla \Delta_\tau Z^\varepsilon \|_{L^2(\Omega_f^\varepsilon \times [0, T-\Delta\tau])^n} \\
&\quad + \| B_2^\varepsilon \|_{L^2(\Omega_w^\varepsilon \times (0,T))^n} \| \varepsilon \nabla \Delta_\tau z^\varepsilon \|_{L^2(\Omega_m^\varepsilon \times [0, T-\Delta\tau])^n} \} \leq C|\Delta\tau|.
\end{aligned}$$

Using monotonicity, we get

$$\int_0^{T-\Delta\tau} \int_\Omega [\widetilde{z}^\varepsilon(\cdot, \cdot + \Delta\tau) - \widetilde{z}^\varepsilon(\cdot, \cdot)] [\widetilde{s}^\varepsilon(\cdot, \cdot + \Delta\tau) - \widetilde{s}^\varepsilon(\cdot, \cdot)] \leq C|\Delta\tau| \tag{3.27}$$

and

$$\int_0^{T-\Delta\tau} \int_\Omega |\widetilde{z}^\varepsilon(\cdot, \cdot + \Delta\tau) - \widetilde{z}^\varepsilon(\cdot, \cdot)|^2 \leq C|\Delta\tau|. \tag{3.28}$$

For Z^ε, we get

$$\int_0^{T-\Delta\tau} \int_{\Omega_f^\varepsilon} |Z^\varepsilon(\cdot, \cdot + \Delta\tau) - Z^\varepsilon(\cdot, \cdot)|^2 \leq C|\Delta\tau|,$$

and since our extension is by reflection, we conclude
(3.29)

$$\int_0^{T-\Delta\tau} \int_\Omega [(\Pi_\varepsilon Z^\varepsilon)(\cdot, \cdot + \Delta\tau) - (\Pi_\varepsilon Z^\varepsilon)(\cdot, \cdot)] \left[\overline{S}^\varepsilon(\cdot, \cdot + \Delta\tau) - \overline{S}^\varepsilon(\cdot, \cdot)\right] \leq C|\Delta\tau|$$

and

$$\int_0^{T-\Delta\tau}\int_\Omega |(\Pi_\varepsilon Z^\varepsilon)(\cdot,\cdot+\Delta\tau)-(\Pi_\varepsilon Z^\varepsilon)(\cdot,\cdot)|^2 \le C|\Delta\tau|. \tag{3.30}$$

4. Compactness and convergence results. Using the a priori estimates derived in §3 and the concept of two-scale convergence as in [13] and [24], we get the following compactness results.

PROPOSITION 4.1. *There exists a subsequence such that*

(4.1) $\Pi_\varepsilon P^\varepsilon \longrightarrow P \in L^2((0,T);H^1(\Omega))$ *in the two-scale sense*;

(4.2) $\nabla(\Pi_\varepsilon P^\varepsilon) \longrightarrow \nabla P + \nabla_y P_1(t,x,y)$ *in the two-scale sense, where*

$$P_1 \in L^2((0,T)\times\Omega\ ;\ H^1_{\text{per}}(Y));$$

(4.3) $\widetilde{p}^\varepsilon \longrightarrow p \in L^2((0,T);H^1(\Omega))$ *in the two-scale sense*;

(4.4) $\Pi_\varepsilon Z^\varepsilon \longrightarrow Z \in L^2((0,T);\ H^1(\Omega))$ *in the two-scale sense*;

(4.5) $\nabla(\Pi_\varepsilon Z^\varepsilon) \longrightarrow \nabla Z + \nabla_y Z_1(t,x,y)$ *in the two-scale sense, where*

$$Z_1 \in L^2((0,T)\times\Omega\ ;\ H^1_{\text{per}}(Y));$$

(4.6) $\Pi_\varepsilon Z^\varepsilon \longrightarrow Z$ *strongly in* $L^q((0,T)\times\Omega)$ $\forall q < +\infty$;

(4.7) $\overline{S}^\varepsilon \longrightarrow S$ *strongly in* $L^q((0,T)\times\Omega)$ $\forall q < +\infty$;

(4.8) $\widetilde{s}^\varepsilon \longrightarrow s \in L^2((0,T)\times\Omega\times Y)$ *in the two scale-sense*;

(4.9) $\widetilde{z}^\varepsilon \longrightarrow z \in L^2((0,T)\times\Omega\ ;H^1_{\text{per}}(Y))$ *in the two scale-sense*;

(4.10) $\varepsilon\nabla\widetilde{z}^\varepsilon \longrightarrow \nabla_y z \in L^2((0,T)\times\Omega\times Y)$ *in the two scale-sense.*

Furthermore, let $\widetilde{V}^\varepsilon$ *be equal to* V^ε *in* $\Omega_f^\varepsilon\times(0,T)$ *and* v^ε *in* $\Omega_m^\varepsilon\times(0,T)$. *Then we have*

(4.11) $\widetilde{V}^\varepsilon \longrightarrow V \in L^2((0,T)\times\Omega\times Y)$ *in the two scale-sense*;

(4.12) $\text{div}_y V = 0$ *and* $\text{div}_x \int_Y V(x,y)dy = 0$.

Proof. The proof is a direct consequence of the a priori estimates and of Theorem 1.2 and Proposition 1.4 in Allaire [3]. □

Again as in Allaire [3], we plug into (2.34) a test function of the form

$$\varphi(x,t)+\varepsilon\varphi_1(x,x/\varepsilon,t)+\Psi(x,x/\varepsilon,t),$$

where $\varphi\in\mathcal{D}(Q_T)$, $\varphi_1\in\mathcal{D}(Q_T\ ;C^\infty_{\text{per}}(Y))$, and $\Psi\in\mathcal{D}(Q_T\ ;C^\infty_{\text{per}}(Y))$ with $\Psi=0$ for $y\in Y_f$.

We get

$$\int_0^T\int_{\Omega_f^\varepsilon} K^*\Lambda(S^\varepsilon)[\nabla P^\varepsilon - E(S^\varepsilon)g][\nabla\varphi+\varepsilon\nabla_x\varphi_1^\varepsilon+\nabla_y\varphi_1^\varepsilon]$$
$$+\int_0^T\int_{\Omega_m^\varepsilon}\varepsilon^2 k^\varepsilon\lambda(s^\varepsilon)[\nabla p^\varepsilon - e(s^\varepsilon)g]\left[\nabla\varphi+\varepsilon\nabla_x\varphi_1^\varepsilon+\nabla_y\varphi_1^\varepsilon+\nabla_x\Psi^\varepsilon+\frac{1}{\varepsilon}\nabla_y\Psi^\varepsilon\right]=0.$$

Passing to the two-scale limit yields

$$\int_0^T\int_\Omega\int_{Y_f} K^*\Lambda(S)[\nabla P+\nabla_y P_1 - E(S)g][\nabla\varphi(x,t)+\nabla_y\varphi_1(x,y,t)]$$
$$+\lim_{\varepsilon\to 0}\int_0^T\int_{\Omega_m^\varepsilon} k^\varepsilon\lambda(s^\varepsilon)\varepsilon\nabla p^\varepsilon\nabla_y\Psi^\varepsilon=0.$$

Choosing $\Psi = 0$ gives

$$-\mathrm{div}_x \left\{ \int_{Y_f} K^*(x)\Lambda(S)[\nabla P + \nabla_y P_1 - E(S)g] \right\} = 0 \quad \text{in} \quad Q_T \tag{4.13}$$

and

$$-\mathrm{div}_y \left\{ K^*(x)\Lambda(S)[\nabla P + \nabla_y P_1 - E(S)g] \right\} = 0 \quad \text{in} \quad Y_f$$

with

$$K^*(x)\Lambda(S)[\nabla P + \nabla_y P_1 - E(S)g] \cdot \nu = 0 \quad \text{on} \quad \partial Y_m$$

for a.e. $x, t \in Q_T$.

We shall use Θ, the tensor whose (i,j) component is $\partial\xi_j/\partial y_i$, where ξ_j is a periodic solution in Y of the auxiliary problem

$$\begin{cases} \Delta_y \xi_j = 0 & \text{in} \quad Y_f, \\ \nabla_y \xi_j \cdot \nu = -e_j \cdot \nu & \text{on} \quad \partial Y_m. \end{cases} \tag{4.14}$$

Since S is independent of y from (4.6) and (4.7), P_1 is given by the product

$$P_1 = \sum_j \xi_j(y) \left(\frac{\partial P}{\partial x_j} - E(S)g\delta_{j3} \right).$$

Finally, (4.13) reduces to

$$\begin{cases} \overline{V} = K^{*H} A(S) \left[\nabla_x P - E(S)g \right] = \dfrac{1}{|Y|} \displaystyle\int_Y V dy, \\ -\mathrm{div}_x \overline{V} = 0 \quad \text{in} \quad Q_T \end{cases} \tag{4.15}$$

with

$$K^{*H} = \frac{1}{|Y|} \int_{Y_f} K^*(I + \Theta) dy = \frac{|Y_f|}{|Y|} K^* \left(I + \frac{1}{|Y_f|} \int_{Y_f} \Theta(y) dy \right).$$

In our next step, we plug in a test function of the same form as above into (2.33). We have

$$\begin{aligned}
&\int_0^T \int_{\Omega_f^\varepsilon} \phi^* S^\varepsilon \left[\frac{\partial \varphi}{\partial t} + \varepsilon \frac{\partial \varphi_1^\varepsilon}{\partial t} \right] + \int_0^T \int_{\Omega_m^\varepsilon} \varphi^\varepsilon s^\varepsilon \left[\frac{\partial \varphi}{\partial t} + \varepsilon \frac{\partial \varphi_1^\varepsilon}{\partial t} + \frac{\partial \Psi}{\partial t} \right] \\
&- \int_0^T \int_{\Omega_f^\varepsilon} \sqrt{A(S^\varepsilon)} K^* \nabla Z^\varepsilon [\nabla \varphi + \varepsilon \nabla_x \varphi_1^\varepsilon + \nabla_y \varphi_1^\varepsilon] \\
&- \int_0^T \int_{\Omega_m^\varepsilon} \varepsilon \sqrt{a(s^\varepsilon)} K^* \varepsilon \nabla z^\varepsilon \left[\nabla \varphi + \varepsilon \nabla_x \varphi_1^\varepsilon + \nabla_y \varphi_1^\varepsilon + \nabla_x \Psi^\varepsilon + \frac{1}{\varepsilon} \nabla_y \Psi^\varepsilon \right] \\
&- \int_0^T \int_{\Omega_f^\varepsilon} B(S^\varepsilon) K^* g [\nabla \varphi + \varepsilon \nabla_x \varphi_1^\varepsilon + \nabla_y \varphi_1^\varepsilon] \\
&- \int_0^T \int_{\Omega_m^\varepsilon} \varepsilon^2 b(s^\varepsilon) k^\varepsilon g \left[\nabla \varphi + \varepsilon \nabla_x \varphi_1^\varepsilon + \nabla_y \varphi_1^\varepsilon + \nabla_x \Psi^\varepsilon + \frac{1}{\varepsilon} \nabla_y \Psi^\varepsilon \right] \\
&+ \int_0^T \int_{\Omega_m^\varepsilon} d(s^\varepsilon) v^\varepsilon \left[\nabla \varphi + \varepsilon \nabla_x \varphi_1^\varepsilon + \nabla_y \varphi_1^\varepsilon + \nabla_x \Psi^\varepsilon + \frac{1}{\varepsilon} \nabla_y \Psi^\varepsilon \right] \\
&+ \int_0^T \int_{\Omega_f^\varepsilon} D(S^\varepsilon) V^\varepsilon [\nabla \varphi + \varepsilon \nabla_x \varphi_1^\varepsilon + \nabla_y \varphi_1^\varepsilon] = 0.
\end{aligned}$$

Passing to the two-scale limit yields

$$\int_0^T \int_\Omega \int_{Y_f} \phi^* S \frac{\partial \varphi(x,t)}{\partial t} + \int_0^T \int_\Omega \int_{Y_m} \varphi^0(x) s(x,y) \cdot \left[\frac{\partial \varphi(x,t)}{\partial t} + \frac{\partial}{\partial t} \Psi(x,y,t) \right]$$
$$- \int_0^T \int_\Omega \int_{Y_f} \sqrt{A(S)} K^* [\nabla Z + \nabla_y Z_1][\nabla \varphi + \nabla_y \varphi_1(x,y,t)]$$
$$- \lim_{\varepsilon \to 0} \int_0^T \int_{\Omega_m^\varepsilon} \sqrt{a(s^\varepsilon)} K^* \varepsilon \nabla z^\varepsilon \nabla_y \Psi^\varepsilon - \int_0^T \int_\Omega \int_{Y_f} B(S) K^* g [\nabla \varphi + \nabla_y \varphi_1(x,y)]$$
$$- \int_0^T \int_\Omega \int_{Y_f} D(S) V [\nabla \varphi + \nabla_y \varphi] + \lim_{\varepsilon \to 0} \int_0^T \int_{\Omega_m^\varepsilon} d(s) \frac{v^\varepsilon}{\varepsilon} \nabla_y \Psi^\varepsilon = 0.$$

The choice $\Psi = 0$ gives

$$|Y_f| \phi^* \frac{\partial S(x,t)}{\partial t} + \int_{Y_m} \varphi^0(y) \frac{\partial s(x,y,t)}{\partial t} dy$$
$$- \mathrm{div}_x \left\{ \sqrt{A(S)} \int_{Y_f} K^*(x) [\nabla_x Z + \nabla_y Z_1] dy + |Y_f| B(S) K^* g - D(S) \int_{Y_f} V(x,y) dy \right\}$$
$$= 0 \quad \text{in} \quad Q_T \tag{4.16}$$

and

$$\left\{ \begin{array}{c} \mathrm{div}_y \{ \sqrt{A(S)} K^*(x) [\nabla_x Z + \nabla_y Z_1] + B(S) K^* g - D(S) V(x,y) \} = 0 \\ \text{in} \quad Y_f \quad \text{for (a.e.)} \quad x,t \in Q_T, \\ \sqrt{A(S)} K^* [\nabla_x Z + \nabla_y Z_1] \nu + B(S) K^* g \cdot \nu = 0 \quad \text{on} \quad \partial Y_m \end{array} \right. \tag{4.17}$$

since

$$V \cdot \nu = K^* \Lambda(S) [\nabla P + \nabla_y P_1 + E(S) g] \cdot \nu = 0 \quad \text{on} \quad \partial Y_m.$$

Using (as for the pressure equation (4.15)) Θ defined by (4.14) and

$$K^{*H} = \frac{1}{|Y|} \int_{Y_f} K^* (I + \Theta) dy = \frac{|Y_f|}{|Y|} K^* \left(I + \frac{1}{|Y_f|} \int_{Y_f} \Theta(y) dy \right), \tag{4.18}$$

we obtain $Z_1 = \sum_j \xi_j(y) \big(\partial Z / \partial x_j + B(S) / \sqrt{A(S)} g \delta_{j3} \big)$.

Finally, (4.16) reduces to

$$\frac{|Y_f|}{|Y|} \phi^* \frac{\partial S}{\partial t} - \mathrm{div}_x \left\{ K^{*H} \left(A(S) \nabla_x S + B(S) g \right) - D(S) \overline{V} \right\}$$
$$= - \frac{1}{|Y|} \int_{Y_m} \varphi^0(y) \frac{\partial s(x,y,t)}{\partial t} \quad \text{in} \quad Q_T. \tag{4.19}$$

However, in (4.16) we have the term

$$\frac{\partial}{\partial t} \int_{Y_m} \varphi^0(y) s(x,y,t) dy,$$

which belongs to $H^{-1}(Q_T)$ a priori.

In fact, we will show that this term is more regular.

LEMMA 4.2.

$$\partial_t \int_{Y_m} \varphi^0(y) s(x,y,t) dy \in L^2(Q_T).$$

Proof. We introduce F_ε as

$$F_\varepsilon = \varepsilon^2 \sqrt{a(s^\varepsilon)} k^\varepsilon(x) \nabla z^\varepsilon - d(s^\varepsilon) v^\varepsilon = \varepsilon \overline{F}_\varepsilon$$

with $\| \overline{F}_\varepsilon \|_{L^2(\Omega^\varepsilon_m \times (0,T))^n} \leq C$.

Obviously, $\overline{F}_\varepsilon \longrightarrow F^*$ in the two-scale sense and

$$\varphi^0(y) \partial_t s - \operatorname{div}_y F^*(x,y) = 0 \quad \text{in} \quad H^{-1}(Q_T \times Y_m).$$

Consequently, for all $\xi \in H^1_0(0,T)$, we have

$$-\int_0^T \varphi^0(y) s \partial_t \xi - \operatorname{div}_y \left\{ \int_0^T F^* \xi \right\} = 0 \quad \text{in} \quad H^{-1}(\Omega \times Y_m),$$

which gives

$$-\operatorname{div}_y \left\{ \int_0^T F^* \xi \right\} \in L^\infty(Y_m) \quad \text{for (a.e)} \quad x \in \Omega$$

and for all $\xi \in H^1_0(0,T)$.

Hence

$$\int_0^T \xi F^* \in L^2(\Omega \times Y_m)^n \quad \text{and} \quad \operatorname{div}_y \left\{ \int_0^T \xi F^* \right\} \in L^2(\Omega \times Y_m),$$

and we conclude that $\int_0^T \xi F^* \cdot \nu \in L^2\left(\Omega\, ; H^{-1/2}(\partial Y_m)\right)$.

Now we obtain

$$-\int_0^T \left[\int_{Y_m} \varphi^0(y) s(x,y,t) \right] \partial_t \xi dt = \int_0^T \int_{\partial Y_m} \xi F^* \cdot \nu \in L^2(\Omega).$$

The right-hand side is well defined for all $\xi \in L^2(0,T)$. Therefore,

$$\partial_t \int_{Y_m} \varphi^0(s) s(x,y,t) dy \in L^2(Q_T). \qquad \square$$

Our next step is to show that $p = P$. We have the following result.

LEMMA 4.3. *Let P and p be defined by* (4.1)–(4.2) *and* (4.3), *respectively. Then we have*

$$P(x,t) = p(x,t) \quad \textit{for (a.e.)} \quad (x,t) \in Q_T. \tag{4.20}$$

Proof. By the continuity of the pressures we have

$$\int_0^T \int_{\Omega^\varepsilon_m} |p^\varepsilon - \Pi_\varepsilon P^\varepsilon|^2 dx\, dt \leq C\varepsilon^2 \int_0^T \int_{\Omega^\varepsilon_m} |\nabla(P^\varepsilon - \Pi_\varepsilon P^\varepsilon)|^2 dx\, dt \leq C\varepsilon^2. \tag{4.21}$$

Passing to the limit $\varepsilon \to 0$ yields (4.20). $\square$

It remains to find the equations satisfied by s and z. The first result is the identification of initial conditions.

LEMMA 4.4. *Let S and s be defined by* (4.8) *and* (4.9), *respectively. Then we have*

$$|Y_f|\phi^* S(x,0) + \int_{Y_m} \varphi^0(y) s(x,y,0)\,dy \\ = \phi^* S_w^0(x)|Y_f| + s_w^0(x) \int_{Y_m} \varphi^0(y)dy \quad (a.e.) \quad x \in \Omega \tag{4.22}$$

and

$$s(x,y,0) = s_w^0(x) \quad (a.e.) \quad (x,y) \in \Omega \times Y_m. \tag{4.23}$$

Our next step is to look for an equation for s. First, it should be noticed that by (2.24) and (3.16),

$$\| v^\varepsilon \|_{L^2(\Omega_m^\varepsilon \times (0,T))} \leq C\varepsilon^2 \tag{4.24}$$

and

$$\left| \int_0^T \int_{\Omega_m^\varepsilon} d(s^\varepsilon) v^\varepsilon \nabla \Psi \right| \leq C\varepsilon. \tag{4.25}$$

Consequently, there will be no transport term in the equation for s. The remaining nonlinear term in (2.23) is monotone. Therefore, it would be natural to try passing to the limit using monotonicity. This approach would work nicely if the nonlinearity was not degenerate. The degeneration makes the standard monotonicity argument (see, e.g., Lions [22, pp. 190–204]) fairly complicated if not impossible, and we have chosen a different approach, that is, the periodic modulation. The approach through the periodic modulation makes it possible to go to the limit in the degenerate terms that depend nonlinearly on s^ε.

DEFINITION 4.5. *For a given $\varepsilon > 0$, we define a dilation operator D^ε mapping measurable functions on $\Omega_m^\varepsilon \times (0,T)$ to measurable functions on $\Omega \times Y_m \times (0,T)$ by*

$$(D^\varepsilon u)(x,y,t) = u(c^\varepsilon(x) + \varepsilon y, t), \qquad y \in Y_m, \quad (x,t) \in \Omega \times (0,T), \tag{4.26}$$

where $c^\varepsilon(x)$ denotes the lattice translation point of the ε-cell domain containing x. We extend $(D^\varepsilon u)$ from Y_m to $\bigcup_k (Y_m + k)$ periodically.

It is easy to see that

$$\begin{cases} \| D^\varepsilon u \|_{L^q(Q_T \times Y_m)} = \| u \|_{L^q(Q_T)}, & 1 \leq q < +\infty, \\ \nabla_y D^\varepsilon u = \varepsilon D^\varepsilon \nabla_x u \quad \text{(a.e.)} & \text{in } Q_T \times Y_m \end{cases} \tag{4.27}$$

for all $u \in L^2(0,T\ ; W^{1,q}\ (\Omega_m^\varepsilon))$. For more properties of the dilation operator D^ε, we refer to Arbogast, Douglas, and Hornung [6, pp. 828–831].

The connection between the periodic modulation and H-measures is discussed in Tartar [29, pp. 203–204].

To proceed, we have to establish in the following proposition the link between the two-scale convergence and the weak convergence of a periodically modulated sequence.

PROPOSITION 4.6. *Let $\{u^\varepsilon\}$ be a uniformly bounded sequence in $L^2(\Omega^\varepsilon_m \times (0,T))$ that satisfies the conditions*

$$D^\varepsilon u^\varepsilon \rightharpoonup u^0 \quad \textit{weakly in} \quad L^2(Q_T; L^2_{\mathrm{per}}(Y_m))$$

and

$$\chi_{\Omega^\varepsilon_m} u^\varepsilon \longrightarrow u^* \in L^2(Q_T\ ; L^2_{\mathrm{per}}(Y)) \quad \textit{in the two-scale sense.}$$

Then we have

$$u^0 = u^* \quad \textit{(a.e.) in} \quad Q_T \times Y_m. \tag{4.28}$$

Proof. Let $\psi \in C^\infty_0(Q_T)$ and $h \in C^\infty_{\mathrm{per}}(\overline{Y}_m)$. Then we have

$$\begin{aligned}
&\int_{Q_T}\int_{Y_m} (D^\varepsilon u^\varepsilon)(x,y,t)\psi(x,t)h(y)dydxdt \\
&\quad = \int_{Q_T}\int_{Y_m} u^\varepsilon(c^\varepsilon(x)+\varepsilon y,t)\psi(x,t)h(y)dydxdt \\
&\quad = \sum_k \int_0^T \left\{\int_{\varepsilon(Y+k)}\int_{Y_m} \psi(x,t)u^\varepsilon(\varepsilon k+\varepsilon y,t)h(y)dydx\right\} dt.
\end{aligned}$$

Now we use the estimate

$$\left|\varepsilon^{-n}\int_{\varepsilon(Y+k)} \psi(x,t)dx - \psi(z,t)J\right| \le C\varepsilon \quad \forall z \in \varepsilon(Y_m+k)$$

and get

$$\begin{aligned}
&\int_{Q_T}\int_{Y_m} (D^\varepsilon u^\varepsilon)(x,y,t)\psi(x,t)h(y)dy\ dx\ dt = \sum_k \int_0^T \left[\int_{\varepsilon(Y+k)} \psi(x,t)dx\right], \\
&\quad \left\{\varepsilon^{-n}\int_{\varepsilon(Y_m+k)} u^\varepsilon(z,t)h(z/\varepsilon)dz\right\} dt = O(\varepsilon) + \sum_k \int_0^T \int_{\varepsilon(Y_m+k)} u^\varepsilon(z,t), \\
&\quad h(z/\varepsilon)\psi(z,t)dz\ dt = O(\varepsilon) + \int_{Q_T} u^\varepsilon(z,t)h(z/\varepsilon)\chi_{\Omega^\varepsilon_m}(z)\psi(z,t)dzdt \\
&\quad \to \int_{Q_T}\int_{Y_m} u^*(z,y,t)h(y)\psi(z,t)dz\ dy\ dt = \int_{Q_T}\int_{Y_m} u^0(x,y,t)h(y)\psi(x,t)dx\ dy\ dt.
\end{aligned}$$

Hence $u^0 = u^*$ (a.e.) in $Q_T \times Y_m$. □

Some additional useful properties of the dilation operator are given by the following lemma and corollary from Arbogast, Douglas, and Hornung [6].

LEMMA 4.7. *Let $\varphi,\ \psi \in L^2(0,T\ ; H^1(\Omega^\varepsilon_m))$. Then*

$$\begin{aligned}
(D^\varepsilon\varphi, D^\varepsilon\psi)_{L^2(Q_T\times Y_m)} &= (\varphi,\psi)_{L^2(\Omega^\varepsilon_m\times(0,T))}, \\
\| \nabla_y D^\varepsilon\varphi \|_{L^2(Q_T\times Y_m)^n} &= \varepsilon \ \| D^\varepsilon \nabla_x\varphi \|_{L^2(\Omega^\varepsilon_m\times(0,T))^n}, \\
(D^\varepsilon\varphi,\psi)_{L^2(Q_T\times Y)} &= (\varphi, D^\varepsilon\psi)_{L^2(Q_T\times Y)}.
\end{aligned}$$

COROLLARY 4.8. *With the above notation, we have*

$$\| D^\varepsilon s^\varepsilon \|_{L^2(Q_T;\ L^2_{\mathrm{per}}(Y_m))} \leq C,$$
$$\| D^\varepsilon z^\varepsilon \|_{L^2(Q_T;\ L^2_{\mathrm{per}}(Y_m))} \leq C.$$

Having established the link between the two-scale convergence and the weak convergence of the periodically modulated sequence $\{s^\varepsilon\}$, our strategy is simple; we find an equation for $D^\varepsilon s^\varepsilon$ and pass to the limit.

Now from assumptions (A11) on the geometry of the cell, we may prove the following proposition.

PROPOSITION 4.9. *Let* $\varepsilon > 0$ *and let* $D^\varepsilon s^\varepsilon$ *be defined by* (4.26). *Then* $D^\varepsilon s^\varepsilon = \overline{s}^\varepsilon$ *satisfies the equation*

$$\varphi^0(y)\partial_t(D^\varepsilon s^\varepsilon) - \mathrm{div}_y\{k(y)\nabla_y \mathcal{A}(\overline{s}^\varepsilon)\} = \tag{4.29}$$
$$\varepsilon \mathrm{div}_y\{k(y)b(\overline{s}^\varepsilon)g\} - \varepsilon^{-1}\mathrm{div}_y\{d(\overline{s}^\varepsilon)D^\varepsilon v^\varepsilon\} \ \ \textit{in}\ \ L^2(0,T;H^{-1}(Y_m))\ \ \textit{for (a.e.)}\ x \in \Omega^\varepsilon_m.$$

Proof. Let $\xi \in H^{-1}(0,T\ ;\ L^2(Y_m)) \cap L^2(0,T\ ;\ C^\infty_0(Y_m))$. Then we define

$$(D^\varepsilon \xi)^\tau(x,z,t) = \begin{cases} \xi\left(\frac{z-c^\varepsilon(x)}{\varepsilon}, t\right) & \text{for } z \in \varepsilon Y_m + c^\varepsilon(x), \\ 0 & \text{otherwise.} \end{cases} \tag{4.30}$$

Let $\xi_k(y,t) = \xi(y+k,t), k \in \mathbb{Z}^n$. Obviously, ξ_k is defined on $Y_m + k$ and $\xi_k \in H^1(0,T\ ;\ L^2(Y_m+k)) \cap L^2(0,T\ ;\ C^\infty_0(Y_m+k))$. □

Now we plug $(D^\varepsilon \xi_k)^\tau(x,x,t)$ into (2.33) as a test function. Since supp $(D^\varepsilon \xi_k)^\tau \subset \varepsilon(Y_m+k) \times (0,T)$ and the components of Ω^ε_m are strictly separated, we get

$$\int_0^T \int_{\varepsilon(Y_m+k)} \varphi^\varepsilon s^\varepsilon \partial_t((D^\varepsilon \xi_k)^\tau) - \int_0^T \int_{\varepsilon(Y_m+k)} \varepsilon^2 k^\varepsilon a(s^\varepsilon)\nabla s^\varepsilon \nabla (D^\varepsilon \xi_k)^\tau$$
$$- \int_0^T \int_{\varepsilon(Y_m+k)} \varepsilon^2 b(s^\varepsilon) k^\varepsilon g \nabla (D^\varepsilon \xi_k)^\tau + \int_0^T \int_{\varepsilon(Y_m+k)} d(s^\varepsilon) v^\varepsilon \nabla (D^\varepsilon \xi_k)^\tau$$
$$= \int_{\varepsilon(Y_m+k)} \varphi^\varepsilon s^\varepsilon(\cdot,T)(D^\varepsilon \xi_k)^\tau(\cdot,T)dx - \int_{\varepsilon(Y_m+k)} \varphi^\varepsilon s^0_w (D^\varepsilon \xi_k)^\tau(\cdot,0)dx.$$

Moreover, for $x \in \varepsilon(Y_m+k)$, $c^\varepsilon(x) = \varepsilon k$, and the change of variables $y = \varepsilon^{-1}(x - c^\varepsilon(x))$ gives

$$\int_0^T \int_{Y_m} \varphi^0(y)(D^\varepsilon s^\varepsilon)(x,y,t)\partial_t \xi_k(y,t)dy\,dt \tag{4.31}$$
$$- \int_0^T \int_{Y_m} k(y)a(D^\varepsilon s^\varepsilon)\nabla_y(D^\varepsilon s^\varepsilon)\nabla_y \xi_k - \int_0^T \int_{Y_m} \varepsilon b(D^\varepsilon s^\varepsilon)k(y)g\nabla_y \xi_k$$
$$+ \int_0^T \int_{Y_m} \varepsilon^{-1} d(D^\varepsilon s^\varepsilon) D^\varepsilon v^\varepsilon \nabla_y \xi_k$$
$$= \int_{Y_m} \varphi^0(y)(D^\varepsilon s^\varepsilon)(\cdot,T)\xi_k(\cdot,T) - \int_{Y_m} \varphi^0(y) D^\varepsilon s^0_w \xi_k(\cdot,0)dy$$

for (a.e.) $x \in \varepsilon(Y_m+k)$. (4.29) is a simple consequence of (4.31). □

Remark. It should be noticed that

$$D^\varepsilon Z^\varepsilon = D^\varepsilon z^\varepsilon \ \text{ in } \ H^{1/2}(\partial Y_m) \quad \text{for (a.e.)} \quad (x,t) \in \Omega^\varepsilon_m \times (0,T), \tag{4.32}$$

where $D^\varepsilon z^\varepsilon = \int_0^{D^\varepsilon s^\varepsilon} \sqrt{a(\eta)}d\eta$.

Having obtained the equations for $D^\varepsilon s^\varepsilon$ we are now able to establish the a priori estimates for $D^\varepsilon z^\varepsilon$ and $D^\varepsilon s^\varepsilon$, analogous to those obtained for Z^ε and S^ε for all $k \in \mathbb{Z}^n$. Then we could pass to the limit for fixed k and, finally, by a density argument the limit equations for s would have been obtained. However, the limit equation for s has already been obtained in Arbogast, Douglas, and Hornung [6] by formal asymptotic expansion, and since it is decoupled from the pressure equation, we prefer comparing the formal limit equation with the equation for the $D^\varepsilon s^\varepsilon$.

We consider the following problem. Find $s^* \in L^2(Q_T\ ;\ L^2_{\text{per}}(Y_m)) \cap L^\infty(Q_T \times Y_m)$, $0 \le s^* \le 1$ (a.e.) on $Q_T \times Y_m$, $a(s^*)\nabla s^* \in L^2(Q_T\ ;\ L^2_{\text{per}}(Y_m))^n$, $\partial_t s^* \in L^2(Q_T; H^{-1}_{\text{per}}(Y_m))$ such that

$$\tag{4.33} \begin{gathered} \int_0^t\int_{Y_m} \varphi^0(y)s^*(x,y,\tau)\partial_t\psi - \int_0^t\int_{Y_m} k(y)a(s^*)\nabla_y s^*\nabla_y\psi \\ = \int_{Y_m} \varphi^0(y)s^*(x,y,t)\psi(y,t)dy - \int_{Y_m} \varphi^0(y)s^0_w(x)\psi(y,0)dy \\ \text{(a.e)}\ (x,t) \in Q_T \quad \forall\psi \in H^1(Q_T \times Y_m), \end{gathered}$$

$$\tag{4.34} z^* = \int_0^{s^*} \sqrt{a(\eta)}d\eta = Z = \int_0^S \sqrt{a(\eta)}d\eta \text{ in } L^2(0,T\ ;\ H^{1/2}(\partial Y_m))$$

for (a.e.) $x \in \Omega$.

For fixed $Z \in L^2(0,T;\ H^1(\Omega))$, $S \in L^\infty(Q_T)$, $0 \le S \le 1$ (a.e.) on Q_T and $S^0_w \in L^\infty(\Omega)$, $0 \le S^0_w \le 1$ (a.e.) on Ω, the classical theory (see, e.g., Lions [22]) gives existence of a unique solution s^* for (4.33)–(4.34).

Our final step is to compare the problems (4.33)–(4.34) and (4.31)–(4.32). We have the following result.

PROPOSITION 4.10. *Let $D^\varepsilon s^\varepsilon$ be defined by* (4.26), *and let s^* be a solution for* (4.33)–(4.34). *Then we have*

$$\tag{4.35} D^\varepsilon s^\varepsilon \longrightarrow s^* \quad in \quad L^2(\Omega \times Y_m \times (0,T))$$

and

$$s^* = s \quad (a.e.)\ on \quad \Omega \times Y_m \times (0,T),$$

where s is defined by assumptions (A11)–(A12).

Proof. Since a can vanish for $D^\varepsilon s^\varepsilon = 0$ or 1, we choose the test function by a change of pivot. Therefore, we introduce w^ε by

$$\tag{4.36} \begin{cases} -\text{div}_y\{k(y)\nabla_y w^\varepsilon\} &= \varphi^0(y)\{D^\varepsilon s^\varepsilon - s^*\} \text{ in } Y_m, \\ w^\varepsilon &= 0 \text{ on } \partial Y_m \end{cases}$$

for (a.e.) $(x,t) \in Q_T$.

Obviously, $w^\varepsilon \in H^1(Q_T \times Y_m)$,

$$\|\ \nabla_y w^\varepsilon\ \|_{L^2(Q_T;L^2(Y_m))} \le C \ \|\ D^\varepsilon s^\varepsilon - s^*\ \|_{L^2(Q_T\times Y_m)},$$

and

$$\tag{4.37} \begin{cases} -\text{div}_y\{k\nabla_y w^\varepsilon(x,y,0)\} &= \varphi^0(y)\{D^\varepsilon s^0_w - s^0_w\}, \\ w^\varepsilon(x,\cdot,0) &= 0 \text{ on } \partial Y_m. \end{cases}$$

Now we introduce the function ζ^ε by $\zeta^\varepsilon = D^\varepsilon s^\varepsilon - s^*$ and $\mathcal{A}$ by $\mathcal{A}(s) = \int_0^s a(\eta)\ d\eta$.

Obviously, ζ^ε satisfies the variational equation

$$\begin{aligned}(4.38)\quad &\int_0^T\int_{Y_m}\int_\Omega \varphi^0(y)\zeta^\varepsilon\partial_t\Psi - \int_{Q_T}\int_{Y_m} k(y)\nabla_y\{\mathcal{A}(D^\varepsilon s^\varepsilon)-\mathcal{A}(s^*)\}\nabla_y\Psi\\ &-\int_\Omega\int_{Y_m}\varphi^0(y)\zeta^\varepsilon(\cdot,T)\Psi\,(\cdot,T)+\int_\Omega\int_{Y_m}\varphi^0(y)\{D^\varepsilon s^0_w - s^0_w\}\Psi(\cdot,0)\\ &=\varepsilon\int_{Q_T}\int_{Y_m}k(y)b(D^\varepsilon s^\varepsilon)g\nabla_y\Psi - \varepsilon^{-1}\int_{Q_T}\int_{Y_m}d(D^\varepsilon s^\varepsilon)(D^\varepsilon v^\varepsilon)\nabla_y\Psi\\ &\qquad\forall\Psi\in H^1(Q_T\times Y_m),\quad \Psi=0\quad\text{on}\quad\partial Y_m\end{aligned}$$

and the boundary condition

$$(4.39)\quad \int_0^{\zeta^\varepsilon}\sqrt{a(\eta)}d\eta=\int_{S(x,t)}^{D^\varepsilon S^\varepsilon}\sqrt{a(\eta)}d\eta\ \text{ in } L^2(0,T\ ;\ H^{1/2}(\partial Y_m))\ (\text{a.e.})\,x\in\Omega.$$

To estimate the L^2-norm of ζ^ε we choose $\Psi=w^\varepsilon$ as the test function. We have

$$\begin{aligned}(4.40)\quad &\int_{Q_T}\int_{Y_m}\varphi^0(y)\zeta^\varepsilon\partial_t w^\varepsilon-\int_\Omega\int_{Y_m}\varphi^0(y)\zeta^\varepsilon(\cdot,T)w^\varepsilon(\cdot,T)\\ &+\int_\Omega\int_{Y_m}\varphi^0(y)\{D^\varepsilon s^0_w-s^0_w\}w^\varepsilon(\cdot,0)-\int_{Q_T}\int_{Y_m}k(y)\nabla_y\{\mathcal{A}(D^\varepsilon s^\varepsilon)-\mathcal{A}(s^*)\}\nabla_y w^\varepsilon\\ &=\varepsilon\int_{Q_T}\int_{Y_m}k(y)b(D^\varepsilon s^\varepsilon)g\nabla_y w^\varepsilon-\varepsilon^{-1}\int_{Q_T}\int_{\Omega_\varepsilon}d(D^\varepsilon s^\varepsilon)(D^\varepsilon v^\varepsilon)\nabla_y w^\varepsilon.\end{aligned}$$

Using the uniform boundedness of $\nabla_y w^\varepsilon$ in $L^2(Q_T\times Y_m)$ we find out that the right-hand side is bounded by $C\varepsilon$.

Furthermore,

$$\begin{aligned}\int_{Q_T}\int_{Y_m}\varphi^0(y)\partial_t w^\varepsilon\zeta^\varepsilon&=\int_{Q_T}\int_{Y_m}k\nabla_y w^\varepsilon\nabla_y\partial_t w^\varepsilon\\ &=\frac12\int_\Omega\int_{Y_m}k|\nabla_y w^\varepsilon(\cdot,T)|^2-\frac12\int_\Omega\int_{Y_m}k|\nabla_y w^\varepsilon(\cdot,0)|^2,\end{aligned}$$

$$\int_\Omega\int_{Y_m}\varphi(y)\zeta^\varepsilon(\cdot,T)w^\varepsilon(\cdot,T)=\int_\Omega\int_{Y_m}k|\nabla_y w^\varepsilon(\cdot,T)|^2,$$

and

$$\int_\Omega\int_{Y_m}\varphi(y)\zeta^\varepsilon(\cdot,0)w^\varepsilon(\cdot,0)=\int_\Omega\int_{Y_m}k|\nabla_y w^\varepsilon(\cdot,0)|^2.\qquad\square$$

Therefore, we write (4.40) in the form

$$\begin{aligned}&\left|\int_{Q_T}\int_{Y_m}k(y)\nabla_y\{\mathcal{A}(D^\varepsilon s^\varepsilon)-\mathcal{A}(s^*)\}\nabla_y w^\varepsilon\right|\\ &\le C\varepsilon+\frac12\int_\Omega\int_{Y_m}|\nabla_y w^\varepsilon(\cdot,0)|^2\le C\left\{\varepsilon+\parallel D^\varepsilon s^0_w-s^0_w\parallel_{L^2(\Omega\times Y_m)}\right\}.\end{aligned}$$

It remains to find an estimate from below for the diffusion term. We have

$$\begin{aligned}0\le&\int_{Q_T}\int_{Y_m}k(y)\nabla_y\{\mathcal{A}(D^\varepsilon s^\varepsilon)-\mathcal{A}(s^*)\}\nabla_y w^\varepsilon=\int_{Q_T}\int_{\partial Y_m}\{\mathcal{A}(D^\varepsilon s^\varepsilon)\\ &-\mathcal{A}(S)\}k(y)\nabla_y w^\varepsilon\cdot\nu-\int_{Q_T}\int_{Y_m}\{\mathcal{A}(D^\varepsilon s^\varepsilon)-\mathcal{A}(S)\}\mathrm{div}_y\{k(y)\nabla_y w^\varepsilon\}\\ &=\int_{Q_T}\int_{Y_m}\nabla_y\mathcal{A}(\Pi_\varepsilon D^\varepsilon S^\varepsilon)k(y)\nabla_y w^\varepsilon+\int_{Q_T}\int_{Y_m}\{\mathcal{A}(D^\varepsilon s^\varepsilon)-\mathcal{A}(s^*)\}\varphi^0(y)\zeta^\varepsilon.\end{aligned}$$

Therefore,

$$\begin{aligned}
&\int_{Q_T}\int_{Y_m} \{\mathcal{A}(D^\varepsilon s^\varepsilon) - \mathcal{A}(s^*)\}\{D^\varepsilon s^\varepsilon - s^*\} \\
&\leq C\varepsilon + C \parallel D^\varepsilon s_w^0 - s_w^0 \parallel_{L^2(\Omega\times Y_m)} + C \parallel \nabla_y \mathcal{A}(\Pi_\varepsilon D^\varepsilon S^\varepsilon) \parallel_{L^2(Q_T\times Y_m)^n} \\
&= C\varepsilon + C \parallel D^\varepsilon s_w^0 - s_w^0 \parallel_{L^2(\Omega\times Y_m)} + C\varepsilon \parallel \nabla_x \mathcal{A}(\Pi_\varepsilon D^\varepsilon S^\varepsilon) \parallel_{L^2(Q_T\times Y_m)^n}
\end{aligned}$$

and consequently,

$$\{\mathcal{A}(D^\varepsilon s^\varepsilon) - \mathcal{A}(s^*)\}\{D^\varepsilon s^\varepsilon - s^*\} \to 0 \quad \text{(a.e.) in} \quad Q_T \times Y_m.$$

Hence $D^\varepsilon s^\varepsilon - s^* \to 0$ (a.e.) in $Q_T \times Y_m$ and $D^\varepsilon s^\varepsilon - s^* \to 0$ (a.e.) in $L^q(Q_T \times Y_m)$ for all $q \in [1, +\infty[$ by Lebesgue's dominated convergence theorem. □

We summarize our results in Theorem 4.11.

THEOREM 4.11 (convergence theorem). *Let the assumptions* (A1)–(A10) *be satisfied. Let* $\{P^\varepsilon, p^\varepsilon, S^\varepsilon, s^\varepsilon, V^\varepsilon, v^\varepsilon\}_{\{\varepsilon>0\}}$ *be weak solutions to the problem* (2.20)–(2.32). *Then there exists a subsequence, denoted by the same subscript, such that*

$$\Pi_\varepsilon P^\varepsilon \to P \in L^2(0,T\,; H^1(\Omega)) \quad \textit{in the two-scale sense};\tag{4.41}$$

$$\nabla(\Pi_\varepsilon P^\varepsilon) \to \nabla P + \nabla_y P_1 \quad \textit{in the two-scale sense};\tag{4.42}$$

$$\Pi_\varepsilon Z^\varepsilon = \Pi_\varepsilon \int_0^{S^\varepsilon} \sqrt{A(\eta)}d\eta \to Z \in L^2(0,T\,; H^1(\Omega)) \quad \textit{in the two-scale sense};\tag{4.43}$$

$$\nabla(\Pi_\varepsilon Z^\varepsilon) \to \nabla Z + \nabla_y Z_1 \quad \textit{in the two-scale sense}\tag{4.44}$$

with $Z_1, P_1 \in L^2(Q_T\ ;\ H^1_{\text{per}}(Y))$;

$$\overline{S}^\varepsilon = (Z^\varepsilon)^{-1}(\Pi_\varepsilon Z^\varepsilon) \to S \quad \textit{strongly in} \quad L^q(Q_T) \quad \forall q \in [1, +\infty[;\tag{4.45}$$

$$\Pi_\varepsilon Z^\varepsilon \to Z \quad \textit{strongly in} \quad L^q(Q_T) \quad \forall q \in [1, +\infty[;\tag{4.46}$$

$$\widetilde{s}^\varepsilon \to s \in L^2(Q_T \times Y) \quad \textit{in the two-scale sense};\tag{4.47}$$

$$D^\varepsilon s^\varepsilon \to s \quad \textit{strongly in} \quad L^q(Q_T \times Y_m),\tag{4.48}$$

where the extension operator Π_ε *is defined by* (3.10)–(3.12), $\widetilde{s}^\varepsilon$ *is equal to* s^ε *in* $\Omega_m^\varepsilon \times]0,T[$ *and to* S^ε *in* $\Omega_f^\varepsilon \times]0,T[$, $D^\varepsilon s^\varepsilon$ *is defined by* (4.26), *and* $\{P, S, s, V\}$ *are solutions to the nonlinear system*

$$\left\{\begin{aligned}
&\frac{1}{|Y|}\int_Y V dy = \overline{V} = -K^{*H}\Lambda(S)[\nabla_x P - E(S)g] \quad \textit{in} \quad Q_T, \\
&\qquad\qquad -\text{div}_x \overline{V} = 0 \quad \textit{in} \quad Q_T, \\
&\overline{V}\cdot\nu = 0 \quad \textit{on} \quad \Gamma^2 \times (0,T), \qquad P = P_T^D = \tfrac{P_w^D + P_o^D}{2} \quad \textit{on} \quad \Gamma^1 \times (0,T);
\end{aligned}\right.\tag{4.49}$$

$$\left\{\begin{aligned}
&S(x,0) = S_w^0(x) \quad \textit{(a.e.) in} \quad \Omega, \qquad S = 0 \quad \textit{(a.e.) on} \quad \Gamma^1 \times (0,T), \\
&\frac{|Y_f|}{|Y|}\phi^*\partial_t S - \text{div}_x\left\{K^{*H}(A(S)\nabla_x S + B(S)g) - D(S)\overline{V}\right\} \\
&\qquad = -\frac{1}{|Y|}\int_{\partial Y_m} k(y)a(s)\nabla_y s\cdot\nu d\eta \quad \textit{in} \quad Q_T, \\
&K^{*H}(A(S)\nabla_x S + B(S)g)\cdot\nu = g_w \quad \textit{on} \quad \Gamma^2 \times (0,T)
\end{aligned}\right.\tag{4.50}$$

with K^{*H} *defined by* (4.18).

For (a.e.) $x \in \Omega$, *the right-hand side of* (4.50) *is defined from*

$$(4.51)\quad \begin{cases} z = \int_0^s \sqrt{a(\eta)}d\eta = Z = \int_0^S \sqrt{A(\eta)}d\eta \quad \textit{on} \quad \partial Y_m, \\ \varphi^0(y)\partial_t s - \operatorname{div}_y\{k(y)a(s)\nabla_y s\} = 0 \quad \textit{in} \quad Q_T \times Y_m, \\ s(x,y,0) = s_w^0(x) \quad \textit{in} \quad \Omega \times Y_m. \end{cases}$$

From (4.49)–(4.51) above, due to the strong convergence of S^ε and hence of all functions of S^ε, doing the inverse of what we have done in §2, we conclude the following:

- There is a macroscopic fracture system driven by equations in all $\Omega \times (0,T)$ similar to (2.1), (2.2) and with an effective absolute rock permeability K^{*H} given by (4.18) and an effective porosity $\frac{|Y_f|}{|Y|}\Phi^*$ but with an additional right-hand side source-like term $\frac{1}{|Y|}\int_{Y_m} \varphi^0(y)\partial_t s dy = \frac{1}{|Y|}\int_{\partial Y_m} k(y)a(s)\nabla_y s \cdot \eta d\eta$.
- For each $x \in \Omega$ there is a matrix block, the flow in which is described by (4.51) and produces the source-like term.

Remark. In the definition (4.18) of K^{*H}, the effective permeability, the term $I + \frac{1}{|Y_f|}\int_{Y_f} \Theta(y)dy$ plays the role of a tortuosity factor like that defined in [14] or [20].

REFERENCES

[1] E. Acerbi, V. Chiadò piat, G. Dal Maso, and D. Percivale, *An extension theorem from connected sets, and homogenization in general periodic domains*, Nonlinear Anal., 18 (1992), pp. 481–496.

[2] I. Aganović and A. Mikelić, *Homogenization of nonstationary flow of a two constituant mixture through a porous medium*, Asymptotic Anal., 6 (1992), pp. 173–189.

[3] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[4] H. W. Alt and E. DiBenedetto, *Nonsteady flow of water and oil through inhomogeneous porous media*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 12 (1985), pp. 335–392.

[5] S. N. Antontsev, A. V. Kazhikhov, and V. N. Monakhov, *Boundary Value Problems in Mechanics of Nonhomogeneous Fluids*, Elsevier, Amsterdam, 1990.

[6] T. Arbogast, J. Douglas, U. Hornung, *Derivation of the double porosity model of single phase flow via homogenization theory*, SIAM J. Math. Anal., 21 (1990), pp. 823–836.

[7] ———, *Modelling of naturally fractured reservoirs by formal homogenization techniques*, in Frontiers in Pure and Applied Mathematics, R. Dautray, ed., Elsevier, Amsterdam, 1991, pp. 1–19.

[8] G. I. Barenblatt, I. P. Zheltov, I. N. Kochina, *Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks*, Prikl. Mat. Meh., 24 (1960), pp. 852–864.

[9] G. I. Barenblatt and A. A. Gilman, *Mathematical model of non equilibrium countercurrent capillary inhibition*, Inzh.-Fiz. Zh., 52 (1987), pp. 456–461 (in Russian).

[10] G. I. Barenblatt, V. M. Entov, and V. M. Rizhik, *Theory of Fluid Flows Through Natural Rocks*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[11] A. Bourgeat, *Homogenized behavior of diphasic flow in naturally fissured reservoir with uniform fractures*, Comput. Methods Appl. Mech. Engrg., 47 (1984), pp. 205–217.

[12] A. Bourgeat, S. Kozlov, and A. Mikelić, *Effective equations of two-phase in random media*, Calculus of Variations and Partial Differential Equations, 3 (1995), pp. 385–406.

[13] A. Bourgeat, A. Mikelić, and S. Wright, *Stochastic two-scale convergence in the mean and applications*, J. Reine Angew. Math., 456 (1994), pp. 19–51.

[14] P. C. Carman, J. Agricultural Sci., 29 (1939), p. 262.

[15] G. Chavent and J. Jaffre, *Mathematical Models and Finite Elements for Reservoir Simulation*, Elsevier Science Publishers, Amsterdam, 1986.

[16] J. Douglas and T. Arbogast, *Dual porosity models for flow in naturally fractured reservoirs*, in Dynamics of Fluids in Hierarchical Porous Media, J. H. Cushman, ed., Academic Press, New York, 1990, pp. 177–220.

[17] R. Ewing, *The mathematics of reservoir simulation*, in Frontiers in Applied Mathematics, Society for Indutrial and Applied Mathematics, Philadelphia, PA, 1983.

[18] R. A. Greenkorn, *Flow Phenomena in Porous Media*, M. Dekker, New York, 1983.

[19] H. Kazemi, L. S. Merril, Jr., K. L. Porterfield, and P. R. Zeman, *Numerical simulation of water-oil flow in naturally fractured reservoirs*, Soc. Pet. Engrg. J., Trans. AIME, 267 (1976), pp. 317–326.

[20] J. S. Kozeni, Ber. Wiener Akad. Abt. IIa, 136 (1967), p. 271.

[21] D. Kröner and S. Luckhaus, *Flow of oil and water in a porous medium*, J. Differential Equations, 55 (1984), pp. 276–288.

[22] J. L. Lions, *Quelques méthodes de résolution des problemes aux limites non linéaires*, Dunod, Paris, 1969.

[23] C. M. Marle, *Multiphase Flow in Porous Media*, Editions Technip, Paris, 1981.

[24] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–628.

[25] M. B. Panfilov, *Mean mode of porous flow in highly inhomogeneous media*, Dokl. Akad. Nauk SSSR, 311 (1990), pp. 313–317 (in Russian); Soviet Phys. Dokl., 35 (1990), pp. 225–227 (in English).

[26] D. W. Peaceman, *Fundamentals of Numerical Reservoir Simulation*, Elsevier, New York, 1977.

[27] ———, *Convection in fractured reservoir: The effect of matrix fissure transfer on the instability of a density inversion in a vertical fissure*, Soc. Petroleum Engrg. J. Trans. AIME, 261 (1976), pp. 269–280.

[28] L. H. Reiss, *The Reservoir Engineering Aspects of Fractured Formations*, Editions Technip, Paris, 1980.

[29] L. Tartar, *H-measures, a new approach for studying homogenization, oscillations and concentration effects in partial differential equation*, Proc. Royal Soc. Edinburgh Sect. A, 115 (1990), pp. 193–230.

[30] C. Vogt, *A homogenization theorem leading to a Volterra integro-differential equation for permeation chromotography*, Preprint 155, SFB 123, University of Heidelberg, Heidelberg, Germany, 1982.

[31] J. E. Warren and P. J. Root, *The behavior of naturally fractured reservoirs*, Soc. Petroleum Engrg. J. Trans. AIME, 228 (1963), pp. 245–255.

SIAM J. MATH. ANAL.
Vol. 27, No. 6, pp. 1544–1558, November 1996

003

TRAVELING WAVES AS LIMITS OF SOLUTIONS ON BOUNDED DOMAINS*

GIORGIO FUSCO[†], JACK K. HALE[‡], AND JIANPING XUN[§]

Abstract. This paper is concerned with the asymptotic behavior as $\epsilon \to 0$ of solutions of the reaction-diffusion equation $u_t = \epsilon^2 u_{xx} - (u+a)(u^2-1)$ defined in $(-1,1)$ with Neumann boundary conditions. For $a = 0$, this equation has a monotone equilibrium solution u^ϵ with the property that $u^\epsilon(x) \to -1$ (resp. $+1$) on $[-1,0)$ (resp. $(0,1]$) as $\epsilon \to 0$; that is, the solution has a sharp transition layer if $a = 0$. Also, it is known that u^ϵ has a one-dimensional unstable manifold $\mathcal{M}(u^\epsilon)$. Solutions near $\mathcal{M}(u^\epsilon)$ decrease exponentially to $\mathcal{M}(u^\epsilon)$ and move with a speed $O(e^{-c/\epsilon})$ along $\mathcal{M}(u^\epsilon)$.

This paper considers the case where a is small and fixed. For each fixed ϵ, $a \neq 0$, small, there is an equilibrium solution $u^{\epsilon a}$ with unstable manifold of dimension one, but $u^{\epsilon a}$ approaches either the function 1 or -1 as $\epsilon \to 0$; that is, there is no monotone equilibrium solution with a sharp transition layer. If we rescale x to ϵx and consider the rescaled equation on $(-\infty, \infty)$, then there is a unique (except for translation) monotone traveling-wave solution on $(-\infty,\infty)$ with wave speed $-\sqrt{2}a$. Using a geometric approach, we prove that there are positive constants ϵ_0 and a_0 such that, for $0 < \epsilon < \epsilon_0$ and $|a| < a_0$, solutions of the rescaled equations on $(-\frac{1}{\epsilon}, \frac{1}{\epsilon})$ in a neighborhood of size $C\sqrt{a_0}$ of a monotone traveling-wave solution decrease exponentially fast before they enter a neighborhood of size $O(\epsilon^k)$ of such a solution, where k can be any positive integer. Along the traveling-wave direction, solutions move with the traveling-wave speed plus an error term $O(\epsilon^k)$. It also is proved that the L^∞-norm between the solution and a translation of the traveling wave is of order $O(\epsilon^k)$ for $C_1 k \log \frac{1}{\epsilon} < t < \frac{C_2}{\epsilon}$.

Key words. transition layers, phase separation, unstable invariant manifold, traveling-wave solutions

AMS subject classifications. 35B30, 35B25, 35K55

1. Introduction.

It is generally accepted folklore that traveling-wave solutions of parabolic partial differential equations are representative of typical behavior of solutions of the same partial differential equation on a large unbounded domain. The intent of this paper is to make this rigorous for a scalar reaction-diffusion equation in one space variable. More specifically, we consider the equation

$$u_t = \epsilon^2 u_{xx} - f_a(u), \quad x \in (-1,1), \tag{1}$$

with homogeneous Neumann boundary conditions

$$u_x = 0, \quad x = \pm 1. \tag{2}$$

In this equation, $\epsilon > 0$ is a small parameter, a is a small parameter independent of ϵ, and

$$f_a(u) =: (u+a)(u^2-1) =: f_0(u) + ag(u), \tag{3}$$

*Received by the editors October 11, 1994; accepted for publication (in revised form) June 21, 1995.

[†] 2 Via Parioli, 00197 Rome, Italy (fusco%40676.hepnet@csa4.lbl.gov). The research of this author was partially supported by the CNR.

[‡] Center for Dynamical Systems and Nonlinear Studies, Georgia Institute of Technology, Atlanta, GA 30332 (hale@math.gatech.edu). The research of this author was partially supported by NSF grant DMS-9005420 and DARPA grant 70NANB8H0860.

[§] 400 East Riverside Drive, Apartment 113, St. George, UT 84770 (pingx@strata3d.com). The research of this author was partially supported by the CDSNS, Georgia Institute of Technology.

$$f_0(u) =: u(u^2-1), \qquad g(u) =: u^2-1. \tag{4}$$

Equation (1) is a gradient system corresponding to the Liapunov functional

$$J(u) =: \int_{-1}^{1} \left(\frac{\epsilon^2}{2} u_x^2 + F(u) \right) dx,$$

where $F(u)$ is a function such that $f_a(u) = F'(u)$. Equation (1) is perhaps the simplest mathematical model for the dynamical phase transition. In this context, u is an order–disorder parameter which is related to the microscopic structure of the matter in such a way that u near -1 corresponds to one of the two phases (solid) and u near 1 corresponds to the other phase (liquid). The parameter a corresponds to the temperature. For $a < 0$, the constant solution $u = -1$ (solid) is the unique global minimizer of $J(u)$. If $a = 0$, then both $u = \mp 1$ minimize $J(u)$ and this corresponds to the situation where two phases of the same substance can coexist at the transition temperature. For $a > 0$, $u = 1$ (liquid) is the unique global minimizer of $J(u)$.

It is known that equation (1) generates a dissipative semiflow in several function spaces and that it possesses a global attractor $\mathcal{A}_\epsilon$ (see [H]).

If $a = 0$, the number of fixed points (stationary solutions of (1)) increases without bounds as $\epsilon \to 0$. When the equilibria are hyperbolic, the attractor $\mathcal{A}_\epsilon$ is the set of equilibria together with their unstable manifolds. In the view of [F-H] (or [C-P]), the solution quickly "*lands*" near $\mathcal{A}_\epsilon$ and there it is strongly attracted toward an unstable manifold; it then enters a slow stage during which it moves along the unstable manifold and, when it gets close to the boundary of this unstable manifold, the motion is quick, to be followed again by slow movement along another unstable manifold, etc. This phenomenon has been observed numerically and studied rigorously by Carr and Pego [C-P], Fusco and Hale [F-H], Fusco [F], and Bronsard and Kohn [B-K].

For $a = 0$, there is an $\epsilon_1 > 0$ such that, for $0 < \epsilon < \epsilon_1$, there is an equilibrium solution u^ϵ of (1) which is monotone increasing on $(-1, 1)$ and, for ϵ very small, is in an ϵ-neighborhood of $+1$ in $(\epsilon \log \frac{1}{\epsilon}, 1)$ and in an ϵ-neighborhood of -1 in $(-1, -\epsilon \log \frac{1}{\epsilon})$.

It is known [C-P], [F-H], [F] that, for $\epsilon \ll 1$, the stationary solution u^ϵ is hyperbolic and unstable with the unstable manifold $\mathcal{M}^\epsilon$ having dimension one. Moreover, as $\epsilon \to 0$, any compact part of $\mathcal{M}^\epsilon$ approaches the manifold defined by the translates of the standing waves of the equation

$$u_t = \epsilon^2 u_{xx} - f_0(u), \quad x \in (-\infty, \infty), \tag{5}$$

and the flow on $\mathcal{M}^\epsilon$ is extremely slow (speed of $O(e^{-c/\epsilon})$). More precisely, the manifold $\mathcal{M}^\epsilon$ approaches the manifold $\mathcal{M}^0$ defined by the map

$$h \to U\left(\frac{\cdot - h}{\epsilon}\right),$$

where U is the unique solution of the problem

$$\begin{cases} U_{xx} - f_0(U) = 0, \\ \lim\limits_{x \to \pm\infty} U(x) = \pm 1, \\ U(0) = 0. \end{cases} \tag{6}$$

Since the equilibrium point referred to above is hyperbolic, we can find a function $a = a(\epsilon)$ approaching zero as $\epsilon \to 0$ such that there is an equilibrium point $u^{\epsilon a}$ of (1) which is monotone increasing with a profile which is close to a step function if ϵ is sufficiently small. The dynamics in this case will be similar to those when $a = 0$.

On the other hand, if we fix $a > 0$ small but independent of ϵ and let $\epsilon \to 0$, the equilibrium solution $u^{\epsilon a}$ will not develop an interior layer and approach a step function but instead will develop a boundary layer at $x = 1$ and approach -1 uniformly in any compact interval in $[-1, 1)$. On the other hand, the dimension of the unstable manifold of $u^{\epsilon a}$ is one and it is exponentially attracting. Thus we would expect that this one-dimensional manifold $\mathcal{M}^{\epsilon a}$ would converge to some one-dimensional stable object as $\epsilon \to 0$. This object should correspond to the invariant manifold $\mathcal{M}^{0a}$ of the monotone traveling wave (and its translates) for equation (5) with f_a replacing f_0. To give a complete proof of this conjecture is beyond the scope of this paper. However, we prove in the following that, if $\epsilon > 0$ is small, solutions of (1) and (2) with initial condition close in a certain sense to $\mathcal{M}^{0a}$ are attracted to a very small neighborhood of $\mathcal{M}^{0a}$ and remain near $\mathcal{M}^{0a}$ for a time of $O(\epsilon^{-1})$ drifting along $\mathcal{M}^{0a}$ with a speed which is almost exactly that of the traveling wave. We now make this more precise.

If we rescale the space variable $x \to \epsilon x$, then we obtain an equation with diffusion coefficient 1 but defined on the interval $(-\frac{1}{\epsilon}, \frac{1}{\epsilon})$:

$$u_t = u_{xx} - f_a(u), \quad x \in \left(-\frac{1}{\epsilon}, \frac{1}{\epsilon}\right), \tag{1'}$$

and

$$u_x = 0 \quad \text{for } x = \pm\frac{1}{\epsilon}. \tag{2'}$$

It is known that, for given initial data $u_0(x)$ which is less than a in $I_1 =: [-1, \xi)$ and greater than a in $I_2 =: (\xi, 1]$, where ξ is some number between -1 and 1, the solution $u(x,t)$ of (1) and (2) which begins at $t = 0$ with this initial data will be attracted to -1 in I_1 and 1 in I_2 in a short time and a sharp interface will be formed near $x = \xi$. This is the so-called phase-generation stage. In the next stage, the configuration of the solution generated in the first stage will stay almost the same but the interface will drift slowly toward the boundary (see [A-B-F], [A-Mc], [B-F1], [B-F2], [C-P], [D-S], [F-H], [B-X1], and [B-X2]).

The main interest of this paper is the limiting behavior of the solutions of (1) and (2) as $\epsilon \to 0$ for large but finite t. In other words, we are interested in the second stage of the motion described above.

Let $U(x + \sqrt{2}at)$ be the unique (except for translation) monotone traveling-wave solution of (1′) on $(-\infty, \infty)$. Our main goal in this paper is to prove the following theorem.

THEOREM. *There exist positive constants C and a_0 such that, for any integer k, there are positive constants ϵ_0, C_1, C_2, and C_3 such that, if $0 < \epsilon \leq \epsilon_0$ and $|a| \leq a_0$, then, for any solution $u^\epsilon(x,t)$ of (1′) and (2′) with initial data in a $C\sqrt{a_0}$ L^∞-neighborhood of $U(\cdot)$, there is a positive constant h_0 such that*

$$\left|u(x,t) - U(x + \sqrt{2}at - h_0)\right| \leq C_1\epsilon^{k-2}$$

for $x \in [-\frac{1}{\epsilon}, \frac{1}{\epsilon}]$, $t \in (C_2 k \log\frac{1}{\epsilon}, \frac{C_3}{\epsilon})$.

We organize the paper as follows. In §2, we construct an approximate manifold such that the traveling-wave solution of (1) is nearby. We also give some estimates about the base manifold and derive equation (1) in the new coordinate system. In §3, we give an estimate of the motion speed of a solution of (1) and (2). In §4, we prove that the approximate manifold attracts solutions which start in a small neighborhood of size independent of ϵ. Finally, in §5, we prove that solutions of (1′) with initial data in a small neighborhood of the approximate manifold approach the traveling-wave solution of (1′) as $\epsilon \to 0$.

2. The tubular coordinates and the equations of motion. We consider equation (1) together with the Neumann boundary conditions (2), assuming throughout that a in equation (1) is small but independent of ϵ. We study the limiting behavior of the solutions of equation (1) when $\epsilon \to 0$. We construct the approximate manifold by patching together solutions $\phi(x, \ell, \pm 1)$ of the equation

$$\begin{cases} \epsilon^2 \phi_{xx} - f_0(\phi) = 0, & |x| < \dfrac{\ell}{2}, \\ \phi = 0, \quad x = \pm\dfrac{\ell}{2}, \end{cases} \tag{7}$$

as was done in [C-P], where it was proved that, if $\frac{\epsilon}{\ell}$ is small, (7) has a unique positive solution $\phi(x, \ell, +1)$ which is exponentially close (in terms of ϵ) to 1 at $x = 0$ and a unique negative solution $\phi(x, \ell, -1)$ which is exponentially close to -1 at 0.

Let $\chi : R \to [0, 1]$ be a C^∞ cutoff function with $\chi(x) = 0$ for $x \leq -1$ and $\chi(x) = 1$ for $x \geq 1$.

We define

$$\Omega_{\rho_0} =: \{h \in R : \rho_0 - 1 < h < 1 - \rho_0\}, \tag{8}$$

where ρ_0 is a small but fixed positive number. For any $h \in \Omega_{\rho_0}$, let

$$\begin{aligned} u^h(x) =: &\left[1 - \chi\left(\frac{x-h}{\epsilon}\right)\right] \phi\left(x+1, 2(h+1), -1\right) \\ &+ \chi\left(\frac{x-h}{\epsilon}\right) \phi\left(x-1, 2(1-h), 1\right), \end{aligned} \tag{9}$$

and we take the approximate manifold to be

$$\mathcal{M} =: \{u^h : h \in \Omega_{\rho_0}\}. \tag{10}$$

Remark. From our definition of the approximate manifold, it is clear that $\mathcal{M}$ does not depend on a. This may look strange at first but becomes reasonable when one takes into account that the special nonlinearity f_a that we are considering has the remarkable property that, as a varies, the speed of the traveling wave changes but its profile remains unchanged and coincides with the standing wave corresponding to $a = 0$. On the other hand, as we discuss in Propositions 4 and 5 below, if ϵ is small, the function ϕ we used for constructing the approximate manifold is extremely close to the standing wave.

It is not necessary to assume that the function f_a has the particular form (3). In fact, we could have considered a more general double-well potential $F_a(u)$ depending smoothly upon a parameter a with the property that $F_0(u)$ has only three critical points—two nondegenerate minima u_0^- and u_0^+ with $F_0(u_0^-) = F_0(u_0^+)$ and a nondegenerate maximum at u_0^0. The same type of result as above is true. Of course, in this case, it is not possible to use such a simple approximate manifold. An appropriate one can be chosen following the approach in Fusco and Hale [F-H]. We have chosen to present only the simplest case.

The following notation will be used throughout the paper:

$\mathcal{L}_a(\phi) =: \epsilon^2\phi_{xx} - f_a(\phi)$;

$L_a^h(\phi) =: \epsilon^2\phi_{xx} - f_a'(u^h)\phi$; that is, L_a^h is $\mathcal{L}_a^h$ linearized at u^h;

$\tau^h(x) =: -[\chi((x-\epsilon+1)/\epsilon)(x)\chi((1-x-\epsilon)/\epsilon)(x)]u_x^h(x)$;

where $\langle\cdot,\cdot\rangle$ is the inner product in $L^2(-1,1)$. Also, we let $\|\cdot\|$ be the norm in $L^2(-1,1)$ induced by $\langle\cdot,\cdot\rangle$

It is possible to show (cf. Lemmas 7.8 and 7.9 and Proposition 3.4 in [C-P]) that the tangent vector u_h^h to $\mathcal{M}$, aside from a small term of $O(e^{-c/\epsilon})$, coincides with $-u_x^h$. On the other hand τ^h and $-u_x^h$ are different only in intervals of size 2ϵ near the boundary, where u_x^h is $O(e^{-c/\epsilon})$. Therefore, τ^h is almost tangent to $\mathcal{M}$ and unlike u_h^h satisfies Neumann boundary conditions.

Our objective is to show that, in a very precise sense, $\mathcal{M}$ can be considered as an approximate invariant manifold. To do this, we employ a new coordinate system

$$u \longrightarrow (h,v) \quad \text{meaning} \quad u = u^h + v, \quad \text{where } u^h \in \mathcal{M},$$

$$\langle v, \tau^h\rangle = 0, \qquad v_x = 0 \quad \text{at } x = 0, 1. \tag{11}$$

If we define

$$\mathcal{B}_\sigma =: \left\{ u \in L^\infty : \inf_{h\in\Omega_{\rho_0}} \|u - u^h\|_\infty < \sigma \right\}, \tag{12}$$

it is proved in [C-P] that, if $u \in \mathcal{B}_{\sigma_0}$, the coordinate representation (11) is valid for some fixed σ_0.

Now let $u(x,t)$ be a classical solution of (1) and (2) having the coordinate representation

$$u(x,t) = u^{h(t)}(x) + v(x,t). \tag{13}$$

If we differentiate the identity $\langle v(\cdot,t), \tau^{h(t)}\rangle = 0$ with respect to t and use the expression for v_t which is obtained by inserting (13) into (1), we obtain the equations

$$[\langle u_h^h, \tau^h\rangle - \langle v, \tau_h^h\rangle]h' = \langle \mathcal{L}_a(u^h + v), \tau^h\rangle, \tag{14}$$

$$v_t = \mathcal{L}_a(u) - u_h^h h', \tag{15}$$

where, for simplicity of notation, we suppress the dependence on t. Equations (14) and (15) will be of primary concern in the remainder of this paper.

In the next section, we will need some estimates about the approximate manifold that we have just constructed explicitly. Here we list some results, the proofs of which can be found in [C-P] (cf. Proposition 2.3 and Theorem 5.2).

PROPOSITION 1. *There are constants $A_0 > 0$, C, c, and $\epsilon_0 > 0$ such that, if $\epsilon < \epsilon_0$, we have, for $h \in \Omega_{\rho_0}$,*

$$\begin{aligned} A_0\epsilon^{-1/2} &\le \|u_h^h\| \le 2A_0\epsilon^{-1/2}, \\ A_0\epsilon^{-1/2} &\le \|\tau^h\| \le 2A_0\epsilon^{-1/2}, \\ A_0\epsilon^{-1} &\le \langle u_h^h, \tau^h\rangle \le 2A_0\epsilon^{-1}, \\ \|\tau_h^h\| &\le C\epsilon^{-3/2}, \\ \|\tau_h^h\|_{L^1} &\le C\epsilon^{-1}, \\ \|\mathcal{L}_0(u^h)\| &= O\left(\exp\left(-\frac{c}{\epsilon}\right)\right), \\ \|L_0^h(\tau^h)\| &= O\left(\exp\left(-\frac{c}{\epsilon}\right)\right). \end{aligned} \tag{16}$$

As was proved in [C-P] or in the appendix of [ABF], the last estimate in (16) is a consequence of the fact that τ^h is exponentially close to the first eigenfunction of L_0^h and the first eigenvalue is exponentially small.

PROPOSITION 2. *There exist constants Λ and ϵ_1 such that, for $h \in \Omega_{\rho_0}$ and $\epsilon < \epsilon_1$, the following assertions hold:*

(i) *if $v \in H^1$ and $\langle v, \tau^h\rangle = 0$, then*

$$\Lambda\epsilon\|v\|_{L^\infty}^2 \le 2\Lambda\int_{-1}^{1}(\epsilon^2 v_x^2 + v^2)dx \le 2\int_0^1(\epsilon^2 v_x^2 + f_0'(u^h)v^2)dx; \tag{17}$$

(ii) *if $v \in H^2$, $v_x = 0$ at $x = 0, 1$, and $\langle v, \tau^h\rangle = 0$, then*

$$\Lambda\int_{-1}^{1}(\epsilon^2 v_x^2 + f_0'(u^h)v^2)dx = \Lambda\langle v, -L_0^h v\rangle \le \|L_0^h v\|^2. \tag{18}$$

3. Estimate on $h'(t)$. The goal of this section is to give an estimate on $h' =: h'(t)$. Later, we will refine this estimate and show that h' approaches the speed of the monotone traveling-wave solution of (1) on $(-\infty, \infty)$. Throughout the remainder of the paper, C will designate a generic constant independent of ϵ and a.

We estimate $\langle v, \tau_h^h\rangle$ first. In the following, we take $0 < \epsilon \le \epsilon_0$ and $|a| \le a_0$ with ϵ_0 and a_0 fixed small constants and assume the following:

I. $u(x,t)$ is a classical solution of (1) and (2) such that $u(\cdot, 0) \in \mathcal{B}_{\sigma_0}$ so that the coordinate representation (13) is valid in a neighborhood of $t = 0$.

II. The function $v(x,t)$ defined in (13) satisfies

$$\|v(\cdot, t)\|_{L^\infty} < C\sqrt{a_0}.$$

The continuity of the representation (13) and the continuity of the solution of (1) and (2) with respect to the initial data imply that this inequality holds for t in some interval $[0, T)$ provided that it holds at $t = 0$. We shall see that, if ϵ_0 and a_0 are small enough, it holds until $h(t)$ leaves Ω_{ρ_0}. We will prove this by obtaining a priori estimates under this assumption and then show that our estimates imply that we can choose ϵ_0 and a_0 so that II is always satisfied.

By assumption II and Proposition 1, we obtain the inequality

$$|\langle v, \tau_h^h\rangle| \le C\|v\|_{L^\infty}\|\tau_h^h\|_{L^1} \le C\sqrt{a_0}\epsilon^{-1}. \tag{19}$$

From this estimate and Proposition 1, we see that

$$\langle u_h^h, \tau^h\rangle - \langle v, \tau_h^h\rangle > C\epsilon^{-1}. \tag{20}$$

From (19) and (20), we can see that the coefficient of h' in formula (14) is nonzero. Therefore, we can rewrite (14) as

$$h' = X(h, v), \tag{14$'$}$$

where

$$X(h, v) =: \frac{\langle \mathcal{L}_a(u^h + v), \tau^h\rangle}{\langle u_h^h, \tau^h\rangle - \langle v, \tau_h^h\rangle}. \tag{14$''$}$$

We estimate $\langle \mathcal{L}_a(u), \tau^h\rangle$ as follows:

$$\begin{aligned}\langle \mathcal{L}_a(u), \tau^h\rangle &= \langle \mathcal{L}_0(u^h), \tau^h\rangle + \langle \epsilon^2 v_{xx} - f_0'(u^h)v, \tau^h\rangle \\ &\quad - \left\langle \frac{1}{2}f_0''v^2 + ag(u^h) + ag'v, \tau^h\right\rangle \\ &=: \mathrm{I} + \mathrm{II} + \mathrm{III},\end{aligned} \tag{21}$$

where f_0'' and g' are evaluated at points between u and u^h.

By using Proposition 1, we have

$$|\mathrm{I}| = O(e^{-c/\epsilon}); \tag{22}$$

$$\begin{aligned}|\mathrm{II}| &= |\langle v, L_0^h \tau^h\rangle| \\ &= O(e^{-c/\epsilon}).\end{aligned} \tag{23}$$

On the other hand, if a_0 is sufficiently small, assumption II implies that, for $|a| \le a_0$,

$$|\mathrm{III}| \le Ca_0 \|\tau^h\|_{L_1},$$

and therefore

$$|\mathrm{III}| \le Ca_0. \tag{24}$$

In fact, the definition of τ^h implies that

$$\|\tau^h\|_{L^1} = \|u_x^h\|_{L^1} + O(e^{-c/\epsilon}) < C.$$

Combining formulas (19) through (24) yields the following proposition.

PROPOSITION 3. *If assumptions* I *and* II *hold,* $0 < \epsilon \le \epsilon_0$, *and* $|a| \le a_0$, *then there is a constant* $C > 0$ *such that the following estimate is valid:*

$$|h'| \le Ca_0\epsilon.$$

4. Estimates in the direction orthogonal to $\mathcal{M}$. In the previous section, we gave an estimate in the direction tangential to $\mathcal{M}$. In this section, we obtain an estimate in the direction orthogonal to $\mathcal{M}$. In order to do this, we need a number of estimates on the functions $\phi(x, \ell, \pm 1)$ that we used for constructing $\mathcal{M}$. We present the estimates in a sequence of propositions.

PROPOSITION 4. *For any integer k, there exist positive constants ϵ_1 and C such that, for $0 < \epsilon < \epsilon_1$, we have*

$$|\phi(x, 2\ell, -1) - \Phi(x-\ell)| \leq C\epsilon^k, \quad x \in [0, \ell],$$

$$|\phi(x, 2\ell, 1) + \Phi(x-\ell)| \leq C\epsilon^k, \quad x \in [0, \ell],$$

where $\phi(x, 2\ell, \pm 1)$ are solutions of equation (7) *with ℓ replaced by 2ℓ and $\Phi(x) = U(\frac{x}{\epsilon})$ is the standing-wave solution of*

$$\begin{cases} \epsilon^2 \Phi_{xx}(x) = f_0(\Phi(x)) & \text{for } x \in (-\infty, \infty) \\ \Phi(x) \to \pm 1 & \text{as } x \to \pm\infty \\ \Phi(0) = 0. \end{cases} \tag{25}$$

Proof: Step 1. Let δ be a fixed positive number such that

$$q^2 =: \min\{f_0'(s) : |s \pm 1| < 2\delta\} > 0. \tag{26}$$

If $\phi(x)$ is either $\phi(x, 2\ell, +1)$ or $\phi(x, 2\ell, -1)$, then ϕ depends on ϵ and ℓ through the ratio $\frac{\epsilon}{\ell} =: r$ and there is a positive number $H = H(\delta)$ such that

$$|\phi(x) - \phi(0)| < \delta \quad \text{for } x \in [0, \ell - \epsilon H). \tag{27}$$

Let $\alpha =: \alpha(r) =: F(\phi(0))$, where F is defined by $F' = f_0$, $F \geq 0$ and $F(\pm 1) = 0$. It is known and actually not hard to show that $\phi(0) + 1 = O(e^{-c/\epsilon})$ and therefore that $\alpha(r) = O(e^{-c/r})$(see [C-P]).

Let $\tilde{H} > H$ be such that

$$|\Phi(x-\ell) + 1| < \delta \quad \text{for } x \in (-\infty, \ell - \epsilon\tilde{H}] \tag{28}$$

and $\epsilon > 0$ sufficiently small. This is possible since $\Phi(-\epsilon\tilde{H})$ is ϵ independent and approaches -1 as $\tilde{H} \to \infty$. If we define $p =: \phi(\ell - \epsilon\tilde{H})$ and $\bar{p} =: \Phi(-\epsilon\tilde{H})$, then $\bar{p}$ is ϵ independent.

Using the fact that

$$\int_p^0 \frac{du}{\sqrt{2(F(u) - \alpha(r))}} = \int_{\bar{p}}^0 \frac{du}{\sqrt{2F(u)}} = \epsilon\tilde{H}$$

and $\alpha(r) > 0$, we know that $-1 < \bar{p} < p \leq \phi(x) \leq 0$ for $x \in [\ell - \epsilon\tilde{H}, \ell]$. Therefore, for $x \in [\ell - \epsilon\tilde{H}, \ell]$, we have $F(\phi(x)) \geq C > 0$ for some constant C depending only on F and $\tilde{H}$. It follows that if we let $\bar{V}(x) =: \phi(x) - \Phi(x-\ell)$, then, for $x \in [\ell - \epsilon\tilde{H}, \ell]$,

$$\begin{cases} \epsilon\bar{V}_x(x) = \sqrt{2}((F(\phi(x)) - \alpha)^{1/2} - F(\Phi(x))^{1/2}) \geq -C(\alpha + \bar{V}), \\ \bar{V}(\ell) = 0, \quad \bar{V}(x) \geq 0, \end{cases}$$

where C is a positive constant depending only on F.

By integration, we deduce that $|\Phi(x-\ell) - \phi(x)| \leq C_1\alpha(r)$ for $x \in [\ell - \epsilon\tilde{H}, \ell]$, where $C_1 = e^{C\tilde{H}}$. Therefore, using the fact that $\alpha(r)$ is exponentially small, we obtain

$$|\Phi(x-\ell) - \phi(x)| \leq C_1'\epsilon^k \quad \text{for } x \in [\ell - \epsilon\tilde{H}, \ell] \tag{29}$$

and any fixed $C_1' \ge 1$ provided that $\epsilon > 0$ is sufficiently small.

Step 2. In this step, we will prove (29) in the interval $\mathrm{I} =: [0, \ell - \tilde{H}\epsilon]$. We denote $\bar{V}(x) = \phi(x) - \Phi(x - \ell)$ as in Step 1.

Let $\tilde{v} =: \Phi(x - \ell) + 1$. Using (28) and the definition of q in (26), we can compare $\tilde{v}$ with the solution of

$$\begin{cases} \epsilon^2 y_{xx} = q^2 y & \text{for } x \in (-\infty, \ell - \epsilon\tilde{H}), \\ y(-\infty) = 0, & y(\ell - \epsilon\tilde{H}) = \delta, \end{cases}$$

to obtain

$$0 \le \tilde{v}(x) \le y(x) = \delta e^{(x - (\ell - \epsilon\tilde{H}))\frac{q}{\epsilon}}.$$

This is bounded above by $C\epsilon^k$ provided that $x \le \ell - \tilde{C}_1 \epsilon k \log \frac{1}{\epsilon}$ for ϵ sufficiently small and where $\tilde{C}_1' = \frac{2}{q}$.

Then, for $x \in \mathrm{I} = [0, \ell - \epsilon\tilde{H}]$, we have

$$\begin{aligned} \epsilon^2 \bar{V}_{xx} &= F'(\phi(x)) - F'(\Phi(x - \ell)) \\ &= F''(\xi)\bar{V}(x) \ge q^2 \bar{V}(x), \end{aligned}$$

where (27) and (28) are used.

At the endpoints of I, we have $\bar{V}(0) = \tilde{a}$ and $\bar{V}(\ell - \tilde{H}\epsilon) = \tilde{b}$. We have already recalled in Step 1 that $\phi(0) + 1 = O(e^{-c/\epsilon})$. Moreover, $\Phi(0) + 1 = O(e^{-c/\epsilon})$. It follows that, for $\epsilon > 0$ small,

$$|\tilde{a}| = |[\phi(0) + 1)] - [\Phi(-\ell) + 1]| \le C\epsilon^k.$$

On the other hand, by (29), $|\tilde{b}| \le C\epsilon^k$ for ϵ sufficiently small.

If we solve

$$\begin{cases} \epsilon^2 y'' = q^2 y & \text{for } x \in \mathrm{I}, \\ y(0) = \tilde{a}, & y(\ell - \tilde{H}\epsilon) = \tilde{b}, \end{cases}$$

and use a comparison argument, it is not hard to conclude that

$$|\Phi(x - \ell) - \phi(x)| = |\bar{V}(x)| \le |y| \le \max\left\{|\tilde{a}|, |\tilde{b}|\right\} \le C_1'\epsilon^k$$

for $x \in [0, \ell - \epsilon\tilde{H}]$. The proof follows if we adjust the constant $\tilde{C}_1$. □

PROPOSITION 5. *For any positive integer k, positive constants ϵ_1 and C as in Proposition* 4, *and any positive integer $0 < n < k$, there exists $C_1 > 0$, which may depend on n, such that, for $0 < \epsilon < \epsilon_1$, we have*

$$|\phi^{(n)}(x) - \Phi^{(n)}(x - \ell)| \le C_1 \epsilon^{k-n} \quad \textit{for } x \in [0, \ell]. \tag{30}$$

Proof. We give the proof only for $n = 1$, and the general case follows immediately from the following formula:

$$\frac{d^n}{dx^n} f(g) = \sum_{\substack{\alpha_j \ge 0 \\ \sum \alpha_j = n}} c^\alpha(g) \frac{d^{\alpha_1}}{dx^{\alpha_1}} g \frac{d^{\alpha_2}}{dx^{\alpha_2}} g \cdots \frac{d^{\alpha_n}}{dx^{\alpha_n}} g,$$

where f and g can be any smooth functions and the $c^\alpha(g)$'s in the sum are functions of g.

For $x \in [0, \ell]$, we start with the relation

$$\epsilon^2[\phi_{xx}(x) - \Phi_{xx}(x-\ell)] = F'(\phi) - F'(\Phi).$$

Integrating, we obtain

$$\epsilon^2[\phi_x(x) - \Phi_x(x-\ell)] = \int_0^x [F'(\phi) - F'(\Phi)]dx + \epsilon^2 O(\epsilon^{k-1}).$$

As a consequence,

$$\begin{aligned}\epsilon^2|\phi_x(x) - \Phi_x(x-\ell)| &\leq \epsilon C_1'' k \log\frac{1}{\epsilon}[|F''(\xi)||\phi - \Phi|] + O(\epsilon^{k+1})\\ &\leq C_1'\epsilon^{k+1}\log\frac{1}{\epsilon}\end{aligned}$$

and the proposition now follows. □

We now show that $\mathcal{M}$ is "*close*" to being invariant under the flow generated by equations (1) and (2). Actually, the following propositions show that $\mathcal{M}$ attracts nearby solutions exponentially fast into a neighborhood of $\mathcal{M}$ of size $O(\epsilon^k)$, where k can be any positive integer provided $0 < \epsilon < \epsilon^0$ and ϵ_0 is small.

To make the above statement rigorous, we denote

$$\tilde{h}(t) =: -\sqrt{2}a\epsilon t, \qquad \zeta(x,t) =: u^{\tilde{h}(t)}(x). \tag{31}$$

Note that we need to require that the t in (31) is such that $\tilde{h}(t) \in \Omega_{\rho_0}$ and therefore $0 < t < \frac{C}{\epsilon}$.

PROPOSITION 6. *There is a positive constant a_0 such that, for any positive integer k, there are positive constants ϵ_0 and C_1 such that, for $0 < \epsilon < \epsilon_0$ and $|a| \leq a_0$, we have*

$$\left|\frac{\partial}{\partial t}\zeta(x,t) = \mathcal{L}_a(\zeta(x,t))\right| \leq C_1\epsilon^k\log\frac{1}{\epsilon} \tag{32}$$

for $x \in [-1,1]$ and $0 < t < \frac{C}{\epsilon}$.

Proof. For simplicity, we denote the function $\chi(\frac{x-h}{\epsilon})$ by $\chi^\epsilon(x-h)$. We have (cf. Lemma 7.8 in [C-P])

$$u^h = (1-\chi^\epsilon(x-h))\phi(x+1, 2(h+1), -1) + \chi^\epsilon(x-h)\phi(x-1, 2(h-1), 1), \tag{33}$$

$$\begin{aligned}u_h^h = &-[(1-\chi^\epsilon(x-h))\phi_x(x+1, 2(h+1), -1)\\ &+ \chi^\epsilon(x-h)\phi_x(x-1, 2(h-1), 1)] + O(e^{-c/\epsilon}),\end{aligned} \tag{34}$$

and

$$\begin{aligned}u_{xx}^h = &(1-\chi^\epsilon(x-h))\phi_{xx}(x+1, 2(h+1), -1)\\ &+ \chi^\epsilon(x-h)\phi_{xx}(x-1, 2(h-1), 1) + O(e^{-c/\epsilon}).\end{aligned} \tag{35}$$

The traveling-wave solution $U(x,t)$ of equation (1) on $(-\infty, \infty)$ can be calculated explicitly as

$$\tilde{\Phi}(x,t) =: U\left(\frac{x}{\epsilon} + \sqrt{2}at\right) =: \frac{-1 + \exp(\sqrt{2}(\frac{x}{\epsilon} + \sqrt{2}at))}{1 + \exp(\sqrt{2}(\frac{x}{\epsilon} + \sqrt{2}at))}. \tag{36}$$

Then we have the following estimates:

$$\begin{aligned}|U(x,t)-\zeta(x,t)| &\leq (1-\chi^{\epsilon})|\tilde{\Phi}(x,t)-\phi(x+1,2(\tilde{h}+1),-1)| \\ &\quad +\chi^{\epsilon}|\tilde{\Phi}(x,t)-\phi(x-1,2(\tilde{h}-1),1)| \\ &\leq C\epsilon^{k},\end{aligned} \tag{37}$$

where we have used the fact that $\tilde{\Phi}$ is simply the translation of the standing wave of (1) with $a=0$, and therefore Proposition 4 applies. Similarly, using Propositions 4 and 5 and formulas (34) and (35), we obtain

$$\epsilon^2|\tilde{\Phi}_{xx}(x,t)-\zeta_{xx}(x,t)| \leq C\epsilon^k \log\frac{1}{\epsilon},$$

$$|\tilde{\Phi}_t(x,t)-\zeta_t(x,t)| \leq Ca_0\ \epsilon^k \log\frac{1}{\epsilon}. \tag{38}$$

The proof of the proposition follows from (37), (38), and the fact that $\tilde{\Phi}$ is a traveling-wave solution of (1). □

We need an estimate on the Fréchet derivative of the function $X(h,v)$ defined by

$$X(h,v)=\frac{\langle\mathcal{L}_a(u^h+v),\tau^h\rangle}{\langle u^h_h,\tau^h\rangle-\langle v,\tau^h_h\rangle}. \tag{39}$$

PROPOSITION 7. *If v satisfies assumption* II, *then there is a positive constant C such that*

$$\left\|\frac{\partial}{\partial v}X(h,v)\right\| \leq C\sqrt{a_0}\epsilon^{1/2}. \tag{40}$$

Proof. If μ is a smooth function satisfying boundary condition (2), then it can be verified that

$$\frac{\partial}{\partial v}X(h,v)\mu=\frac{\mathrm{I}+\mathrm{II}}{(\langle u^h_h,\tau^h\rangle-\langle v,\tau^h_h\rangle)^2}, \tag{41}$$

where

$$\mathrm{I}=:[\langle u^h_h,\tau^h\rangle-\langle v,\tau^h_h\rangle]\langle\epsilon^2\mu_{xx}-f_a'(u^h+v)\mu,\tau^h\rangle,$$
$$\mathrm{II}=:\langle\mu,\tau^h_h\rangle\langle\epsilon^2(u^h+v)_{xx}-f_a(u^h+v),\tau^h\rangle.$$

Using (20) and (41), we obtain

$$\left|\frac{\partial}{\partial v}X(h,v)\mu\right| \leq C\epsilon^2[|\mathrm{I}|+|\mathrm{II}|]. \tag{42}$$

We have

$$\begin{aligned}|\mathrm{I}| &= |\langle u^h,\tau^h\rangle-\langle v,\tau^h_h\rangle||\langle\mu,\epsilon^2\tau_{xx}-f_a'(u^h+v)\tau\rangle| \\ &\leq C\epsilon^{-1}\|\mu\|\|\epsilon^2\tau_{xx}-f_a'(u^h+v)\tau\| \\ &\leq C\epsilon^{-1}\|\mu\|\|L_0^h(\tau^h)+ag'(u^h)\tau^h-f_a''v\tau^h\| \\ &\leq C\epsilon^{-1}\|\mu\|[Ce^{-c/\epsilon}+Ca_0\epsilon^{-1/2}+C\sqrt{a_0}\epsilon^{-1/2}] \\ &\leq C\sqrt{a_0}\epsilon^{-3/2}\|\mu\|,\end{aligned} \tag{43}$$

where we have used the fact that $\|L_0^h(\tau^h)\| = O(e^{-c/\epsilon})$ by Proposition 1.

Similarly, we derive the following estimate for II:

$$\begin{aligned}|\mathrm{II}| &= |\langle \mathcal{L}_0^h(u^h), \tau^h\rangle + \langle v, L_0^h \tau^h\rangle + \langle -ag(u^h) - ag'(u^h)v \\ &\qquad + \frac{1}{2} f_a'' v^2, \tau^h\rangle||\langle \mu, \tau_h^h\rangle| \\ &\leq [Ce^{-c/\epsilon} + Ca_0\|\tau^h\|_{L^1}]\epsilon^{-3/2}\|\mu\| \\ &\leq C\sqrt{a_0}\epsilon^{-3/2}\|\mu\|.\end{aligned} \tag{44}$$

The proof of the proposition follows from (42), (43), and (44). □

PROPOSITION 8. *There exists a positive constant* a_0 *such that, for any integer* k, *there are positive constants* ϵ_0 *and* C *such that, if* $0 < \epsilon \leq \epsilon_0$ *and* v *satisfies assumption* II, *then*

$$\|v(\cdot, t)\|^2 \leq \|v(\cdot, 0)\|^2 e^{-\Lambda t} + C\epsilon^k \tag{45}$$

for $x \in [-1, 1]$ *and* $0 < t < \frac{C}{\epsilon}$.

Proof. We have

$$\begin{aligned}\frac{1}{2}\frac{\partial}{\partial t}\|v\|^2 &= \langle v, v_t\rangle \\ &= \langle v, \mathcal{L}_a(u) - u_h^h h'\rangle \\ &= \Big\langle v, [\mathcal{L}_a(u^h) - u_h^h X(h, 0)] + L_0^h v - ag'(u^h)v \\ &\qquad + \frac{1}{2} f_a'' v^2 + u_h^h [X(h, 0) - X(h, v)]\Big\rangle \\ &=: \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV} + \mathrm{V}.\end{aligned}$$

By Propositions 4 and 5 and the fact that $\tilde{\Phi}_t = \mathcal{L}_a(\tilde{\Phi}) = -\sqrt{2}at\epsilon$, it follows that $X(h, 0) = -\sqrt{2}a\epsilon + O(\epsilon^{k+1})$ for any h. Therefore, if we let $s =: h/(-\sqrt{2}a\epsilon)$, then $h(t) = \tilde{h}(s)$, where $\tilde{h}$ is defined in (31), and

$$\mathcal{L}_a(u^h) - u_h^h X(h, 0) = \mathcal{L}_a(u^{\tilde{h}}) - \frac{\partial}{\partial s} u^{\tilde{h}} + O(e^{-c/\epsilon}).$$

From this, Proposition 6, and assumption II, it follows that $|\mathrm{I}| \leq C\epsilon^k$. By Proposition 2, $-\mathrm{II} \geq \Lambda\|v\|^2$.

Using assumption II, we have $|\mathrm{III}|, |\mathrm{IV}| \leq C\sqrt{a_0}\|v\|^2$.

For V, we use Proposition 7 to deduce that

$$\begin{aligned}|\mathrm{V}| \leq \|v\|^2 \|u_h^h\| \left\|\frac{\partial}{\partial v} X(h, \bar{v})\right\| &\leq C\epsilon^{-1/2}\|v\|^2 \sqrt{a_0}\epsilon^{1/2} \\ &= C\sqrt{a_0}\|v\|^2,\end{aligned}$$

where $\bar{v}(x, t)$ is a suitable value between 0 and $v(x, t)$.

Combining all of these estimates, we obtain

$$\frac{\partial}{\partial t}\|v\|^2 + \Lambda\|v\|^2 \leq C\epsilon^k$$

provided that a_0 is chosen sufficiently small. The proof of the proposition follows immediately from Gronwall's inequality. □

In order to obtain an L^∞ estimate on v, we need the following proposition.

PROPOSITION 9. *There is a positive constant a_0 such that, for any integer k, there are positive constants ϵ_0, C, C_1, and C_2 such that, if $0 < \epsilon < \epsilon_0$ and assumption* II *is satisfied, we have*

$$\langle -L_0^h v(\cdot,t), v(\cdot,t)\rangle \le C\epsilon^k \tag{46}$$

for $x \in [-1,1]$ and $C_1 k \log\frac{1}{\epsilon} < t < \frac{C_2}{\epsilon}$.

Proof. We have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\langle -L_0^h v, v\rangle = &-\|L_0^h v\|^2 + \langle -L_0^h v, \mathcal{L}_0(u^h)\rangle \\ &+ \left\langle L_0^h v, \frac{1}{2}(f_0'' + ag'')v^2\right\rangle + \langle L_0^h v, ag(u^h)\rangle \\ &+ \langle L_0^h v, ag'(u^h)v\rangle + \langle L_0^h v, u_h^h h'\rangle \\ &+ \frac{1}{2}\langle f_0''(u^h)u_h^h h' v, v\rangle \\ =:& -\|L_0^h v\|^2 + \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV} + \mathrm{V} + \mathrm{VI}.\end{aligned} \tag{47}$$

There exist positive constants $b_i, i = 1, \ldots, 4$, whose sum is $\frac{1}{2}$ such that the six terms in (47) are estimated as follows:

$$\begin{aligned}|\mathrm{I}| &\le \|L_0^h v\|\|\mathcal{L}_0(u^h)\| \le b_1\|L_0^h v\|^2 + C\|\mathcal{L}_0(u^h)\|^2 \\ &\le b_1\|L_0^h v\|^2 + O(e^{-c/\epsilon}),\end{aligned}$$

where, in the last inequality, we have used Proposition 1.

By the second part of (17), we know that $\|v\|^2 \le C\|L_0^h v\|\|v\|$ and therefore $\|v\|^2 \le C\|L_0^h\|^2$, and it follows from this that

$$|\mathrm{II}| \le C\|v\|_\infty\|L_0^h v\|^2 \le b_2\|L_0^h v\|^2.$$

We estimate III by integration by parts. Noticing that the boundary term contribution is $O(e^{-c/\epsilon})$, we obtain using Proposition 8 that for $0 \le t \le T$, T such that assumption II is satisfied in $[0,T]$,

$$\begin{aligned}|\mathrm{III}| &\le |\langle L_0^h(ag(u^h)), v\rangle| + Ce^{-c/\epsilon} \\ &\le C|a|\|v(\cdot,t)\| \\ &\le C|a| + Ce^{-c/\epsilon}\left(\|v(\cdot,0)\|e^{-\Lambda t/2} + C\epsilon^{k/2}\right).\end{aligned}$$

Therefore, we can take ϵ small such that

$$|\mathrm{III}| \le C|a|\left(\|v(x,0)\|e^{-\Lambda t/2} + \epsilon^k\right).$$

$$|\mathrm{IV}| \le C|a|\|v\|\|L_0^h v\| \le b_3\|L_0^h v\|^2.$$

Similarly to estimating III, we have

$$|\mathrm{V}| \le C|a|\left(\|v(x,0)\|e^{-\Lambda t/2} + \epsilon^k\right).$$

By using Propositions 2 and 3 and assuming that a_0 is sufficiently small, we estimate VI as follows:

$$|\mathrm{VI}| \leq Ca_0\|v\|^2 \leq Ca_0\|L_0^h v\|^2.$$

Combining (47) and the above estimates, we have

$$\frac{d}{dt}\langle -L_0^h v, v\rangle \leq -\|L_0^h v\|^2 + C\left(\|v(x,0)\|e^{-\Lambda t/2} + \epsilon^k\right).$$

This inequality together with Proposition 2 implies that

$$\frac{d}{dt}\langle -L_0^h v, v\rangle + \Lambda\langle -L_0^h v, v\rangle \leq C\left(\|v(x,0)\|e^{-\Lambda t/2} + \epsilon^k\right),$$

and, from an application of Gronwall's lemma, we conclude that

$$\langle -L_0^h v(\cdot,t), v(\cdot,t)\rangle \leq \langle -L_0^h v(\cdot,0), v(\cdot,0)\rangle e^{-\Lambda t} + C\left(\|v(x,0)\|e^{-\Lambda t/2} + \epsilon^k\right),$$

and the proof follows. □

5. Solutions in bounded domain approach the traveling-wave solution. In this section, we prove the theorem stated in §1. After all the work in the previous sections, this is a fairly easy task.

We may choose C_1 such that $\Lambda C_1 \geq 1$; now, using inequality (45), we see that $C_1 k \log\frac{1}{\epsilon} < t < \frac{C_2}{\epsilon}$ implies that $\|v(\cdot,t)\| \leq C\epsilon^k$. From (14′) and (14″), we have

$$\begin{aligned} h'(t) &= \frac{\langle \mathcal{L}_0(u^h) + ag(u^h) + f_a'v, \tau^h\rangle + \langle v, L_a^h\tau^h\rangle}{\langle u^h, \tau^h\rangle - \langle v, \tau_h^h\rangle} \\ &= \frac{-aA + O(\epsilon^{(k-1)/2})}{\frac{A}{\epsilon\sqrt{2}} + O(\epsilon^{(k-3)/2})} \\ &= -a\epsilon\sqrt{2} + O(\epsilon^{(k+1)/2}), \end{aligned}$$

where A is a constant which can be calculated explicitly.

Therefore, by choosing a different k, we have $h'(t) = -a\epsilon\sqrt{2} + O(\epsilon^k)$ for $t_0 =: C_1 k\log\frac{1}{\epsilon} < t < \frac{C^2}{\epsilon} =: t_1$. Thus, if we let $h_0 =: h(t_0)\epsilon^{-1}$, then

$$h(t) = -a\epsilon\sqrt{2}t + h_0\epsilon + O(\epsilon^{(k-1)}). \tag{48}$$

Also from (46), it is easy to deduce that there is a positive constant C_3 such that

$$\|v(\cdot,t)\|_{L^\infty} \leq C_3\epsilon^{k-2} \tag{49}$$

for $C_1 k\log\frac{1}{\epsilon} < t < \frac{C_2}{\epsilon}$. Estimate (49) was obtained under hypothesis II; that is, as long as II is satisfied, inequality (49) is satisfied. Therefore, if we choose ϵ_0 so that $C_3\epsilon_0^{k-2} \leq C\sqrt{a_0}$, where C is the constant in II, then II will be satisfied for $0 < \epsilon \leq \epsilon_0$ and $|a| \leq a_0$.

Applying Propositions 4 and 5 and formulas (36), (48), and (49), we obtain

$$\begin{aligned} &|u(x,t) - \Phi(x + \epsilon\sqrt{2}at - h_0\epsilon)| \\ &\qquad \leq |u^h - \Phi(x + \epsilon\sqrt{2}at - h_0\epsilon)| + |v| \leq C\epsilon^{k-2} \end{aligned}$$

for $t_0 < t < t_1$, where Φ is the solution of equation (25). Therefore, we have proved the following proposition.

PROPOSITION 10. *There is a positive constant a_0 such that, for any positive integer k, there are positive constants ϵ_0, C, C_1, and C_2 such that if $0 < \epsilon < \epsilon_0$ and $|a| \leq a_0$, then*

$$|u(x,t) - \Phi(x + \epsilon\sqrt{2}at - h_0\epsilon)| \leq C\epsilon^{k-2}$$

for $x \in [-1, 1]$ and $C_1 k \log\frac{1}{\epsilon} < t < \frac{C_2}{\epsilon}$.

The theorem stated in §1 is a direct result of this proposition using the rescaling $x \longrightarrow \epsilon x$.

REFERENCES

[A-B-F] N. D. ALIKAKOS, P. W. BATES, AND G. FUSCO, *Slow motion for the Cahn–Hilliard equation in one space dimension*, J. Differential Equations, 90 (1991), pp. 81–135.

[A-Mc] N. D. ALIKAKOS AND W. R. MCKINNEY, *Remarks on the equilibrium theory for Cahn–Hilliard equation in one space dimension*, in Reaction-Diffusion Equations, K. J. Brown and A. A. Lacey, eds., Oxford Science Publications, Oxford, UK, 1990, pp. 75–93.

[B-F1] P. W. BATES AND P. C. FIFE, *Spectral comparison principles for the Cahn–Hilliard and phase-field equations, and time scales for coarsening*, Phys. D, 43 (1990), pp. 335–348.

[B-F2] ———, *The dynamics of nucleation for the Cahn–Hilliard equation*, SIAM J. Appl Math., 53 (1993), pp. 990–1008.

[B-X1] P. W. BATES AND J. XUN, *Metastable patterns for the Cahn–Hilliard equation, part* I, J. Differential Equation, 111 (1994), pp. 421–457.

[B-X2] ———, *Metastable patterns for the Cahn–Hilliard equation, part* II *Layer dynamics and slow invariant manifold*, J. Differential Equation, 117 (1995), pp. 165–216.

[B-K] B. BRONSARD AND R. V. KOHN, *On the slowness of phase boundary motion in one space dimension*, Comm. Pure Appl. Math., 43 (1990), pp. 983–998.

[C-P] J. CARR AND R. L. PEGO, *Metastable patterns in solutions of $u_t = \epsilon^2 u_{xx} - f(u)$*, Comm. Pure Appl. Math., 42 (1989), pp. 523–576.

[D-S] P. DEMOTTONI AND M. SCHATZMAN, *Development of Interfaces in R^n*, Proc. Roy. Soc. Edinburgh Sect. A, 116 (1990), pp. 207–220.

[F] G. FUSCO, *A geometric approach to the dynamics of $u_t = \epsilon^2 u_{xx} + f(u)$ for small ϵ*, Lecture Notes in Phys. 359, Springer-Verlag, Berlin, New York, 1990, pp. 53–73.

[F-H] G. FUSCO AND J. K. HALE, *Slow motion manifolds, dormant instability and singular perturbations*, Dynamics Differential Equations, 1 (1989), pp. 75–94.

[H] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, New York, Heidelberg, 1993.

SIAM J. MATH. ANAL.
Vol. 27, No. 6, pp. 1559–1596, November 1996

004

ASYMPTOTIC BEHAVIOR OF TWO INTERREACTING CHEMICALS IN A CHROMATOGRAPHY REACTOR*

DANIEL N. OSTROV†

Abstract. The chromatographic separation of two chemical species (c_1 and c_2) that transform into each other with first-order kinetics as they pass through a Langmuir isotherm reactor is governed by the following system of nonlinear hyperbolic conservation equations:

$$\begin{aligned} &\frac{\partial c_1}{\partial x} + \frac{\partial}{\partial t}\left(\frac{c_1}{1+c_1+c_2}\right) = -kc_1 + k'c_2 \\ \text{and}\quad &\frac{\partial c_2}{\partial x} + \frac{\partial}{\partial t}\left(\frac{\gamma c_2}{1+c_1+c_2}\right) = \gamma(kc_1 - k'c_2), \end{aligned}$$

$$\text{where } t \in (-\infty, \infty).$$

An analysis is presented of the two species' asymptotic behavior as they progress down a semiinfinite (i.e., $x \in [0,\infty)$) separation reactor with cyclic (periodic) entering feed concentrations. First it is shown that the method of generalized characteristics can be extended to describe the above system of equations. Then generalized characteristics are applied to show that the ω-limit set for the species concentrations is comprised of a single determined point on the curve of chemical equilibrium and that this point is approached at an exponential rate.

Key words. chromatography, hyperbolic conservation laws, generalized characteristics

AMS subject classification. 35

1. Introduction.

Chromatography is a process used by chemists and engineers to separate two chemical components in a fluid phase. The two components are passed through a tubular reactor packed with solid particles. Even though the fluid velocity may be constant, the fact that the solid will absorb different amounts of the two chemical components will cause the two concentration distributions to move down the reactor at different rates.

When a reaction allowing the two components to transform into each other occurs in the fluid, the separation caused by the chromatographic reactor will be inhibited. We will consider the most common case where the kinetics of this transformation are first order (i.e., the rate of a component's transformation is proportional to the component's concentration). A simple example of chromatography with a first-order transformation is provided by considering a separation of the cis and trans isomeric forms of an organic compound. When the temperature of the chromatography reactor provides enough activation energy to allow the isomers' chemical bonds to break and reconnect, each isomer will transform into the other at a rate proportional to its concentration.

When the solid phase consists of small, densely packed particles and the fluid velocity is relatively slow, each chemical species approaches an equilibrium between its fluid and solid phases at each point in the reactor. If the temperature within the reactor is assumed to be uniformly constant in time and space, this equilibrium allows us to express each species' solid concentration strictly as a function of both species' fluid concentrations. The function describing this relation is called an *adsorption isotherm.* The *Langmuir isotherm* is the most common type of adsorption isotherm

* Received by the editors October 14, 1994; accepted for publication (in revised form) June 1, 1995. This research was supported in part by the National Science Foundation Fellowship Program and Army Research Office grant DAAH04-93-G-0198.

† Department of Mathematics, Santa Clara University, Santa Clara, CA 95053.

used by chemical engineers. It corresponds to a simple kinetic model for adsorption that effectively describes a large range of observed solid–liquid kinetic behavior [7]. For a chromatography reactor in which the two components interreact with first-order kinetics and the solid–liquid equilibrium is governed by the Langmuir isotherm, the mass balances for the two species are

$$\frac{\partial c_1}{\partial x} + \frac{\partial}{\partial t}\left(\frac{c_1}{1+c_1+c_2}\right) = -kc_1 + k'c_2 \tag{1.1a}$$

$$\text{and}\quad \frac{\partial c_2}{\partial x} + \frac{\partial}{\partial t}\left(\frac{\gamma c_2}{1+c_1+c_2}\right) = \gamma(kc_1 - k'c_2), \tag{1.1b}$$

$$\text{where } t \in (-\infty,\infty), \quad x \in [0,\infty).$$

In these balances, c_1 and c_2 are proportional to the two species' concentrations in the liquid phase, k and k' are proportional to the rate constants in the reaction, x is proportional to the length down the reactor, t is related to the time, and γ is a known constant between 0 and 1 dependent on the nature of the Langmuir isotherm. To avoid potential confusion, it should be noted that the mathematical roles played by x and t in the above chromatography equations are the opposite of the mathematical roles that x and t typically play in physical hyperbolic systems (e.g., gas dynamics) and in most hyperbolic systems literature.

It is not immediately clear how the species concentrations will behave as the reactor gets progressively longer (i.e., as $x \to \infty$). On one hand, the chemical reaction will tend to push the concentrations toward chemical equilibrium ($c_2/c_1 = k/k'$); on the other hand, the separation caused by the chromatographic effect will tend to push the concentrations away from equilibrium. Even in cases where the concentrations do approach equilibrium, the fact that the equilibrium is defined by a curve, as opposed to one point, means that the nature of the ω-limit set is still uncertain. Further, we would like to know at what rate the ω-limit set is approached. This rate of approach can be useful in designing reactors of this nature.

In §2, we will start to approach these questions by translating the physical conditions of the reactor into an appropriate mathematical problem. We will begin by deriving the mass balances given in (1.1). It is more convenient to describe the behavior of c_1 and c_2 when these concentrations are transformed into two special state vector coordinates called the Riemann invariants, z and w. We will determine the Riemann invariant forms of the mass balances in (1.1) and also determine which regions of the (z, w) graph are physically possible. Finally, we will note some of the properties of the equilibrium curve on the (z, w) phase portrait.

Section 3 will describe the method of generalized characteristics, which will be the tool used to determine the evolution of z and w. Classical characteristics are curves in the (t, x) plane where the shape of the curve and the evolution of the solution along the curve are defined by a characteristic system of ordinary differential equations (ODEs). For a system of n equations with n unknown functions, there are n families of these characteristic curves. Classical characteristics can be used to solve nonlinear differential equations in the region between their initial condition ($x = 0$) and the value of x corresponding to the first shock formation. At the shock formation, characteristic curves of the same family cross each other, leading to multiple possible solutions.

The method of generalized characteristics allows the characteristic system of ODEs to be extended past shocks to *all* $x \in [0, \infty)$. This is accomplished by considering any specific point in the (t, x) plane and looking at the funnel of all classical characteristic curves and shock curves of each family that emanate backward in x from that point to the initial condition $x = 0$. Each funnel is bounded by a minimal and maximal backward characteristic. For genuinely nonlinear systems (such as chromatography reactors with Langmuir isotherms), the minimal and maximal backward characteristics are not shocks; they propagate with classical characteristic speed. This fact, along with the standard entropy condition (which corresponds, physically, to requiring the concentration c_1 at a fixed location to increase as shocks pass through the location), will allow us to study the solution in regions of the (t, x) plane with shocks.

The nonlinearity caused by the Langmuir isotherm will identify the system in (1.1) as part of a special class of equations defined by Temple in 1983 [9]. This class of equations is remarkable in the sense that shocks formed by the characteristics associated with the family of one Riemann invariant will not induce discontinuities in the other Riemann invariants. This property has the powerful effect of preventing rarefaction waves from occurring when $x \neq 0$. In other words, there can only be one generalized characteristic (classical or shock) from each family emanating in the forward direction from any point (t, x), where $x > 0$.

Section 3 will show how basic properties of the minimal and maximal characteristics can be used to derive one-sided Lipschitz bounds on the variation of the solution in the t direction. These bounds will be used to prove that the Riemann invariants exhibit the properties of Temple equations even though (1.1) contains inhomogeneous reaction terms. From these properties, we will be able to determine the structure of the solution. The approach used in this section will involve modifying the method used by Dafermos and Geng to describe the behavior of characteristics for the homogeneous form of (1.1) [2].

With the tools of generalized characteristics in place, we will be able to explore the long-term behavior of (1.1) in §4. We will consider the case of periodic initial conditions since they physically correspond to the conditions under which a continuously run chromatography reactor would operate. First, we will define regions of the (z, w) phase plane that are invariant (i.e., if the solution is contained by the region at $x = x_o$, then the solution is contained by the region for all $x > x_o$). We will then establish that under periodic conditions the smallest invariant region containing the solution shrinks in proportion to its own size as x increases. It immediately follows from this conclusion that the ω-limit set is comprised of a single point on the chemical equilibrium curve and that this point is approached exponentially as x increases. Finally, by using an integrated form of the mass balances in (1.1), we will determine precisely which point on the equilibrium curve is being approached.

2. Mathematical modeling of the chromatography reactor.

We begin with the mass balances of the two components in the chromatography reactor

$$\frac{\partial c_1}{\partial x} + \frac{\partial n_1}{\partial t} = -kc_1 + k'c_2, \tag{2.1a}$$

$$\frac{\partial c_2}{\partial x} + \frac{\partial n_2}{\partial t} = kc_1 - k'c_2. \tag{2.1b}$$

where c_1, c_2 and n_1, n_2 are the component concentrations in the reactor's liquid and solid phases, respectively, x is the location down the reactor, t is related to time, and

k and k' are the first-order reaction rate constants. The derivation of this system is detailed by Rhee, Aris, and Amundson in [7].

The Langmuir isotherm expresses each solid concentration as a function of the liquid concentrations:

$$n_i = \frac{NK_i c_i}{1 + K_1 c_1 + K_2 c_2}, \qquad i = 1, 2, \tag{2.2}$$

where N, K_1 and K_2 are experimentally determined constants. Inserting (2.2) into (2.1), rescaling t, x, the concentrations, and the rate constants, and defining $\gamma \equiv K_2/K_1$ leads to the forms of the mass balances shown in §1:

$$\frac{\partial c_1}{\partial x} + \frac{\partial}{\partial t}\left(\frac{c_1}{1 + c_1 + c_2}\right) = -kc_1 + k'c_2 \tag{1.1a}$$

and

$$\frac{\partial c_2}{\partial x} + \frac{\partial}{\partial t}\left(\frac{\gamma c_2}{1 + c_1 + c_2}\right) = \gamma(kc_1 - k'c_2), \tag{1.1b}$$

$$\text{where } t \in (-\infty, \infty), \quad x \in [0, \infty).$$

We can label the two chemical components so that $K_1 \geq K_2$ and therefore state that $\gamma \in (0, 1]$. We will consider $\gamma \in (0, 1)$ since the case where $\gamma = 1$ can be solved with the aid of the single homogeneous expression for $c_1 + c_2$ obtained by adding the two equations in (1.1) together. Again, we point out that the mathematical roles played by x and t in (1.1) are the reverse of the mathematical roles that x and t typically play in hyperbolic-systems literature.

The Riemann-invariant form of the system in (1.1) partially decouples the behavior of the two components. This form can be obtained by first transforming (1.1) into a system investigated by Dafermos and Geng [2] by making a change in the state variable coordinates

$$u \equiv (c_1 + c_2 + 1)/\gamma, \qquad v \equiv (\gamma c_1 + c_2 + 1 + \gamma)/\gamma$$

and also making a scale change in the independent variable by defining x/γ to be equal to a new variable x, which yields

$$\frac{\partial u}{\partial x} - \frac{\partial (v/u)}{\partial t} = -au + bv - c, \tag{2.3a}$$

$$\frac{\partial v}{\partial x} - \frac{\partial (1/u)}{\partial t} = 0, \tag{2.3b}$$

$$\text{where } a \equiv \gamma(k + k'\gamma), \quad b \equiv \gamma(k + k'), \quad \text{and } c \equiv (k' + k\gamma).$$

It should be noted that the constants in (2.3) are all positive and have the property that $b^2 - ac = -\gamma kk'(1 - \gamma)^2 < 0$.

The Riemann invariants z and w for the system in (2.3) are

$$z \equiv \frac{1}{\lambda u} \quad \text{and} \quad w \equiv \frac{1}{\mu u}, \tag{2.4}$$

where λ and μ are the characteristic speeds of the system, which can be expressed in (u, v) coordinates or (z, w) coordinates:

$$\lambda = \frac{v - \sqrt{v^2 - 4u}}{2u^2} = \frac{1}{z^2 w}, \qquad \mu = \frac{v + \sqrt{v^2 - 4u}}{2u^2} = \frac{1}{zw^2}. \tag{2.5}$$

In regions where the solution to (2.3) is smooth, the system can be transformed into the desired Riemann-invariant form:

$$\frac{\partial z}{\partial x} + \lambda \frac{\partial z}{\partial t} = g(z, w), \tag{2.6a}$$

$$\frac{\partial w}{\partial x} + \mu \frac{\partial w}{\partial t} = -g(z, w), \tag{2.6b}$$

$$\text{where } g(z, w) \equiv \frac{a(zw) - b(z + w) + c}{z - w}.$$

We will assume in any solution to (2.6) that the Riemann invariants do not stray, for some constant M, from the following restricted ranges:

$$z \in \left[\frac{1}{\gamma}, M\right] \quad \text{and} \quad w \in \left[1, \frac{1}{\gamma}\right]. \tag{2.7}$$

These restrictions physically correspond to requiring that the concentrations, c_1 and c_2, are nonnegative and do not approach infinity.

Further, we assume that the solution remains bounded away from the umbilic point $z = w = 1/\gamma$, which implies that the system is strictly hyperbolic since $0 < \lambda < \mu$.

Both characteristic fields are genuinely nonlinear since

$$\frac{\partial \lambda}{\partial z} < 0 \quad \text{and} \quad \frac{\partial \mu}{\partial w} < 0, \tag{2.8}$$

which is immediately seen from (2.5) and (2.7).

The chemical equilibrium curve ($c_2/c_1 = k/k'$) for (1.1) corresponds in the Riemann-invariant system to the curve defined by $g(z, w) = 0$. If we define

$$z_e(w) \equiv \frac{c - bw}{b - aw} \quad \text{and} \quad w_e(z) \equiv \frac{c - bz}{b - az}, \tag{2.9}$$

then we have that $g(z_e(w), w) = g(z, w_e(z)) \equiv 0$ from the definition of g. Therefore, both z_e and w_e define the equilibrium curve in the Riemann-invariant plane.

We wish to note some of the properties of g and z_e. From (2.7) we see that $z_e(w)$ is defined only for $w \in [1, \frac{c-bM}{b-aM}]$. This restricted domain of z_e implies that the expressions for the first and second derivatives of z_e must be positive:

$$\frac{dz_e}{dw} = \frac{ac - b^2}{(aw - b)^2} > 0, \tag{2.10a}$$

$$\frac{d^2 z_e}{dw^2} = \frac{-2a(ac - b^2)}{(aw - b)^3} > 0. \tag{2.10b}$$

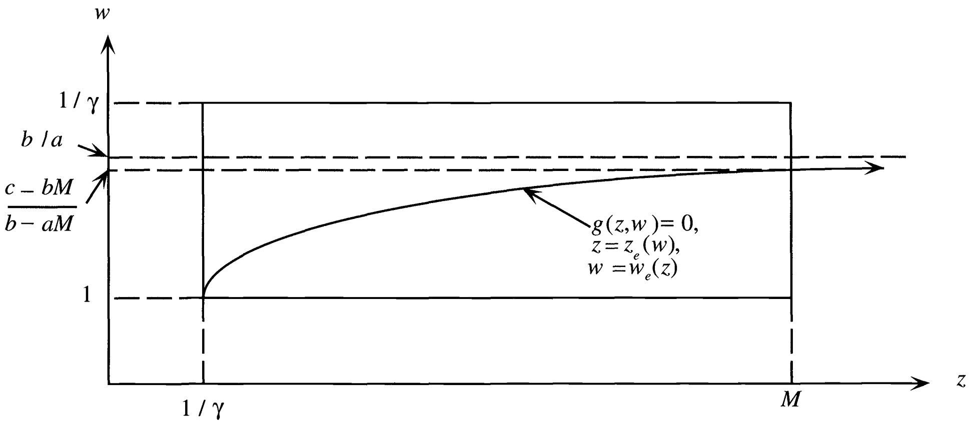


FIG. 1. *The Riemann-invariant phase plane. The invariants can only take values inside the solid rectangle shown above. As $z \longrightarrow \infty$, the equilibrium curve asymptotically approaches $w = b/a$.*

These properties of $z_e(w)$ along with (2.7) allow us to construct the phase plane shown in Fig. 1. Finally, we have that

$$\frac{\partial g}{\partial w} = \frac{az^2 - 2bz + c}{(w-z)^2} > 0, \tag{2.11a}$$

$$\frac{\partial g}{\partial z} = \frac{aw^2 - 2bw + c}{-(w-z)^2} < 0. \tag{2.11b}$$

The speed of shock propagation σ is given by the Rankine–Hugoniot conditions as applied to (2.3). These can be expressed in terms of the Riemann invariants by using the fact that $v = z + w$ and $u = zw$:

$$\sigma(z_+w_+ - z_-w_-) + z_+^{-1} + w_+^{-1} - z_-^{-1} - w_-^{-1} = 0, \tag{2.12a}$$

$$\sigma(z_+ + w_+ - z_- - w_-) + z_+^{-1}w_+^{-1} - z_-^{-1}w_-^{-1} = 0, \tag{2.12b}$$

where the subscripts "+" and "−" refer to the limit as t approaches the shock from the right and the left, respectively. Solving (2.12) leads to two possible shock speeds: 1-shocks, where $z_- \neq z_+$, $w_- = w_+$, and

$$\sigma = z_-^{-1}z_+^{-1}w_\pm^{-1}; \tag{2.13a}$$

and 2-shocks, where $w_- \neq w_+$, $z_- = z_+$, and

$$\sigma = w_-^{-1}w_+^{-1}z_\pm^{-1}. \tag{2.13b}$$

From (2.13), it is clear that shocks associated with one Riemann-invariant family do not induce discontinuities in the other invariant family, and so our system belongs to the class of equations defined by Temple.

The initial-value problem for (2.3) may have many solutions whose shocks satisfy the Rankine–Hugoniot conditions. By using the *entropy*, η, and the *entropy flux*, q, we are able to remove solutions that are physically irrelevant. The entropy conditions used in chromatography are completely analogous to the entropy conditions discussed in the context of gas dynamics [7]. η and q are any two functions satisfying

$$\frac{\partial q}{\partial z} = \lambda \frac{\partial \eta}{\partial z} \quad \text{and} \quad \frac{\partial q}{\partial w} = \mu \frac{\partial \eta}{\partial w}. \tag{2.14}$$

In §3, we will make use of the solution pair $(\eta,\ q)$ to (2.14) of the Lax type considered by Dafermos and Geng [2]:

$$\eta(z,w) = \left[1 - \frac{w}{z} + \frac{2}{l}w - \frac{2}{l^2}zw\right] \exp\left(\frac{l}{z}\right), \tag{2.15a}$$

$$q(z,w) = \left[\left(\frac{1}{z^2 w} - \frac{1}{z^3}\right) - \frac{2}{l}\left(\frac{1}{zw} - \frac{2}{z^2}\right) + \frac{2}{l^2}\left(\frac{1}{w} - \frac{5}{z}\right) + \frac{12}{l^3}\right] \exp\left(\frac{l}{z}\right). \tag{2.15b}$$

In particular, we will use the following two equations, which are determined directly from (2.15):

$$q(z,w) - \lambda(z,w)\eta(z,w) = -\frac{2}{l}\left[\frac{1}{zw}\left(1 - \frac{w}{z}\right) + O\left(\frac{1}{l}\right)\right] \exp\left(\frac{l}{z}\right), \tag{2.16a}$$

$$q(z,w) - [\lambda(z,w) + \varepsilon]\eta(z,w) = -\left[\varepsilon\left(1 - \frac{w}{z}\right) + O\left(\frac{1}{l}\right)\right] \exp\left(\frac{l}{z}\right). \tag{2.16b}$$

Entropies of the type considered by Lax [5] are generally convex only when l is large in one direction. However, the entropies for Temple-class systems are special in that they are convex when l is large in either direction.

Entropy–entropy flux solution pairs also exist that look similar to (2.15) with the roles of z and w interchanged.

Existence of entropy solutions to homogeneous systems of conservation laws in Temple's class under any initial condition with bounded variation has been shown by Serre [8]. Existence for the inhomogeneous system in Temple's class examined in this paper may be established through a routine extension of these methods. The issue of uniqueness of the entropy solution, even in the homogeneous case, has only been partially resolved, and in the inhomogeneous case, the problem is open.

3. Generalized characteristics. We consider a (weak) solution $(c_1(t,x), c_2(t,x))$ of locally bounded variation (BV) to (1.1) defined for $(t,x) \in (-\infty,\infty) \times [0,\infty)$. We assume that the Riemann-invariant fields $(z(t,x), w(t,x))$ induced by this solution take values in a small neighborhood of some fixed state (z_o, w_o) and that the two characteristic fields have well-separated speeds.

We assume that for any fixed $x \geq 0$, the concentrations $c_1(.,x)$ and $c_2(.,x)$—and thereby also the functions $z(.,x)$ and $w(.,x)$—have bounded variation locally on $(-\infty,\infty)$. This ensures that limits of the solution from the left and the right (i.e., $c_1(t\pm,x)$, $c_2(t\pm,x)$, $z(t\pm,x)$, and $w(t\pm,x)$) always exist. If we wish, we can normalize the solution by requiring that the solution always equals its left (or right) limit. This will define the value of the solution on the set of all shocks and, therefore, will require an alteration of the solution on no more than a set of measure zero since BV solutions can have no more than a countable number of shock curves.

We will require that the solution satisfies the standard entropy condition, which can be written in terms of the characteristic speeds

$$\lambda(z(t-,x),w(t-,x)) \geq \lambda(z(t+,x),w(t+,x)),$$

$$\mu(z(t-,x),w(t-,x)) \geq \mu(z(t+,x),w(t+,x)) \tag{3.1}$$

or, equivalently, in terms of the Riemann invariants

$$z(t-,x) \leq z(t+,x), \qquad w(t-,x) \leq w(t+,x). \tag{3.2}$$

These conditions imply that for any convex entropy function, η, and its associated entropy flux, q, the relation

$$\frac{\partial \eta}{\partial x} + \frac{\partial q}{\partial t} \leq g(z,w)\left(\frac{\partial \eta}{\partial z} - \frac{\partial \eta}{\partial w}\right) \tag{3.3}$$

holds in the sense of measures.

The entropy condition in (3.1) has been previously applied to chromatography equations in [7]. It corresponds (see [2]) to the physical requirement

$$c_1(t-,x) \leq c_1(t+,x). \tag{3.4}$$

In other words, as a shock passes through any fixed location in the reactor, c_1, the concentration of the species that can induce higher concentrations in the solid phase (see (2.2)) will increase.

A *classical* 1-characteristic on the interval $[x_1, x_2]$ for the system being studied is an integral curve of the ODE

$$\frac{dt}{dx} = \lambda(z(t,x),w(t,x)). \tag{3.5}$$

Classical characteristics determine the solution in regions where the solution is smooth. To extend characteristics to regions with shocks, it is necessary to consider them as integral curves of (3.5) in the sense of Filippov [3] since the right-hand side of (3.5) can now be discontinuous. It follows that a *generalized* 1-characteristic is defined as a Lipschitz curve $\tau(x)$ such that

$$\tau'(x) \in [\lambda(z(\tau(x)+,x),w(\tau(x)+,x)),\lambda(z(\tau(x)-,x),w(\tau(x)-,x))] \quad \text{a.e.} \tag{3.6}$$

Similarly, a classical 2-characteristic satisfies

$$\frac{dt}{dx} = \mu(z(t,x),w(t,x)), \tag{3.7}$$

and a generalized 2-characteristic is a Lipschitz curve $v(x)$, where

$$v'(x) \in [\mu(z(v(x)+,x),w(v(x)+,x)),\mu(z(v(x)-,x),w(v(x)-,x))] \quad \text{a.e.} \tag{3.8}$$

We state some elementary properties of characteristics which can be determined by extending the proofs in [1]. Every generalized i-characteristic ($i = 1, 2$) propagates with either classical characteristic speed or with the shock speed associated with the i family. From any point (T, X) in the upper half-plane, there emanates in the backward (i.e., decreasing-x) direction a 1-characteristic funnel and a 2-characteristic funnel, which only intersect at (T, X). Each funnel is bounded by minimal and maximal i-characteristics (which may or may not be distinct). These minimal and maximal characteristics always propagate with classical characteristic speed; specifically, we have the following.

LEMMA 3.1. *Consider* $(T, X) \in (-\infty, \infty) \times (0, \infty)$. *If* τ *is the minimal or maximal backward* 1*-characteristic emanating from* (T, X), *then for almost every* $x \in [0, X]$,

$$z(\tau(x)-, x)) = z(\tau(x)+, x)), \qquad w(\tau(x)-, x)) = w(\tau(x)+, x)), \tag{3.9}$$

$$\tau'(x) = \lambda(z(\tau(x)\pm, x), w(\tau(x)\pm, x)). \tag{3.10}$$

If v *is the minimal or maximal backward* 2*-characteristic emanating from* (T, X), *then for almost every* $x \in [0, X]$,

$$z(v(x)-, x)) = z(v(x)+, x)), \qquad w(v(x)-, x)) = w(v(x)+, x)), \tag{3.11}$$

$$v'(x) = \mu(z(v(x)\pm, x), w(v(x)\pm, x)). \tag{3.12}$$

Now we consider the behavior of the Riemann invariants on these minimal and maximal characteristics. Over a classical 1-characteristic, we have that $\frac{dz}{dx} = g(z, w)$, and over a classical 2-characteristic, we have that $\frac{dw}{dx} = -g(z, w)$. We begin to derive similar results for the extremal characteristics by using entropy.

LEMMA 3.2. *Let* τ *and* v *be the minimal and maximal backward* 1*-characteristics emanating from any point* (T, X) *of the upper half-plane. Then for all* $0 \le \xi \le \chi \le X$,

$$z(\tau(\chi)-, \chi) - z(\tau(\xi)-, \xi) \le \int_{\xi}^{\chi} g(z(\tau(x)-, x), w(\tau(x)-, x))dx, \tag{3.13}$$

$$z(v(\chi)+, \chi) - z(v(\xi)+, \xi) \ge \int_{\xi}^{\chi} g(z(v(x)+, x), w(v(x)+, x))dx. \tag{3.14}$$

Similarly, if τ *and* v *are the minimal and maximal backward* 2*-characteristics emanating from the point* (T, X), *then for all* $0 \le \xi \le \chi \le X$,

$$w(\tau(\chi)-, \chi) - w(\tau(\xi)-, \xi) \le \int_{\xi}^{\chi} -g(z(\tau(x)-, x), w(\tau(x)-, x))dx, \tag{3.15}$$

$$w(v(\chi)+, \chi) - w(v(\xi)+, \xi) \ge \int_{\xi}^{\chi} -g(z(v(x)+, x), w(v(x)+, x))dx. \tag{3.16}$$

Proof. We will only show the proof of (3.13), since the proofs of the other three relations are similar.

First, we define τ_ε, an integral curve in the sense of Filippov of the differential equation

$$\frac{dt}{dx} = \lambda(z(t, x), w(t, x)) + \varepsilon \tag{3.17}$$

which emanates from the point $(T-\varepsilon, X)$, where ε is a small, positive, fixed number. It follows that

$$\tau_\varepsilon'(x) \geq \lambda(z(\tau_\varepsilon(x)+, x), w(\tau_\varepsilon(x)+, x)) + \varepsilon \tag{3.18}$$

for almost every $x \in [0, X]$ and that $\tau_\varepsilon(x) < \tau(x)$ for all $x \in [0, X]$. Further, $\tau_\varepsilon(x)$ converges uniformly on $[0, X]$ to $\tau(x)$ as $\varepsilon \to 0$.

We now integrate the entropy inequality in (3.3) over the domain $\{(t,x) : \xi \leq x \leq \chi, \tau_\varepsilon(x) \leq t \leq \tau(x)\}$ and apply the Gauss–Green theorem to obtain

$$\begin{aligned} &\int_{\tau_\varepsilon(\chi)}^{\tau(\chi)} \eta(z(t,\chi), w(t,\chi))dt - \int_{\tau_\varepsilon(\xi)}^{\tau(\xi)} \eta(z(t,\xi), w(t,\xi))dt \\ &+ \int_\xi^\chi q(z(\tau(x)-, x), w(\tau(x)-, x)) - \tau'(x)\eta(z(\tau(x)-, x), w(\tau(x)-, x))dx \\ &- \int_\xi^\chi q(z(\tau_\varepsilon(x)+, x), w(\tau_\varepsilon(x)+, x)) - \tau_\varepsilon'(x)\eta(z(\tau_\varepsilon(x)+, x), w(\tau_\varepsilon(x)+, x))dx \\ &\leq \int_\xi^\chi \int_{\tau_\varepsilon(x)}^{\tau(x)} g(z,w)\left(\frac{\partial \eta}{\partial z} - \frac{\partial \eta}{\partial w}\right) dtdx. \end{aligned} \tag{3.19}$$

By using the expressions for η and q given in (2.15) with l very large and negative, we can use (2.16), (3.10), and (3.18) to show that the last two integrals on the left-hand side of (3.19) are nonnegative and nonpositive respectively. Therefore, both integrals can be removed from (3.19) and the inequality is preserved:

$$\begin{aligned} &\int_{\tau_\varepsilon(\chi)}^{\tau(\chi)} \eta(z(t,\chi), w(t,\chi))dt - \int_{\tau_\varepsilon(\xi)}^{\tau(\xi)} \eta(z(t,\xi), w(t,\xi))dt \\ &\leq \int_\xi^\chi \int_{\tau_\varepsilon(x)}^{\tau(x)} -\frac{lg(z(t,x), w(t,x))}{z^2} \exp\left(\frac{l}{z}\right)\left(1 - \frac{w}{z} + O\left(\frac{1}{l}\right)\right) dtdx. \end{aligned} \tag{3.20}$$

If we define $\bar{z}(x) \equiv \operatorname{ess\,inf}_{[\tau_\varepsilon(x), \tau(x)]} z(.,x)$, $\bar{g}(x) \equiv \operatorname{ess\,sup}_{[\tau_\varepsilon(x), \tau(x)]} g(z(.,x), w(.,x))$, and $\Phi(x) \equiv \int_{\tau_\varepsilon(x)}^{\tau(x)} \eta dt$, we can express (3.20) in terms of Φ:

$$\Phi(\chi) - \Phi(\xi) \leq \int_\xi^\chi \frac{-l\bar{g}(x)}{\bar{z}^2(x)}(1 + O(1/l))\Phi(x)dx. \tag{3.21}$$

Since the integrand in (3.21) is finite, it immediately follows that $\Phi(x)$ is upper Lipschitz continuous. Any upper-Lipschitz-continuous function $f(x)$ is differentiable almost everywhere and $\int_a^b \frac{df(x)}{dx}dx \geq f(b) - f(a)$. Therefore, when we divide (3.21) by $\chi - \xi$ and let $\xi \to \chi$, we get an expression for the derivative of Φ almost everywhere. Further, after rearranging the resulting expression and applying the chain rule, we can integrate between new ξ and χ points and—since $\ln(\Phi(x))$ is upper Lipschitz continuous—we preserve the inequality

$$\ln\left(\frac{\Phi(\chi)}{\Phi(\xi)}\right) \leq \int_\xi^\chi \frac{d\ln(\Phi(x))}{dx}dx \leq \int_\xi^\chi \frac{-l\bar{g}(x)}{\bar{z}^2(x)}(1 + O(1/l))dx. \tag{3.22}$$

Now we multiply (3.22) by $-1/l$ and let $l \to -\infty$, which causes the Φ terms to converge to the L^∞ norm of the exponentials in their entropy expressions given in (2.15), which leads to

$$-\frac{1}{\tilde{z}(\chi)} + \frac{1}{\tilde{z}(\xi)} \leq \int_\xi^\chi \frac{\bar{g}(x)}{\tilde{z}^2(x)} dx, \tag{3.23}$$

where $\tilde{z}(x) \equiv \operatorname{ess\,sup}_{[\tau_\varepsilon(x),\tau(x)]} z(.,x)$. Letting $\varepsilon \to 0$ forces the z terms on both sides of the inequality to converge to the limit of z from the left of the $\tau(x)$ curve:

$$-\frac{1}{\hat{z}(\chi)} + \frac{1}{\hat{z}(\xi)} \leq \int_\xi^\chi \frac{\hat{g}(x)}{\hat{z}^2(x)} dx, \tag{3.24}$$

where $\hat{z}(x) \equiv z(\tau(x)-,x)$ and $\hat{g}(x) \equiv g(z(\tau(x)-,x), w(\tau(x)-,x))$. From the upper-Lipschitz-function argument used to establish (3.22), we can "differentiate" (3.24), cancel terms, and then integrate while still preserving the inequality. This yields

$$\hat{z}(\chi) - \hat{z}(\xi) \leq \int_\xi^\chi \hat{g}(x) dx, \tag{3.25}$$

which is identical to (3.13).

An almost identical method is used to show (3.14), but instead of using a characteristic τ_ε, which is displaced slightly to the left of τ, one uses a characteristic v_ε, which is displaced slightly to the right of v. Also, we let l (as opposed to $-l$) be large. In the proofs of (3.15) and (3.16), we follow the proofs of (3.13) and (3.14) but use the form of the entropy which is similar to (2.15) with the roles of z and w switched.

The next lemma establishes bounds on the widening of extremal backward characteristics of a family. These bounds depend on the total variation of the Riemann invariant associated with the other family along this extremal curve. We will assume in any x interval of finite length that this variation is bounded uniformly over the extremal characteristics; specifically, for any $x_1 < \infty$, we have that

$$K_z(x_1) \equiv \sup_v \left[TV_{x\in[0,x_1]} z(v(x)-,x)\right] < \infty \tag{3.26a}$$

and

$$K_w(x_1) \equiv \sup_\tau \left[TV_{x\in[0,x_1]} w(\tau(x)-,x)\right] < \infty, \tag{3.26b}$$

where v is any extremal 2-characteristic and τ any extremal 1-characteristic that emanates from the line $x = x_1$.

LEMMA 3.3. *Let $-\infty < S \leq T < \infty$ and $X > 0$. If τ is the minimal backward 1-characteristic emanating from (S, X) and v is the maximal backward 1-characteristic emanating from (T, X), then there exist positive constants D, δ, β, α, and L such that for $x \in \{[0, X] : X - x \leq L\}$,*

$$\begin{aligned} v(x) - \tau(x) \leq{} & 2[T-S] \exp\left(\frac{D}{\delta}(\Theta(X) - \Theta(x))\right) \\ & + 2\beta[z(T+,X) - z(S-,X)] \int_x^X \exp\left(\frac{D}{\delta}(\Theta(y) - \Theta(x)) + \alpha(X-y)\right) dy, \end{aligned} \tag{3.27}$$

where $\Theta(\xi) \equiv TV_{x\in[0,\xi]}w(\tau(x)-,x)+\xi$. *Similarly, if* τ *is the minimal backward* 2*-characteristic emanating from* (S,X) *and* v *is the maximal backward* 2*-characteristic emanating from* (T,X)*, then there exist positive constants* D*,* δ*,* β*,* α*, and* L *such that for* $x \in \{[0,X] : X-x \le L\}$*,*

$$\begin{aligned} v(x)-\tau(x) &\le 2[T-S]\exp\left(\frac{D}{\delta}(\Theta(X)-\Theta(x))\right) \\ &\quad + 2\beta[w(T+,X)-w(S-,X)]\int_x^X \exp\left(\frac{D}{\delta}(\Theta(y)-\Theta(x))+\alpha(X-y)\right)dy, \end{aligned} \tag{3.28}$$

where $\Theta(\xi) \equiv TV_{x\in[0,\xi]}z(\tau(x)-,x)+\xi$.

Proof. We will only show (3.27) since the proof for (3.28) is essentially identical. The proof for (3.27) is somewhat involved, so we will employ the following simplifying notation:

$$z(x) \equiv z(\tau(x)-,x), \qquad w(x) \equiv w(\tau(x)-,x),$$

$$Z(x) \equiv z(v(x)+,x), \qquad W(x) \equiv w(v(x)+,x),$$

$$\zeta(x) \equiv z(x)-Z(x), \qquad \omega(x) \equiv w(x)-W(x),$$

$$\phi(x) \equiv v(x)-\tau(x).$$

Since we are looking for bounds on the characteristics' widening, we begin with the expressions for the characteristics' slopes given in (3.10) and (3.12):

$$\phi'(x) = \frac{1}{Z^2W}-\frac{1}{z^2w} = \frac{z+Z}{WZ^2z^2}\zeta(x)+\frac{1}{z^2Ww}\omega(x) \quad \text{a.e.,} \tag{3.29}$$

where z, Z, w, and W are all functions of x. From (2.7), it is clear that the fractions in front of ζ and ω are bounded away from both 0 and ∞, so we next concentrate on looking for bounds on the behavior of ζ and ω.

Lemma 3.2 gives an expression bounding ζ:

$$\begin{aligned} \zeta(y)-\zeta(x) &\le \int_x^y -\frac{aZW-b(Z+W)+c}{Z-W}+\frac{azw-b(z+w)+c}{z-w}d\xi \\ &= \int_x^y \frac{-aW^2+2bW-c}{(Z-W)(z-W)}\zeta(\xi)-\frac{-az^2+2bz-c}{(z-W)(z-w)}\omega(\xi)d\xi, \end{aligned} \tag{3.30}$$

where z, Z, w, and W are all functions of ξ. Since we are bounded away from the umbilic point where $w(t,x)=z(t,x)$, the denominators in the fractions of (3.30) are bounded away from zero. The numerators in the fractions of (3.30) are also bounded from zero since $b^2-ac<0$. Therefore, there are positive, finite constants a_1, a_2, and b_1 such that $a_1 > a_2$ and

$$\zeta(y)-\zeta(x) \le \int_x^y [-a(\xi)\zeta(\xi)+b_1|\omega(\xi)|]\,d\xi, \tag{3.31}$$

$$\text{where } a(\xi) \equiv \begin{cases} a_1 & \text{if } \zeta(\xi) < 0, \\ a_2 & \text{if } \zeta(\xi) \geq 0. \end{cases}$$

As in Lemma 3.2, the properties of upper-Lipschitz-continuous functions allow us to "differentiate" (3.31), multiply the result by an appropriate integrating factor, and preserve the inequality upon integration between x and X, leading to

$$\zeta(x) \geq \zeta(X) \exp\left(\int_x^X a(\xi)d\xi\right) - b_1 \int_x^X |\omega(\xi)| \exp\left(\int_x^\xi a(\chi)d\chi\right) d\xi, \quad x \in [0, X]. \tag{3.32}$$

From (3.32), we obtain our final form of the bound on $\zeta(x)$:

$$\zeta(x) \geq \zeta(X)e^{\alpha(X-x)} - b_1 e^{a_1(X-x)} \int_x^X |\omega(\xi)| d\xi, \tag{3.33}$$

$$\text{where } \alpha \equiv \begin{cases} a_1 & \text{if } \zeta(X) \leq 0, \\ a_2 & \text{if } \zeta(X) > 0. \end{cases}$$

Next, we derive the following bound on the integral of $\omega(x)$:

$$\int_x^X |\omega(y)| dy \leq \delta\phi(x) + D \int_x^X \phi(y) d\Theta(y). \tag{3.34}$$

We require a somewhat intricate construction to establish (3.34). First, we define the function $P(y)$ in the case where $\omega(y) \leq 0$ to be the x coordinate of the point of intersection of τ with the *minimal* backward 2-characteristic emanating from $(v(y), y)$. In the case where $\omega(y) > 0$, $P(y)$ is defined to be the x coordinate of the point of intersection of τ with the *maximal* backward 2-characteristic emanating from $(v(y), y)$. Next, we define $p(y) \equiv \inf_{\xi \in [y,X]} P(\xi)$ on the domain $y \in (\bar{y}, X]$, where $\bar{y} \equiv \inf\{y \in [0, X] : P(y) \text{ is defined}\}$. The domain of definition of p is extended to the point $\bar{y}$ by using $p(\bar{y}) \equiv \inf_{[\bar{y},X]} P(\xi)$ when $P(\bar{y})$ is defined or $p(\bar{y}) \equiv \inf_{(\bar{y},X]} P(\xi)$ when $P(\bar{y})$ is not defined. p is a monotonically increasing function with the property that $p(y) \leq y$ since the characteristic speeds are positive. We also define an "inverse" for p: $h(\xi) \equiv \sup\{y : p(y-) \leq \xi \leq p(y)\}$.

Now we fix $x \in [0, X]$ and define $\hat{y}$ by $\hat{y} \equiv \bar{y}$ if $x \in [0, p(\bar{y})]$, $\hat{y} \equiv h(x)$ if $x \in (p(\bar{y}), p(X)]$, and $\hat{y} \equiv X$ if $x \in (p(X), X]$.

Since the characteristic speeds are well separated, there is a constant C such that $0 \leq \hat{y} - x \leq C\phi(x)$. Combining this with the requirement that w stays in a small neighborhood of some fixed state yields

$$\int_x^{\hat{y}} |\omega(y)| dy \leq \delta\phi(x), \quad \text{where } \delta \ll 1. \tag{3.35}$$

This estimate also holds, with $\hat{y} = X$, when $\bar{y}$ is undefined.

Now we consider the integral of $|\omega(y)|$ over the remaining region: $y \in (\hat{y}, X]$. Using Lemma 3.2, we can establish that

$$|\omega(y)| \leq \Lambda(y) - \Lambda(P(y)) + ||g||_{L^\infty}(y - P(y)), \tag{3.36}$$

where $\Lambda(y) \equiv TV_{\xi \in [0,y]} w(\tau(\xi)-, \xi)$. $||g||_{L^\infty}$ must exist since the ranges of z and w are restricted by (2.7). From the definition of $p(y)$, it is obvious that $p(y) \leq P(y)$. Using this and the fact that the separation of speeds implies that $y - P(y) \leq C\phi(y)$ allows the integral of (3.36) to be expressed as

$$\int_{\hat{y}}^{X} |\omega(y)|dy \leq \int_{\hat{y}}^{X} \Lambda(y) - \Lambda(p(y))dy + C\|g\|_{L^\infty} \int_{\hat{y}}^{X} \phi(y)dy. \tag{3.37}$$

We wish to reexpress the integral of the total variations in (3.37). This is accomplished by constructing a sequence $\{\Lambda_n\}$ of nondecreasing, absolutely continuous functions which converge pointwise to $\Lambda(y)$ on $[0, X]$ so that we have

$$\int_{\hat{y}}^{X} [\Lambda_n(y) - \Lambda_n(p(y))]dy = \int_{\hat{y}}^{X} \int_{p(y)}^{y} \Lambda_n'(\xi)d\xi dy \tag{3.38}$$

$$= \int_{p(\hat{y})}^{\hat{y}} [h(\xi) - \hat{y}]\Lambda_n'(\xi)d\xi + \int_{\hat{y}}^{p(X)} [h(\xi) - \xi]\Lambda_n'(\xi)d\xi + \int_{p(X)}^{X} [X - \xi]\Lambda_n'(\xi)d\xi.$$

From the separation of the characteristic speeds and the definition of h, we infer that

$$h(\xi) - \hat{y} \leq C\phi(\xi), \qquad \xi \in [p(\hat{y}), \hat{y}],$$

$$h(\xi) - \xi \leq C\phi(\xi), \qquad \xi \in (\hat{y}, p(X)],$$

$$X - \xi \leq C\phi(\xi), \qquad \xi \in (p(X), X].$$

Moreover, by the construction of $\hat{y}$, it follows that $p(\hat{y}) \geq x$. Therefore, (3.38) yields

$$\int_{\hat{y}}^{X} [\Lambda_n(y) - \Lambda_n(p(y))]dy \leq C \int_{x}^{X} \phi(\xi)d\Lambda_n(\xi). \tag{3.39}$$

Passing to the limit as $n \to \infty$ causes (3.39) to converge to

$$\int_{\hat{y}}^{X} [\Lambda(y) - \Lambda(p(y))]dy \leq C \int_{x}^{X} \phi(\xi)d\Lambda(\xi). \tag{3.40}$$

Substitution of (3.40) into (3.37) yields

$$\int_{\hat{y}}^{X} |\omega(y)|dy \leq D \int_{x}^{X} \phi(y)d\Theta(y), \tag{3.41}$$

where D is a constant and $d\Theta(y) \equiv d\Lambda(y) + dy$. This combined with (3.35) establishes (3.34).

With the bounds on ζ and ω established in (3.33) and (3.34), we return to the bound on ϕ in (3.29). Substitution of (3.33) into (3.29) yields

$$\phi'(y) \geq -A|\omega(y)| + \beta\zeta(X)e^{\alpha(X-y)} - Be^{a_1(X-y)} \int_{y}^{X} |\omega(\xi)|d\xi \quad \text{a.e.}, \tag{3.42}$$

where A, B, and β are all positive constants. Now we integrate (3.42) between x and X. The double integral in the resulting relation can be reexpressed by the following change in the order of integration:

$$\int_{x}^{X} \int_{y}^{X} |\omega(\xi)|e^{a_1(X-y)}d\xi dy = \int_{x}^{X} |\omega(\xi)| \int_{x}^{\xi} e^{a_1(X-y)}dyd\xi$$

$$\leq (1/a_1)(e^{a_1(X-x)}-1)\int_x^X |\omega(\xi)|d\xi \tag{3.43}$$

so that (3.42) yields

$$\phi(x) \leq \phi(X) - (\beta/\alpha)\zeta(X)(e^{\alpha(X-x)}-1) + [A + (B/a_1)(e^{a_1(X-x)}-1)]\int_x^X |\omega(\xi)|d\xi. \tag{3.44}$$

Next, we substitute (3.34) into (3.44). Further, we restrict $X-x \leq L$, where L is some positive constant such that the bracketed term in (3.44) is less than $1/(2\delta)$ whenever $X-x \leq L$. This process yields

$$\phi(x) \leq 2\phi(X) - 2(\beta/\alpha)\zeta(X)(e^{\alpha(X-x)}-1) + \frac{D}{\delta}\int_x^X \phi(y)d\Theta(y). \tag{3.45}$$

Application of a generalized form of Gronwall's inequality [2], [4] to (3.45) leads to

$$\begin{aligned}\phi(x) &\leq 2\phi(X)\exp\left[\frac{D}{\delta}(\Theta(X)-\Theta(x))\right] \\ &\quad - \zeta(X)2\beta\int_x^X \exp\left[\frac{D}{\delta}(\Theta(y)-\Theta(x)) + \alpha(X-y)\right]dy,\end{aligned} \tag{3.46}$$

which is equivalent to (3.27), the claim of the lemma.

From Lemma 3.3, we can quickly deduce the following.

LEMMA 3.4. *If (T,X) is any point of the upper half-plane where $z(T-,X) = z(T+,X)$, then a unique backward 1-characteristic, τ, emanates from (T,X) and*

$$z(\tau(x)\pm,x) = z(T\pm,X) - \int_x^X g(z(\tau(y)\pm,y),w(\tau(y)\pm,y))dy, \quad x \in (0,X]. \tag{3.47}$$

Similarly, if $w(T-,X) = w(T+,X)$, then a unique backward 2-characteristic v emanates from (T,X) and

$$w(v(x)\pm,x) = w(T\pm,X) + \int_x^X g(z(v(y)\pm,y),w(v(y)\pm,y))dy, \quad x \in (0,X]. \tag{3.48}$$

Proof. Since $z(T+,X) = z(T-,X)$, (3.27) clearly implies that the minimal and maximal backward characteristics must be the same curve, $\tau(y)$, for all $y \in [0,X]$. From (3.9) of Lemma 3.1, we have that

$$\int_x^X g(z(\tau(y)-,y),w(\tau(y)-,y))dy = \int_x^X g(z(\tau(y)+,y),w(\tau(y)+,y))dy; \tag{3.49}$$

therefore, Lemma 3.2 implies that

$$z(\tau(x)+,x) \leq z(T\pm,X) - \int_x^X g(z(\tau(y)\pm,y),w(\tau(y)\pm,y))dy \leq z(\tau(x)-,x). \tag{3.50}$$

However, the entropy condition in (3.2) states that $z(\tau(x)+,x) \geq z(\tau(x)-,x)$, so (3.50) must hold as an equality, which establishes (3.47).

The proof for (3.48) is completely analogous.

Lemma 3.3 also gives us the following one-sided Lipschitz conditions.

LEMMA 3.5. *There are functions $H(X) > 0$ and $\hat{H}(X) > 0$ which are unbounded only as $X \to 0$ such that*

$$\frac{z(T,X) - z(S,X)}{T - S} \geq -H(X), \quad -\infty < S < T < \infty, \quad X > 0, \tag{3.51}$$

$$\frac{w(T,X) - w(S,X)}{T - S} \geq -\hat{H}(X), \quad -\infty < S < T < \infty, \quad X > 0. \tag{3.52}$$

Proof. It suffices to establish (3.51) under the assumptions $z(T-,X) = z(T+,X)$, $z(S-,X) = z(S+,X)$, and $z(S,X) > z(T,X)$. Since the left-hand side of (3.27) is nonnegative, we immediately have for any $x \in [0,X]$, where $x \geq X - L$, that

$$\frac{z(T,X) - z(S,X)}{T - S} \geq \frac{-\exp\left(-\frac{D}{\delta}(\Theta(x) - \Theta(X))\right)}{\beta \int_x^X \exp\left(-\frac{D}{\delta}(\Theta(x) - \Theta(y)) + \alpha(X - y)\right) dy}$$

$$= -\left[\beta \int_x^X \exp\left(-\frac{D}{\delta}(\Theta(X) - \Theta(y)) + \alpha(X - y)\right) dy\right]^{-1}. \tag{3.53}$$

By choosing x judiciously and exploiting the uniform bounds on the total variation of w given in (3.26), we define $H(X)$:

$$H(X) \equiv \begin{cases} \left[\beta \int_{X-L}^X \exp\left(-\frac{D}{\delta}K_w(X) + \left(\alpha - \frac{D}{\delta}\right)(X - y)\right) dy\right]^{-1} & \text{if } X > L \\ \left[\beta \int_0^X \exp\left(-\frac{D}{\delta}K_w(X) + \left(\alpha - \frac{D}{\delta}\right)(X - y)\right) dy\right]^{-1} & \text{if } X \leq L. \end{cases} \tag{3.54}$$

The proof for (3.52) is completely analogous.

The conclusions of Lemmas 3.1–3.5 are collected in the following theorem, which states that forward characteristics are unique and summarizes the properties of minimal and maximal backward characteristics. In particular, we see the classical 1-characteristic behavior ($\frac{dz}{dx} = g(z,w)$) on extremal 1-characteristics (except possibly at the characteristics' endpoints) and the classical 2-characteristic behavior ($\frac{dw}{dx} = -g(z,w)$) on the extremal 2-characteristics (except possibly at the endpoints).

THEOREM 3.1. *Let (T,X) be any point of the upper half-plane, with $X > 0$. A unique forward* 1*-characteristic emanates from the point. Further, if we define τ and v to be the minimal and maximal backward* 1*-characteristics emanating from (T,X), then for $x \in (0,X)$,*

$$z(\tau(0)+,0) + \int_0^x g(z(\tau(y)\pm,y),w(\tau(y)\pm,y))dy \leq z(\tau(x)+,x)$$

$$= z(T-,X) - \int_x^X g(z(\tau(y)\pm,y),w(\tau(y)\pm,y))dy$$

$$= z(\tau(x)-,x) \leq z(\tau(0)-,0) + \int_0^x g(z(\tau(y)\pm,y),w(\tau(y)\pm,y))dy, \tag{3.55}$$

$$z(v(0)+,0)+\int_0^x g(z(v(y)\pm,y),w(v(y)\pm,y))dy \le z(v(x)+,x)$$

$$= z(T+,X)-\int_x^X g(z(v(y)\pm,y),w(v(y)\pm,y))dy$$

$$= z(v(x)-,x) \le z(v(0)-,0)+\int_0^x g(z(v(y)\pm,y),w(v(y)\pm,y))dy. \tag{3.56}$$

When $z(T-,X)=z(T+,X)$, τ and v coincide.

Similarly, a unique forward 2*-characteristic emanates from (T,X), and if we now define τ and v to be the minimal and maximal backward* 2*-characteristics emanating from (T,X), then for $x\in(0,X)$,*

$$w(\tau(0)+,0)-\int_0^x g(z(\tau(y)\pm,y),w(\tau(y)\pm,y))dy \le w(\tau(x)+,x)$$

$$= w(T-,X)+\int_x^X g(z(\tau(y)\pm,y),w(\tau(y)\pm,y))dy$$

$$= w(\tau(x)-,x) \le w(\tau(0)-,0)-\int_0^x g(z(\tau(y)\pm,y),w(\tau(y)\pm,y))dy \tag{3.57}$$

$$w(v(0)+,0)-\int_0^x g(z(v(y)\pm,y),w(v(y)\pm,y))dy \le w(v(x)+,x)$$

$$= w(T+,X)+\int_x^X g(z(v(y)\pm,y),w(v(y)\pm,y))dy$$

$$= w(v(x)-,x) \le w(v(0)-,0)-\int_0^x g(z(v(y)\pm,y),w(v(y)\pm,y))dy. \tag{3.58}$$

When $w(T-,X)=w(T+,X)$, τ and v coincide.

Proof. From (3.51) and (3.52) in Lemma 3.5 and the expressions for λ and μ in (2.5), we determine that

$$\frac{\lambda(z(t,x),w(t,x))-\lambda(z(s,x),w(s,x))}{t-s} \le AH(x)+B\hat{H}(x), \tag{3.59}$$

$$\frac{\mu(z(t,x),w(t,x))-\mu(z(s,x),w(s,x))}{t-s} \le AH(x)+B\hat{H}(x). \tag{3.60}$$

Since H and $\hat{H}$ are bounded everywhere except at $x=0$, it follows from Filippov's theory [3] that the initial-value problems in (3.5) and (3.7) with datum (T,X) have *unique* solutions in the forward spatial direction if $X>0$. In other words, unique forward 1- and 2-characteristics emanate from (T,X).

We now establish (3.55). From Lemma 3.1, we have that $z(\tau(x)+,x) = z(\tau(x)-,x)$ for almost every x. Combining this fact with (3.47) of Lemma 3.4 implies that $z(\tau(x)+,x) = z(\tau(x)-,x)$ is true for all $x \in (0,X)$. Combining this new result with Lemma 3.4, we see that τ is both the minimal and maximal characteristic emanating from $(\tau(x),x)$ for any $x \in (0,X)$. The last inequality in (3.55) now follows directly from (3.13) of Lemma 3.2, and the first inequality in (3.55) follows directly from (3.14) of Lemma 3.2.

All that remains to be proven in (3.55) is that

$$z(\tau(x)-,x) = z(T-,X) - \int_x^X g(z(\tau(y)\pm,y),w(\tau(y)\pm,y))dy. \tag{3.61}$$

This is established by considering an increasing sequence $\{t_n\}$ such that t_n converges to T and $z(t_n-,X) = z(t_n+,X)$ for all n. We let τ_n denote the unique backward 1-characteristic emanating from (t_n,X). From Lemma 3.4, we know that

$$z(\tau_n(x)\pm,x) = z(t_n,X) - \int_x^X g(z(\tau_n(y)\pm,y),w(\tau_n(y)\pm,y))dy, \quad x \in (0,X]. \tag{3.62}$$

As $n \to \infty$, $z(t_n,X) \to z(T-,X)$ and $\tau_n(x) \to \tau(x)$ uniformly on $(0,X]$. The uniform convergence of τ_n establishes the convergence of the integral:

$$\lim_{n\to\infty} \int_x^X g(z(\tau_n(y)\pm,y),w(\tau_n(y)\pm,y))dy = \int_x^X g(z(\tau(y)-,y),w(\tau(y)-,y))dy. \tag{3.63}$$

The uniform convergence also allows us to state that

$$\lim_{n\to\infty} z(\tau_n(y)\pm,y) = z(\tau(y)-,y). \tag{3.64}$$

Therefore, taking the limit as $n \to \infty$ of (3.62) yields (3.61).

The proofs of (3.56), (3.57), and (3.58) are quite similar and therefore are omitted.

The knowledge we have collected concerning the nature of characteristics is applied to the following three theorems, which describe the structure of solutions to our system. The reader can find the explicit proofs of these theorems in [6]; however, they are omitted here due to their similarity to the proofs used in [2] by Dafermos and Geng for the homogeneous form of (1.1).

THEOREM 3.2. *Let (T,X) be any point on the upper half-plane with $X > 0$. Consider the (unique) forward 1-characteristic σ and the (not necessarily distinct) minimal and maximal backward 1-characteristics, τ_- and τ_+, emanating from (T,X). Define the sets*

$$S_- \equiv \{(t,x) : 0 \le x < X,\ t \le \tau_-(x) \text{ or } x \ge X,\ t \le \sigma(x)\}$$

and

$$S_+ \equiv \{(t,x) : 0 \le x < X,\ t \ge \tau_+(x) \text{ or } x \ge X,\ t \ge \sigma(x)\}.$$

Then the restriction of $z(t-,x)$ to S_- and the restriction of $z(t+,x)$ to S_+ are continuous at (T,X). In particular, z is continuous at (X,T) if and only if $z(T-,X) = z(T+,X)$.

Similarly, if σ is the forward 2-characteristic and τ_- and τ_+ are the minimal and maximal backward 2-characteristics emanating from (T,X), then the restriction of $w(t-,x)$ to the set

$$\Sigma_- \equiv \{(t,x) : 0 \le x < X,\ t \le \tau_-(x) \text{ or } x \ge X,\ t \le \sigma(x)\}$$

and the restriction of $w(t+,x)$ to the set

$$\Sigma_+ \equiv \{(t,x) : 0 \le x < X,\ t \ge \tau_+(x) \text{ or } x \ge X,\ t \ge \sigma(x)\}$$

are continuous at (T,X). In particular, w is continuous at (X,T) if and only if $w(T-,X) = w(T+,X)$.

The next proposition states that once a discontinuity develops, it has to propagate all the way to infinity as a shock.

THEOREM 3.3. *If σ is the (unique) forward 1-characteristic emanating from a point (T,X) of the upper half-plane with $z(T-,X) < z(T+,X)$, then $z(\sigma(x)-,x) < z(\sigma(x)+,x)$ for $0 \le X < x < \infty$.*

Similarly, if σ is the forward 2-characteristic emanating from (T,X), where $w(T-, X) < w(T+,X)$, then $w(\sigma(x)-,x) < w(\sigma(x)+,x)$ for $X < x < \infty$.

A point (T,X) of the upper half-plane will be called a *1-shock generation point* if a forward 1-characteristic σ emanating from (T,X) satisfies $z(\sigma(x)-,x) < z(\sigma(x)+,x)$ for $x \in (X,\infty)$ and none of the backward 1-characteristics emanating from (T,X) contains any point of discontinuity of z. The definition of a 2-shock generation point is completely analogous. By virtue of Theorem 3.3, it is easily seen that if (T,X) is a point of discontinuity of z (or w), then at least one backward 1-characteristic (or 2-characteristic) emanating from (T,X) must pass through a 1-shock (or a 2-shock) generation point.

When (T,X) is a 1-shock generation point, either $z(T-,X) = z(T+,X)$ or $z(T-,X) < z(T+,X)$. In the latter case (T,X) is the *focus* of a *1-compression wave.* Similarly, 2-shock generation points (T,X) may be either points of continuity of w, $w(T-,X) = w(T+,X)$, or foci of 2-compression waves when $w(T-,X) < w(T+,X)$.

The following proposition describes the structure of shocks.

THEOREM 3.4. *Let σ be a 1-shock generated at the point $(\sigma(X),X)$. Consider the four functions $z_\pm(x) \equiv z(\sigma(x)\pm,x)$, $w_\pm(x) \equiv w(\sigma(x)\pm,x)$, defined on $[X,\infty)$. Then the following hold:*

(i) *$z_\pm$ are right-continuous functions with bounded variation locally. For $x > X$,*

$$z_-(x) < z_+(x), \qquad z_-(x-) \ge z_-(x+), \qquad z_+(x-) \le z_+(x+). \tag{3.65}$$

When $z_-(x-) = z_-(x+)$, $(\sigma(x),x)$ is a point of continuity of the restriction of z to the set $\{(t,\xi) : \xi > X, t < \sigma(\xi)\}$; otherwise, $(\sigma(x),x)$ is a point of interaction of σ with another 1-shock or it is the focus of a 1-compression wave impinging from the left. When $z_+(x-) = z_+(x+)$, $(\sigma(x),x)$ is a point of continuity of the restriction of z to the set $\{(t,\xi) : \xi > X,\ t > \sigma(\xi)\}$; otherwise, $(\sigma(x),x)$ is a point of interaction of σ with another 1-shock or it is the focus of a 1-compression wave impinging from the right.

(ii) *$w_\pm$ are functions of bounded variation locally*; *w_- is right-continuous while w_+ is left-continuous. For $x > X$,*

$$w_-(x-) \ge w_-(x+), \quad w_+(x-) \ge w_+(x+),$$

$$w_-(x-) = w_+(x-), \quad w_-(x+) = w_+(x+). \tag{3.66}$$

x is point of discontinuity of $w_\pm$ if $(\sigma(x),x)$ is a point of interaction of σ with a 2-shock.

(iii) *σ is right-differentiable at every $x \ge X$ and*

$$\frac{d^+}{dx}\sigma(x) = \frac{1}{z_-(x)z_+(x)w_-(x)}, \quad X \le x < \infty. \tag{3.67}$$

(iv) *If x is a point of continuity of $z_\pm$ and $w_\pm$, then σ is differentiable at x.*

A similar statement holds for 2-shocks with the roles of z and w appropriately interchanged.

Proof. We present a proof of the claim that $z_\pm$ and $w_\pm$ are functions of bounded variation locally since the proofs given in [2] cannot be extended to the inhomogeneous case considered here.

To show that z_- has bounded variation, we first prove that for any $Y \in [0,\infty)$,

$$TV_{t\in(-\infty,\infty)}w(t,Y) + TV_{t\in(-\infty,\infty)}z(t,Y)$$

$$\le [TV_{t\in(-\infty,\infty)}w(t,0) + TV_{t\in(-\infty,\infty)}z(t,0)]e^{2KY}, \tag{3.68}$$

$$\text{where } K \equiv \max\left[\left\|\frac{\partial g}{\partial w}\right\|_{L^\infty}, \left\|\frac{\partial g}{\partial z}\right\|_{L^\infty}\right].$$

We know that K is finite from the expressions for the partial derivatives for g given in (2.11) combined with the restricted variation of z and w in (2.7) and the fact that we are bounded away from the umbilic point where $w = z$. To show (3.68), we pick any mesh $\cdots s_{i-1} < S_{i-1} < s_i < S_i < s_{i+1} < S_{i+1}\cdots$ with the property that $\cdots z(s_i,Y) > z(S_i,Y) < z(s_{i+1},Y)\cdots$. By virtue of (3.9), it suffices to consider only meshes with $z(s_i-,Y) = z(s_i+,Y)$ and $z(S_i-,Y) = z(S_i+,Y)$. Now define τ_i to be the (unique) backward 1-characteristic emanating from (s_i,Y) and v_i to be the (unique) backward 1-characteristic emanating from (S_i,Y). From Theorem 3.1, we know that

$$z(s_i,Y) - z(S_i,Y) \le z(\tau_i(0)-,0) - z(v_i(0)+,0)$$

$$+\int_0^Y g(z(\tau_i(x),x),w(\tau_i(x),x)) - g(z(v_i(x),x),w(v_i(x),x))dx \tag{3.69}$$

and

$$z(s_i,Y) - z(S_{i-1},Y) \le z(\tau_i(0)-,0) - z(v_{i-1}(0)+,0)$$

$$+\int_0^Y g(z(\tau_i(x),x),w(\tau_i(x),x)) - g(z(v_{i-1}(x),x),w(v_{i-1}(x),x))dx. \tag{3.70}$$

Since forward characteristics are unique, we know that $\cdots\tau_i(0) \le v_i(0) \le \tau_{i+1}(0)\cdots$. Therefore, when we sum (3.69) and (3.70) for all i and apply the chain rule to the resulting equation's integrands, we obtain

$$TVz(Y) \le TVz(0) + \int_0^Y K(TVz(x) + TVw(x))dx, \tag{3.71}$$

where $TVz(x) \equiv TV_{t\in(-\infty,\infty)}z(t,x)$ and $TVw(x) \equiv TV_{t\in(-\infty,\infty)}w(t,x)$. We can repeat this entire process for w(.,Y) and add the result to (3.71), yielding

$$TVz(Y) + TVw(Y) \le TVz(0) + TVw(0) + \int_0^Y 2K[TVz(x) + TVw(x)]dx. \tag{3.72}$$

(3.68) follows from (3.72) by Gronwall's inequality.

Now we perform a similar process on σ. We pick any mesh $\cdots y_{i-1} < \xi_{i-1} < y_i < \sigma_i < y_{i+1} < \xi_{i+1} \cdots$ with the property that $z_-(y_i) > z_-(\xi_i) < z_-(y_{i+1})$. We redefine τ_i to be the minimum backward 1-characteristic emanating from $(\sigma(y_i), y_i)$ and v_i to be the minimum backward 1-characteristic emanating from $(\sigma(\xi_i), \xi_i)$. From Theorem 3.1, we know that

$$z_-(y_i) - z_-(\xi_i) \leq z(\tau_i(0)-, 0) - z(v_i(0)+, 0)$$

$$+ \int_0^{y_i} g(z(\tau_i(x), x), w(\tau_i(x), x)) - g(z(v_i(x), x), w(v_i(x), x))dx$$

$$+ \int_{y_i}^{\xi_i} g(z(\tau_i(x), x), w(\tau_i(x), x))dx \tag{3.73}$$

and

$$z_-(y_i) - z_-(\xi_{i-1}) \leq z(\tau_i(0)-, 0) - z(v_{i-1}(0)+, 0)$$

$$+ \int_0^{y_i} g(z(\tau_i(x), x), w(\tau_i(x), x)) - g(z(v_{i-1}(x), x), w(v_{i-1}(x), x))dx$$

$$+ \int_{y_i}^{\xi_i} g(z(\tau_i(x), x), w(\tau_i(x), x))dx. \tag{3.74}$$

As before, we have that $\cdots \tau_i(0) \leq v_i(0) \leq \tau_{i+1}(0) \cdots$, so when we sum (3.73) and (3.74) for all i and apply the chain rule, we obtain

$$TV_{x \in [0,Y]} z_-(x) \leq TVz(0) + \int_0^Y K(TVz(x) + TVw(x))dx + Y||g||_{L^\infty}. \tag{3.75}$$

Now we use (3.68), yielding

$$TV_{x \in [0,Y]} z_-(x) \leq TVz(0) + \frac{1}{2}(e^{2KY} - 1)[TVz(0) + TVw(0)] + Y||g||_{L^\infty}. \tag{3.76}$$

Therefore, the total variation of z_- is bounded locally.

The proofs showing that z_+, w_-, and w_+ have local bounded variation are similar and thus will be omitted.

The proofs of the remaining assertions follow [2] closely and are thus omitted. Again, the reader is directed to [6] for the exact form of these proofs.

4. Asymptotic behavior of the periodic case. With our knowledge of generalized characteristics and the structure of the solution from §3, we are now prepared to establish the asymptotic behavior of the chemical concentrations. We will begin by locating invariant regions of the (z, w) phase plane. Then we will show that these invariant regions shrink exponentially to a single point located on the equilibrium curve when the chemical components are cyclically fed into the chromatography reactor (i.e., periodic conditions at $x = 0$). Finally, we use an integrated form of the mass balance to show which point of the equilibrium curve is approached under these periodic conditions.

We start with the assumption that there exists some spatial location X where $\sup_{t\in(-\infty,\infty)} w(t,X) < b/a$. This assumption allows us to state that $w(t,X)$ is in the domain of the equilibrium curve function $z_e(w)$ given in (2.9). Once we establish the invariant regions, this assumption will also imply that $w(t,x)$ is also in the domain of $z_e(w)$ for all $x > X$.

To determine these invariant regions, we will need the following lemma, which states that as an extremal 1-characteristic passes through a spatial interval where w is constant, the value of z on the 1-characteristic exponentially approaches the equilibrium value associated with that constant state. An analogous result holds for the 2-characteristic.

LEMMA 4.1. *Define $z(x) \equiv z(\tau(x)-,x)$ and $w(x) \equiv w(\tau(x)-,x)$, where τ is the minimal backward 1-characteristic which emanates from a point (T,X) in the upper half plane. If $0 < y \leq Y \leq X$ and there is a constant $W < b/a$ such that $w(x) = W$ for all $x \in [y,Y]$, then there are positive constants $\alpha_1 < \alpha_2$ such that*

$$z(y) \geq z_e(W)$$

implies

$$z_e(W)+(z(y)-z_e(W))e^{-\alpha_2(Y-y)} \leq z(Y) \leq z_e(W)+(z(y)-z_e(W))e^{-\alpha_1(Y-y)} \tag{4.1a}$$

and

$$z(y) \leq z_e(W)$$

implies

$$z_e(W)+(z(y)-z_e(W))e^{-\alpha_1(Y-y)} \leq z(Y) \leq z_e(W)+(z(y)-z_e(W))e^{-\alpha_2(Y-y)}. \tag{4.1b}$$

The above equations also hold if we define $z(x) \equiv z(\tau(x)+,x)$ and $w(x) \equiv w(\tau(x)+,x)$ and if τ is the maximal backward 1-characteristic which emanates from (T,X).

Similarly, if we define $z(x) \equiv z(v(x)-,x)$ and $w(x) \equiv w(v(x)-,x)$, where v is the minimal backward 2-characteristic which emanates from (T,X), and if there is a constant, Z, such that $z(x) = Z$ for $0 < y < x < Y \leq X$, then

$$w(y) \geq w_e(Z)$$

implies

$$w_e(Z)+(w(y)-w_e(Z))e^{-\alpha_2(Y-y)} \leq w(Y) \leq w_e(Z)+(w(y)-w_e(Z))e^{-\alpha_1(Y-y)} \tag{4.1c}$$

and

$$w(y) \leq w_e(Z)$$

implies

$$w_e(Z)+(w(y)-w_e(Z))e^{-\alpha_1(Y-y)} \leq w(Y) \leq w_e(Z)+(w(y)-w_e(Z))e^{-\alpha_2(Y-y)}. \tag{4.1d}$$

As before, the above equations also hold if we define $z(x) \equiv z(v(x)+,x)$ and $w(x) \equiv w(v(x)+,x)$ and if v is the maximal backward 2-characteristic which emanates from (T,X).

Proof. We establish (4.1a). From Theorem 3.1, we know that

$$z(Y) = z(y) + \int_y^Y g(z(x), w(x))dx, \tag{4.2}$$

and therefore, after substituting the definitions of $g(z, w)$ and $z_e(w)$, we have

$$z(Y) = z(y) + \int_y^Y \frac{aW - b}{z(x) - W}[z(x) - z_e(W)]dx. \tag{4.3}$$

From the range of z and w given in (2.7) and the constraint that $W < b/a$, we know that there exist positive constants $C_1 < C_2$, where $-C_2 \le (aW - b)/(z(x) - W) \le -C_1$. If we differentiate (4.3), insert these bounding constants, and solve the resulting differential inequality, we obtain (4.1a) with C_1 and C_2 instead of α_1 and α_2.

We use this same method to obtain (4.1b)–(4.1d) (each with different C_1 and C_2 values in place of α_1 and α_2). Finally, we define α_1 as the minimum of the four C_1's and α_2 as the maximum of the four C_2's.

This same proof (with the same values for α_1 and α_2) holds for the maximal characteristic cases.

Now we can establish the invariant region.

THEOREM 4.1. *When there exists an X such that $\sup_{t\in(-\infty,\infty)} w(t, X) < b/a$, then we can define a closed rectangle I in the (z, w) phase plane which encloses the solution at $x = X$ and is bounded by the lines*

$$w = W_1 \equiv \max\left[\sup_t w(t, X), w_e\left(\sup_t z(t, X)\right)\right],$$

$$w = W_0 \equiv \min\left[\inf_t w(t, X), w_e\left(\inf_t z(t, X)\right)\right],$$

$$z = Z_1 \equiv z_e(W_1) = \max\left[\sup_t z(t, X), z_e\left(\sup_t w(t, X)\right)\right],$$

$$z = Z_0 \equiv z_e(W_0) = \min\left[\inf_t z(t, X), z_e\left(\inf_t w(t, X)\right)\right].$$

This rectangle is invariant. That is, $(z(t, x), w(t, x)) \in I$ for all $x \ge X$.

Proof. We will prove the invariance of the open rectangles $I_n \equiv \{(z, w) : z \in (Z_0 - 1/n, Z_1 + 1/n),\ w \in (w_e(Z_0 - 1/n), w_e(Z_1 + 1/n))\}$. The invariance of I_n implies the invariance of I since $I \subset I_n$ and $I = \bigcap_{n=1}^{\infty} I_n$.

If I_n is not invariant, then there is a point (T, Y), where $Y > X$, such that $(z(t, x), w(t, x)) \in I_n$ for $x \in [X, Y)$ but $(z(T, Y), w(T, Y)) \in \partial I_n$. Without loss of generality, we will assume the case $z(T, Y) = Z_0 - 1/n$. From Theorem 3.1 we know that

$$z(Y) = z(X) + \int_X^Y g(z(\xi), w(\xi))d\xi, \tag{4.4}$$

where τ is the minimal backward 1-characteristic emanating from (T, Y), $z(x) \equiv z(\tau(x)-, x)$, and $w(x) \equiv w(\tau(x)-, x)$. From (2.11a), we know that $\frac{\partial g}{\partial w} > 0$; therefore, decreasing $w(\xi)$ decreases the right-hand side of (4.4):

$$z(Y) \ge z(X) + \int_X^Y g\left(z(\xi), w_e\left(Z_0 - \frac{1}{n}\right)\right) d\xi. \tag{4.5}$$

The right-hand side of (4.5) is equal to the value that $Z(Y)$ would have to take if $\forall x \in [X,Y], w(x) = w_e(Z_0 - 1/n)$. Since $Z(X) \geq Z_0 - 1/n$, we can apply Lemma 4.1 to the right-hand side of 4.5 to obtain

$$z(Y) \geq Z_0 - \frac{1}{n} + \left[z(X) - \left(Z_0 - \frac{1}{n}\right)\right] e^{-\alpha_2(Y-X)}. \tag{4.6}$$

Since, from entropy, we must have that $z(T,Y) \geq z(T-,Y) \equiv z(Y)$, we conclude from (4.6) that $Z_0 - 1/n = z(X)$. But this contradicts the fact that $z(X) \geq Z_0$. Therefore, I_n and thereby I are invariant.

The fact that I is invariant is not particularly surprising given the nature of the vector fields $\frac{dz}{dx} = g(z,w)$ and $\frac{dw}{dx} = -g(z,w)$ shown in Fig. 2.

We now consider the case where the conditions at $x = 0$ are periodic in t with period L, which, of course, induces a solution which is periodic in t for any fixed x. We will show that over a fixed interval of space, the invariant region associated with this solution shrinks in proportion to its own size, which implies that the solution must exponentially approach a single point on the equilibrium curve.

We assume that at some location X, the solution is close enough to the equilibrium curve so that $\sup_{t\in(-\infty,\infty)} w(t,X) < b/a$, and therefore, from Theorem 4.1, we can define an invariant region bounded by the lines $w = W_0$, $w = W_1$, $z = Z_0 = z_e(W_0)$, and $z = Z_1 = z_e(W_1)$. We also consider some arbitrary number $x_0 > X$ and define the lines that bound the (smaller) invariant region at $x = x_0$: $w = w_0$, $w = w_1$, $z = z_0 = z_e(w_0)$, and $z = z_1 = z_e(w_1)$.

In the beginning of §3, we assumed that the characteristic speeds have small oscillations and are well separated. We reexpress this condition by defining s_1, s_2, s_3, and s_4 as explicit bounds on the speeds:

$$\lambda \in [s_1, s_2], \qquad \mu \in [s_3, s_4],$$

$$\text{where } 0 < s_1 \leq s_2 < s_3 \leq s_4 \quad \text{and} \quad \frac{(s_2 - s_1) + (s_4 - s_3)}{s_3 - s_2} \ll 1. \tag{4.7}$$

From this, we can define the spatial interval $H \equiv L/(s_3 - s_2)$, which will be of particular interest to us. H has the property that any backward 1-characteristic emanating from an arbitrary point $(t, x+H)$ must intersect each backward 2-characteristic emanating from the period of points between $(t, x+H)$ and $(t+L, x+H)$ as the 1-characteristic passes through the $[x, x+H]$ region.

We will show that as x progresses through the interval $[x_0, x_0 + 7H + \ln 2/\alpha_2]$, the invariant region must proportionally shrink. To simplify notation, we translate the x-coordinate system so as to set $x_0 = -4H$. (Therefore, the original invariant region is now located at $x = -(x_0 + 4H - X)$ and we wish to show proportional shrinkage as we pass through the interval $[-4H, 3H + \ln 2/\alpha_2]$.)

Our proof is basically by contradiction. If there is no proportional shrinking, then there is at least one backward 1-characteristic emanating from the line $x = 3H$ that stays near the equilibrium point (z_1, w_1) on the phase plane as the trajectory travels back to $x = -4H$, and there is a second backward 1-characteristic emanating from a later time on the $x = 3H$ line that stays near the equilibrium point (z_0, w_0) as it travels back to $x = -4H$. The value of w on a 2-characteristic will not be able to change too drastically as the 2-characteristic progresses between these 1-characteristics. Therefore, the 2-characteristics with the property that w is near

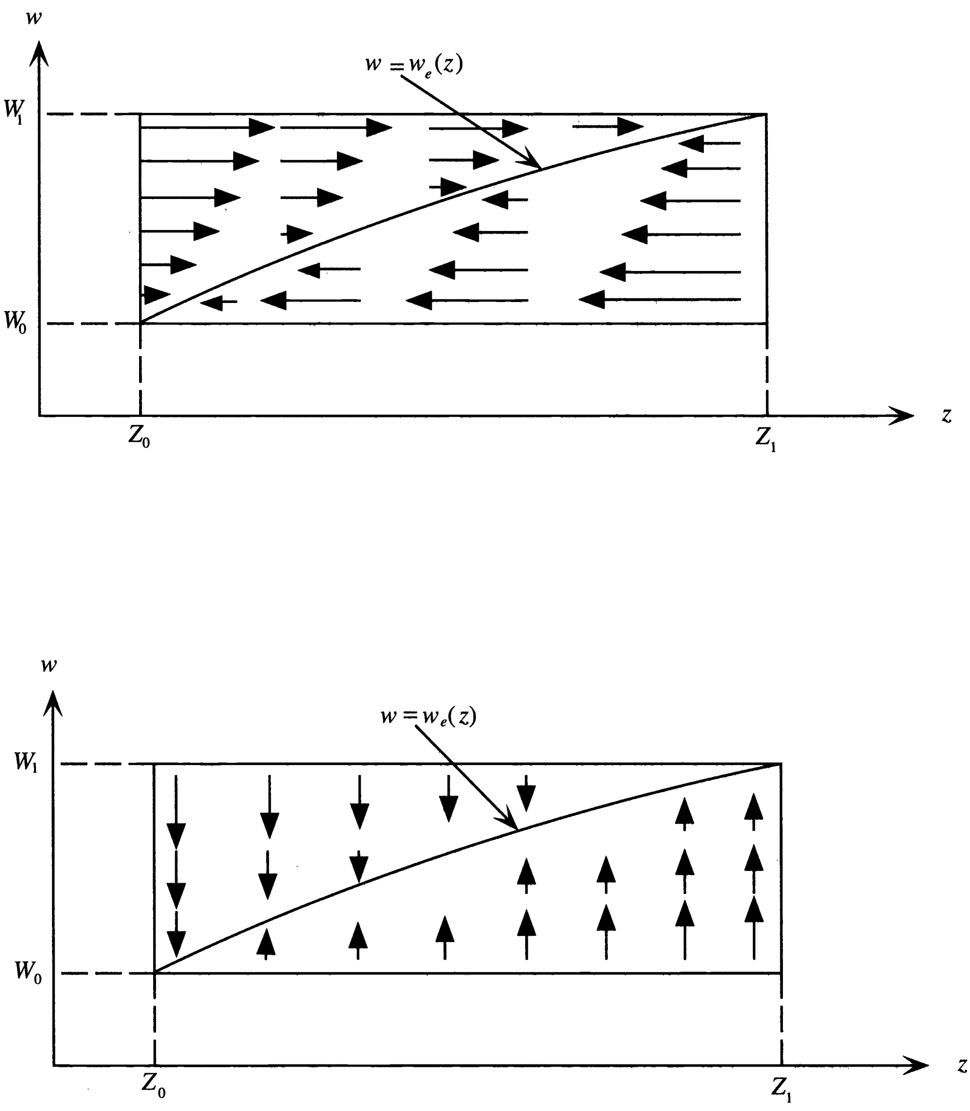


FIG. 2. *The vector fields for classical characterisic behavior. In* (2a)*, we have the vector field* $\frac{dz}{dx} = g(z, w)$*. In* (2b)*, we have the vector field* $\frac{dw}{dx} = -g(z, w)$*.*

w_0 as they cross the first 1-characteristic must significantly expand in measure before they reach the second 1-characteristic. By comparing the slopes over the region $x \in [-4H, 3H]$ of 2-characteristics where w is near w_0 to 2-characteristics where w is near w_1, we will see that this expansion cannot occur. We will conclude in Theorem 4.2 that the range of z values must proportionally shrink over the spatial interval $7H$. This shrinkage of z will induce a proportional shrinkage of the range of w values as we progress through an additional interval of length $\ln 2/\alpha_2$, as will be shown in Theorem 4.3.

We begin to formalize these ideas. First, we label the bounds on the range of values that z takes on the line $x = 3H$:

$$z_M \equiv \sup_t z(t,3H), \qquad z_m \equiv \inf_t z(t,3H). \tag{4.8}$$

We wish to show that there exists some constant $\varepsilon' > 0$ that depends *only* on the period L, the bounds on the characteristic speeds s_i, and the reference invariant region (i.e. Z_0, Z_1, W_0, and W_1) such that $(z_M - z_m)/(z_1 - z_0) < 1 - \varepsilon'$.

To select the appropriate backward 1-characteristics that we wish to analyze, we first choose any two points t_m and t_M which satisfy the conditions

$$\lim_{t\to t_M^+} z(t,3H) = z_M, \qquad \lim_{t\to t_m^-} z(t,3H) = z_m,$$

$$\text{and } t_m - t_M \in (L, 2L]. \tag{4.9}$$

Now we define τ_M as the maximal backward 1-characteristic emanating from $(t_M, 3H)$ and τ_m as the minimal backward 1-characteristic emanating from $(t_m, 3H)$. We also define v_m, the minimal backward 2-characteristic emanating from $(t_m, 3H)$, and v_M, the (unique) forward 2-characteristic emanating from $(\tau_M(0), 0)$. From this construction, we find that the location $\tilde{\xi}$, where $\tau_M(\tilde{\xi}) = v_m(\tilde{\xi})$ (i.e., the two characteristics cross), must have the property that $\tilde{\xi} \geq H$. We also find that the location $\hat{\xi}$, which corresponds to the x coordinate where the τ_m and v_M characteristics cross, must have the property that $3H - \hat{\xi} \geq H$. This implies that the measure of the space under which τ_M is influenced by the 2-characteristics located between v_M and v_m must be no smaller than H; similarly, the measure of the space under which τ_m is influenced by the 2-characteristics between v_M and v_m is also no smaller than H (see Fig. 3).

To investigate how requiring z_M to be close to z_1 affects $w(\tau_M(x), x)$, we will require a lemma whose result is dependent on the nature of the equilibrium curve. From (2.10a), it is clear that there are constants $E_2 \geq E_1 > 0$ that bound the slope of the equilibrium curve: $E_1 \leq \frac{dz_e}{dw} \leq E_2$. The values of these constants depend only on Z_0, Z_1, W_0, and W_1. Since w_e is the inverse of z_e, we also have that $1/E_1 \geq \frac{dw_e}{dz} \geq 1/E_2$. Finally, it will be convenient to define $E \equiv E_1/E_2$. Now we present a lemma whose proof will be used to establish a number of useful results.

LEMMA 4.2. *If we define the function $\gamma(x) \in [0,1]$ by the relation $w(\tau_M(x)+, x) \equiv w_1 - \gamma(x)(w_1 - w_0)$ and define $z(x) \equiv z(\tau_M(x)+, x)$, then*

$$\frac{z_1 - z(X_2)}{z_1 - z_0} \geq E\alpha_1 \int_{X_1}^{X_2} \gamma(x) e^{-\alpha_2(X_2 - x)} dx, \tag{4.10}$$

where $-4H \leq X_1 \leq X_2$. This implies that if $z(X_2)$ is near z_1, then the measure of x values where $w(\tau_M(x)+, x)$ fails to be close to w_1 is small.

Proof. We begin by comparing the actual solution $z(x)$ of (3.56), where τ_M is under the influence of $\gamma(x)$, with the solution $z_I(x)$ of (3.56) that would occur if τ_M were under the influence of a function $\gamma_I(x)$, where $\gamma_I(x) \leq \gamma(x)$. If both solutions start at the same point, $z(X_1) = z_I(X_1) = Z$, then the fact from (2.11a) that $\frac{\partial g}{\partial w} > 0$ implies that $z(x) \leq z_I(x)$ for all $x \geq X_1$. In other words, making $\gamma(x)$ smaller makes $z(x)$ bigger as x increases.

With this in mind, we divide the region $[X_1, X_2]$ into pieces of length Δx and define

$$\gamma^j \equiv \inf\{\gamma(x) : x \in [X_1 + (j-1)\Delta x, X_1 + j\Delta x]\}. \tag{4.11}$$

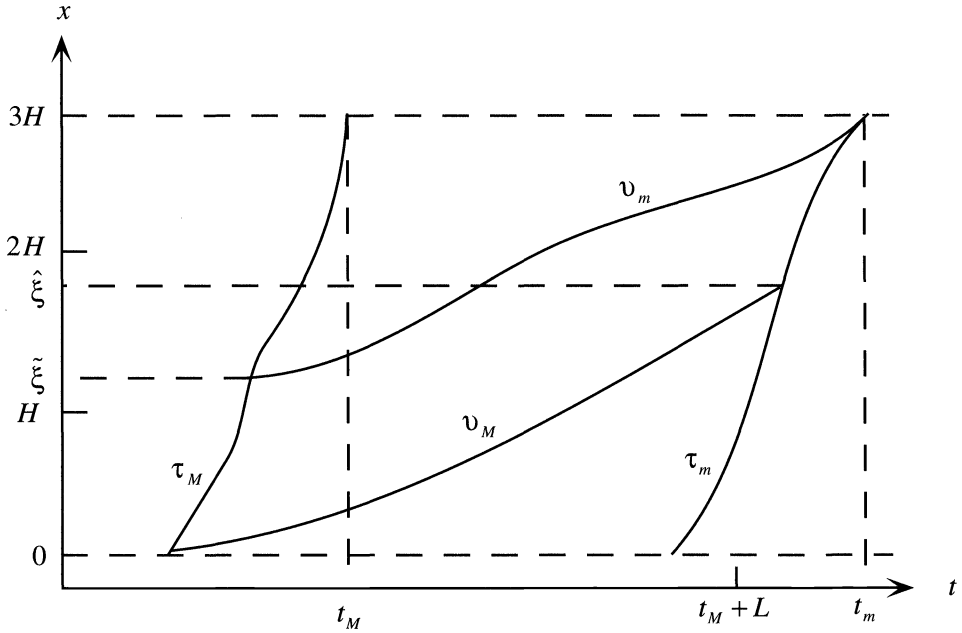


FIG. 3. *The characteristics τ_M, τ_m, v_M, and v_m on the (t, x) plane. We will monitor the variation of z in the regions of τ_M and τ_m located between v_M and v_m.*

We also define

$$z_e^j \equiv z_e(w_1 - \gamma^j(w_1 - w_0)), \tag{4.12}$$

$$z^j \equiv z(X_1 + j\Delta x), \tag{4.13}$$

$$\text{and } \alpha^j \equiv \begin{cases} \alpha_1 & \text{if } z^{j-1} - z_e^j \geq 0, \\ \alpha_2 & \text{if } z^{j-1} - z_e^j < 0, \end{cases} \tag{4.14}$$

where α_1 and α_2 are defined in Lemma 4.1 and depend only on Z_0, Z_1, W_0, and W_1. We use Lemma 4.1 combined with the fact that making $\gamma(x)$ smaller makes $z(x)$ bigger to obtain

$$z^1 \leq z_e^1 + (Z - z_e^1)e^{-\alpha^1 \Delta x}. \tag{4.15}$$

Using the fact that $Z \leq z_1$, which follows from Theorem 4.1, and that $z_e^j \leq z_1 - E\gamma^j(z_1 - z_0)$, we can transform (4.15) into

$$z^1 \leq z_1 - E\gamma^1(z_1 - z_0)(1 - e^{-\alpha^1 \Delta x}). \tag{4.16}$$

Repeating this method and using (4.16), we can obtain an expression for z^2:

$$z^2 \leq z_1 - E(z_1 - z_0)[\gamma^2(1 - e^{-\alpha^2 \Delta x}) + \gamma^1(1 - e^{-\alpha^1 \Delta x})e^{-\alpha^2 \Delta x}] \tag{4.17}$$

and, continuing in this fashion, we obtain an expression for z^n:

$$z^n \leq z_1 - E(z_1 - z_0) \sum_{i=1}^{n} \left[\gamma^i (1 - e^{-\alpha^i \Delta x}) \exp\left[-\Delta x \sum_{j=1}^{n-i} \alpha^{(n-j+1)} \right] \right]. \tag{4.18}$$

By using the definition of α^j, we can simplify (4.18):

$$z^n \leq z_1 - E(z_1 - z_0) \sum_{i=1}^{n} \left[\gamma^i(1 - e^{-\alpha_1 \Delta x}) \exp[-\Delta x \alpha_2 (n-i)]\right]. \tag{4.19}$$

By the nature of BV solutions, the number of 2-shocks is countable; therefore, $\gamma(x)$ can be discontinuous only on a set of measure zero, which implies that $\gamma(x)$ is Riemann integrable. Therefore, as $\Delta x \to 0$, (4.19) yields the desired integral, and, after some simple algebra, we obtain (4.10).

Now we are prepared to quantify the effect on $w(\tau_M(x)+, x)$ when z_M is near z_1.

LEMMA 4.3. *If we define* $\varepsilon_M \equiv \sqrt{(z_1 - z_M)/(z_1 - z_0)}$ *and define the region of* τ_M *between* v_M *and* v_m *where* w *is bounded away from* w_1 *by* $A_{\varepsilon_M} \equiv \{x \in [0, \tilde{\xi}] : w(\tau_M(x)+, x) \in [w_0, w_1 - \varepsilon_M(w_1 - w_0)]\}$, *then the Lebesgue measure,* m, *of* A_{ε_M} *is bounded by* ε_M; *specifically,*

$$\frac{e^{3H\alpha_2}}{E\alpha_1} \varepsilon_M \geq m(A_{\varepsilon_M}). \tag{4.20}$$

Proof. From Lemma 4.2, we have that

$$(\varepsilon_M)^2 = \frac{z_1 - z_M}{z_1 - z_0} \geq E\alpha_1 e^{-3H\alpha_2} \int_0^{3H} \gamma(x)dx. \tag{4.21}$$

By the definition of A_{ε_M}, we have that $\int_0^{3H} \gamma(x)dx \geq \varepsilon_M m(A_{\varepsilon_M})$, and therefore, (4.21) yields

$$\frac{e^{3H\alpha_2}}{E\alpha_1} (\varepsilon_M)^2 \geq \varepsilon_M m(A_{\varepsilon_M}), \tag{4.22}$$

and the result of the lemma is obvious.

Using a similar process, we can analyze the behavior on τ_m.

LEMMA 4.4. *If we define* $\varepsilon_{\bar{m}} \equiv \sqrt{(z_m - z_0)/(z_1 - z_0)}$ *and define the region of* τ_m *between* v_M *and* v_m *where* w *is close to* w_0 *by* $A_{\varepsilon_m} \equiv \{x \in [\hat{\xi}, 3H] : w(\tau_m(x)+, x) \in [w_0, w_0 + \varepsilon_m(w_1 - w_0)]\}$, *then the parameter* ε_m *is bounded by* $\varepsilon_{\bar{m}}$; *specifically,*

$$\frac{e^{3H\alpha_2}}{E\alpha_1(H - m(A_{\varepsilon_m}))} (\varepsilon_{\bar{m}})^2 \geq \varepsilon_m. \tag{4.23}$$

Proof. The proof is similar to the proof of Lemma 4.3. The method used to establish Lemma 4.2 can be reemployed to verify the following analogous result:

$$(\varepsilon_{\bar{m}})^2 = \frac{z_m - z_0}{z_1 - z_0} \geq E\alpha_1 e^{-3H\alpha_2} \int_0^{3H} \gamma(x)dx, \tag{4.24}$$

where $\gamma(x)$ is defined by $w(\tau_m(x)+, x) \equiv w_0 + \gamma(x)(w_1 - w_0)$. By the definition of A_{ε_m}, we have that

$$\int_0^{3H} \gamma(x)dx \geq \varepsilon_m(3H - \hat{\xi} - m(A_{\varepsilon_m})), \tag{4.25}$$

which, when combined with the fact that $3H - \hat{\xi} \geq H$, can be inserted into (4.24) to yield the result of the theorem.

With the above definition of ε_M and $\varepsilon_{\bar{m}}$, we see that our goal of finding $\varepsilon' > 0$ such that $(z_M - z_m)/(z_1 - z_0) < 1 - \varepsilon'$ is equivalent to finding $\varepsilon' > 0$ such that $\varepsilon_M^2 + \varepsilon_{\bar{m}}^2 > \varepsilon'$.

Next, we use Lemma 4.1 to determine some bounds on the evolution of a Riemann invariant as it progresses on an extremal characteristic of its own family.

LEMMA 4.5. *Consider any maximal backward* 2*-characteristic v which emanates from a point (T, X), where $X > -4H$. If we define $w(x) \equiv w(v(x)+, x)$, then for $-4H \leq y \leq Y \leq X$, we have*

$$w(Y) \geq w_0 + (w(y) - w_0)e^{-\alpha_2(Y-y)}, \tag{4.26a}$$

$$w(Y) \leq w_1 + (w(y) - w_1)e^{-\alpha_2(Y-y)}. \tag{4.26b}$$

(4.26) *also holds if v is a minimal backward* 2*-characteristic and $w(x) \equiv w(v(x)-, x)$.*

Similarly, if we consider any maximal backward 1*-characteristic τ which emanates from the point (T, X) and we define $z(x) \equiv z(\tau(x)+, x)$, then for $-4H \leq y \leq Y \leq X$, we have*

$$z(Y) \geq z_0 + (z(y) - z_0)e^{-\alpha_2(Y-y)}, \tag{4.27a}$$

$$z(Y) \leq z_1 + (z(y) - z_1)e^{-\alpha_2(Y-y)}. \tag{4.27b}$$

(4.27) *also holds if τ is a minimal backward* 1*-characteristic and $z(x) \equiv z(\tau(x)-, x)$.*

Proof. We prove (4.26a). From Theorem 3.1, we know that

$$w(Y) = w(y) - \int_y^Y g(z(v(x)+, x), w(x))dx. \tag{4.28}$$

From (2.11b), we know that $\frac{\partial g}{\partial z} < 0$; therefore, decreasing $z(v(x)+, x)$ in (4.28) decreases the right-hand side of (4.28):

$$w(Y) \geq w(y) - \int_y^Y g(z_0, w(x))dx. \tag{4.29}$$

Lemma 4.1 implies that

$$w(y) - \int_y^Y g(z_0, w(x))dx \geq w_0 + (w(y) - w_0)e^{-\alpha_2(Y-y)}, \tag{4.30}$$

which we combine with (4.29) to obtain (4.26a).

The remaining statements in the lemma have analogous proofs.

To prove the proportional shrinkage of the invariant regions (i.e., to find an $\varepsilon' > 0$ such that $\varepsilon_M^2 + \varepsilon_{\bar{m}}^2 > \varepsilon'$), we require only one more result, which will be the goal of Lemmas 4.6 and 4.7: we need to show that if ε_M and ε_m are in a sufficiently small neighborhood of zero, which we will call $[0, \bar{\varepsilon}]$, then there is a constant C such that $m(A_{\varepsilon_M}) \geq Cm(A_{\varepsilon_m})$. In other words, the measure of the x values where $w(\tau_M(x)+, x)$ is not near w_1 bounds the measure of the x values where $w(\tau_m(x)+, x)$ is near w_0.

The proofs of Lemmas 4.6 and 4.7 will make use of two new small-valued variables, ε_1 (which will be closely related to ε_m) and ε_2 (which will be closely related to ε_M).

The proofs will also require the following construction, which divides the region of the (t, x) plane bounded by τ_M, τ_m, $x = 0$, and $x = 3H$ into a countable number of "blocks."

We start with $x_1^t \equiv \max\{x : x \leq 3H \text{ and } x \in A_{\varepsilon_m}\}$. The point $(\tau_m(x_1^t), x_1^t)$ is (generally) located near the top of the τ_m characteristic. We define v_1^t to be the maximal backward 2-characteristic emanating from $(\tau_m(x_1^t), x_1^t)$ and follow it back until it intersects τ_M. We define the x coordinate of this point of intersection as X_1^t. Now we move down τ_M until we hit the point $(\tau_M(X_1^c), X_1^c)$, where $X_1^c \equiv \sup\{x : x < X_1^t \text{ and } x \in \text{the complement of } A_{\varepsilon_M}\}$. Next, we follow v_1^c, the forward characteristic emanating from $(\tau_M(X_1^c), X_1^c)$, until it intersects τ_m, and we define x_1^c as the x coordinate of the point of intersection. This completes the first block. Now we define $x_2^t \equiv \max\{x : x \leq x_1^c \text{ and } x \in A_{\varepsilon_m}\}$ and continue constructing these blocks until we reach $x = 0$. Because the variation of w on 1-characteristics is locally bounded, there can be at most a countably infinite number of these blocks.

Since $x_i^t \in A_{\varepsilon_m}$, we can use (4.26a) of Lemma 4.5 to obtain

$$w(\tau_M(X_i^t)+, X_i^t) \leq w_0 + \varepsilon_1^2(w_1 - w_0), \tag{4.31}$$

where $\varepsilon_1 \equiv \sqrt{\varepsilon_m e^{3H\alpha_2}}$. The form of the construction also provides the following facts:

$$A_{\varepsilon_m} \subset \bigcup_i [x_i^c, x_i^t], \qquad \bigcup_i [X_i^c, X_i^t] \subset A_{\varepsilon_M}, \tag{4.32}$$

$$x_i^c < x_i^t, \qquad X_i^c < X_i^t. \tag{4.33}$$

We wish to find some constant C, where $x_i^t - x_i^c \leq C(X_i^t - X_i^c)$. This will allow us to use (4.32) to show that the measure of A_{ε_M} bounds the measure of A_{ε_m}. To find C, we will need to add some additional elements to our construction. We define $X_i^b \equiv \sup\{x : x < X_i^c \text{ and } w(\tau_M(x)+, x) \leq w_0 + \varepsilon_1(w_1 - w_0)\}$. X_i^b has the property $X_{i+1}^t < X_i^b < X_i^c$. We also define $v_i^b(x)$ as the forward 2-characteristic emanating from $(\tau_M(X_i^b), X_i^b)$ when $x \geq X_i^b$ and as the maximal backward 2-characteristic emanating from $(\tau_M(X_i^b), X_i^b)$ when $x < X_i^b$. The x coordinate of the point of intersection of $v_i^b(x)$ with $\tau_m(x)$ will be called x_i^b. Finally, we extend the previous definition of v_i^c by defining v_i^c as the maximal backward 2-characteristic emanating from $(\tau_M(X_i^c), X_i^c)$ when $x \leq X_i^c$.

Note that in this construction, the capital letter X is used to denote locations that are significant on the τ_M characteristic, and the lower-case letter x denotes important locations on the τ_m characteristic. The superscripts t, c, and b refer to the "top," "center," and "bottom" of a block. A constructed block along with some of its properties is shown in Fig. 4.

By analyzing v_i^c and v_i^b, we will determine bounds on the slope of v_i^c. With these bounds in place, we will analyze v_i^t and v_i^c, which will yield the desired relationship: $x_i^t - x_i^c \leq C(X_i^t - X_i^c)$. The bounds on the slope of v_i^c will come from the following lemma.

LEMMA 4.6. *If ε_m and ε_M are sufficiently small, then $v_i^c(x)$ cannot intersect $v_i^b(x)$ at any $x \in [-4H, 3H]$.*

Proof. We wish to show that the minimum possible value of $v_i^c(x) - v_i^b(x)$ is bounded away from zero in the region $x \in [0, 3H]$. To accomplish this, we consider the largest possible values for the slope (i.e., derivative) of $v_i^c(x)$ and the smallest possible values for the slope of $v_i^b(x)$ in the region $x \in [-4H, 3H]$.

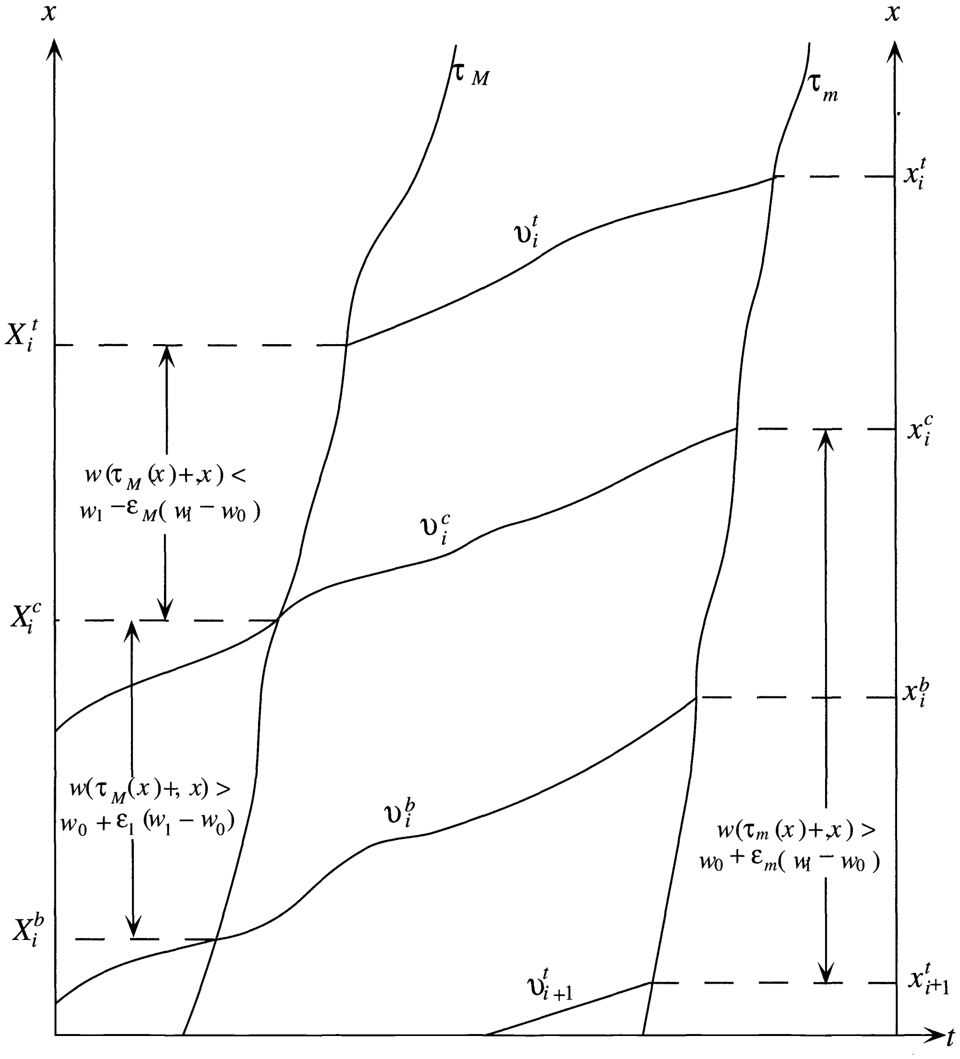


FIG. 4. *A constructed block in the region between τ_M and τ_m. w is bounded away from w_1 in the region of τ_M located between v_i^t and v_i^c, whereas w is bounded away from w_0 in the region of τ_M located between v_i^c and v_i^b. w is also bounded away from w_0 in the region of τ_m located between v_i^c and v_{i+1}^t.*

First, we analyze $v_i^c(x)$ in the region $x \in [-4H, X_i^c]$. From (2.5) and (3.12), we have

$$v_i^c(X_i^c) - v_i^c(-4H) = \int_{-4H}^{X_i^c} \frac{1}{w^2(x)z(x)}dx, \tag{4.34}$$

where $w(x) \equiv w(v_i^c(x)\pm, x)$ and $z(x) \equiv z(v_i^c(x)\pm, x)$. Adding and subtracting $1/(w_1^2 z_1)$ to the integrand of (4.34) and manipulating the result yields

$$v_i^c(X_i^c) - v_i^c(-4H) = \frac{1}{w_1^2 z_1}(X_i^c + 4H)$$

$$+\int_{-4H}^{X_i^c}\frac{w_1+w(x)}{w^2(x)w_1^2z_1}(w_1-w(x))dx+\int_{-4H}^{X_i^c}\frac{1}{w^2(x)z(x)z_1}(z_1-z(x))dx. \tag{4.35}$$

We can use the reference invariant region to partially bound the integrals in (4.35):

$$\int_{-4H}^{X_i^c}\frac{w_1+w(x)}{w^2(x)w_1^2z_1}(w_1-w(x))dx+\int_{-4H}^{X_i^c}\frac{1}{w^2(x)z(x)z_1}(z_1-z(x))dx$$

$$\leq\frac{2W_1}{W_0^4Z_0}\int_{-4H}^{X_i^c}(w_1-w(x))dx+\frac{1}{W_0^2Z_0^2}\int_{-4H}^{X_i^c}(z_1-z(x))dx; \tag{4.36}$$

however, we still require bounds on the integrals which appear on the right-hand side of (4.36).

Since $w(\tau_M(x)+,x)$ is left-continuous (from Theorem 3.2), we know from the definition of X_i^c that

$$w(v_i^c(X_i^c)+,X_i^c)\geq w_1-\varepsilon_M(w_1-w_0). \tag{4.37}$$

Therefore, Lemma 4.5 implies that

$$w(v_i^c(x)+,x)\geq w_1-\varepsilon_2(w_1-w_0),\quad x\in[-4H,X_i^c], \tag{4.38}$$

where $\varepsilon_2\equiv\varepsilon_M\exp[7H\alpha_2]$. This allows us to bound the first integral on the right-hand side of (4.36):

$$\int_{-4H}^{X_i^c}(w_1-w(x))dx\leq\int_{-4H}^{X_i^c}\varepsilon_2(w_1-w_0)dx\leq 7H\varepsilon_2(w_1-w_0). \tag{4.39}$$

The value of $z(x)$ cannot be too small in the region $x\in[-4H,X_i^c]$ or (4.37) will fail to be satisfied. By applying the logic used in Lemma 4.2, we can explicitly express this bound on the behavior of $z(x)$ as

$$\frac{\varepsilon_2}{E\alpha_1}\geq\int_{-4H}^{X_i^c}\gamma(x)dx, \tag{4.40}$$

where $\gamma(x)$ is defined by the equation $z(x)\equiv z_1-\gamma(x)(z_1-z_0)$. This allows us to bound the second integral on the right-hand side of (4.36):

$$\int_{-4H}^{X_i^c}(z_1-z(x))dx=(z_1-z_0)\int_{-4H}^{X_i^c}\gamma(x)dx\leq\frac{\varepsilon_2}{E\alpha_1}(z_1-z_0). \tag{4.41}$$

We combine (4.39) and (4.41) with (4.35) and (4.36) and use the fact that $(z_1-z_0)\leq E_2(w_1-w_0)$ to obtain the final form of the bound on $v_i^c(x)$:

$$v_i^c(X_i^c)-v_i^c(-4H)\leq\frac{1}{w_1^2z_1}(X_i^c+4H)+K\varepsilon_2(w_1-w_0). \tag{4.42}$$

In the above equation—and throughout the rest of this paper—K will be used to denote any positive constant that depends strictly on the period, the bounds on the characteristic speeds, and the reference invariant region.

Using analogous steps, we can establish a similar bound on $v_i^b(x)$ in the region $x \in [-4H, X_i^b]$:

$$v_i^b(X_i^b) - v_i^b(-4H) \geq \frac{1}{w_0^2 z_0}(X_i^b + 4H) - K\varepsilon_1(w_1 - w_0). \tag{4.43}$$

Now we turn our attention to the region past the τ_M characteristic. Because (4.42) and (4.43) pull v_i^c and v_i^b so far away from each other, we can let the two characteristics approach each other as quickly as possible in this region. From (2.5) and (3.12), we have

$$v_i^c(x) - v_i^c(X_i^c) = \int_{X_i^c}^{x} \frac{dy}{w_0^2(v_i^c(y), y) z_0(v_i^c(y), y)} \leq \frac{1}{w_0^2 z_0}(x - X_i^c), \quad x \in [X_i^c, 3H], \tag{4.44}$$

$$v_i^b(x) - v_i^b(X_i^b) = \int_{X_i^b}^{x} \frac{dy}{w_0^2(v_i^b(y), y) z_0(v_i^b(y), y)} \geq \frac{1}{w_1^2 z_1}(x - X_i^b), \quad x \in [X_i^b, 3H]. \tag{4.45}$$

We now have bounds for v_i^c and v_i^b throughout our region of interest. Since backward 2-characteristics cannot cross, we know that $v_i^c(x)$ and $v_i^b(x)$ cannot cross in the region $x \in [-4H, X_i^c]$. We now show that $v_i^c(x)$ and $v_i^b(x)$ cannot cross in the region $x \in [X_i^c, 3H]$. First, we combine (4.42) with (4.44) and use the fact that $X_i^c \geq 0$ to obtain

$$\begin{aligned} v_i^c(x) - v_i^c(-4H) &\leq \frac{X_i^c + 4H}{w_1^2 z_1} + \frac{x - X_i^c}{w_0^2 z_0} + K\varepsilon_2(w_1 - w_0) \\ &\leq \frac{4H}{w_1^2 z_1} + \frac{x}{w_0^2 z_0} + K\varepsilon_2(w_1 - w_0). \end{aligned} \tag{4.46}$$

Similarly, we combine (4.43) with (4.45) and use the fact that $X_i^b \geq 0$ to obtain

$$\begin{aligned} v_i^b(x) - v_i^b(-4H) &\geq \frac{X_i^b + 4H}{w_0^2 z_0} + \frac{x - X_i^b}{w_1^2 z_1} - K\varepsilon_1(w_1 - w_0) \\ &\geq \frac{4H}{w_0^2 z_0} + \frac{x}{w_1^2 z_1} - K\varepsilon_1(w_1 - w_0). \end{aligned} \tag{4.47}$$

Subtracting (4.46) from (4.47) and using the fact that $v_i^b(-4H) - v_i^c(-4H) > 0$ yields

$$v_i^b(x) - v_i^c(x) \geq \left[\frac{1}{w_0^2 z_0} - \frac{1}{w_1^2 z_1}\right](4H - x) - K(w_1 - w_0)(\varepsilon_1 + \varepsilon_2). \tag{4.48}$$

Using the reference invariant region, the fact that $x \leq 3H$, and the bounds on the slope of the equilibrium curve, we can rearrange and minimize the right-hand side of (4.48) so as to obtain

$$\begin{aligned} v_i^b(x) - v_i^c(x) &\geq \left[\frac{2W_0 H}{W_1^4 Z_1} + \frac{E_1 H}{W_1^2 Z_1^2} - K(\varepsilon_1 + \varepsilon_2)\right](w_1 - w_0) \\ &= [K - K(\varepsilon_1 + \varepsilon_2)](w_1 - w_0). \end{aligned} \tag{4.49}$$

For sufficiently small ε_m and ε_M, ε_1 and ε_2 become small enough to cause the $K(\varepsilon_1 + \varepsilon_2)$ term in (4.49) to be small compared to the K term. Therefore, for sufficiently small ε_m and ε_M, we have that

$$v_i^b(x) - v_i^c(x) \geq K(w_1 - w_0), \quad x \in [X_i^c, 3H]. \tag{4.50}$$

We conclude from (4.50) that the v_i^b and v_i^c characteristics never cross. This conclusion, of course, assumes that $w_1 > w_0$, but this is not a problem since $w_1 = w_0$ is just the trivial case where the invariant region is already a single point on the chemical-equilibrium curve, leaving nothing more to prove.

From Lemma 4.6, we get the following useful fact.

COROLLARY 4.1. *For sufficiently small ε_m and ε_M, the limit of w from the right of v_i^c is bounded away from w_0; specifically,*

$$w(v_i^c(x)+, x) \geq w_0 + \varepsilon_1 e^{-3H\alpha_2}(w_1 - w_0), \tag{4.51}$$

where $x \in [X_i^c, x_i^c]$.

Proof. Consider any $Y \in [X_i^c, x_i^c]$. Define v as the maximal backward 2-characteristic emanating from $(v_i^c(Y), Y)$, and define y as the x-coordinate of the point where v and τ_M cross. From Lemma 4.6, we have that $X_i^b < y \leq X_i^c$. Therefore, the nature of the construction implies that

$$w(v(y)+, y) \geq w_0 + \varepsilon_1(w_1 - w_0). \tag{4.52}$$

We can relate $w(v(y)+, y)$ to $w(v(Y)+, Y)$ by using (4.26a) of Lemma 4.5:

$$w(v(Y)+, Y) \geq w_0 + (w(v(y)+, y) - w_0)e^{-\alpha_2(Y-y)}. \tag{4.53}$$

Combining (4.52) and (4.53) with the fact that $Y - y \leq 3H$ establishes the claim of the corollary.

Now we analyze how wide the distance between the v_i^t and v_i^c characteristics can become as x increases. Specifically, we wish to relate $x_i^t - x_i^c$ to $X_i^t - X_i^c$.

LEMMA 4.7. *There is a constant $C > 0$ that depends strictly on the bounds on the characteristic speeds s_i such that*

$$x_i^t - x_i^c \leq C(X_i^t - X_i^c). \tag{4.54}$$

Proof. To obtain the lemma's claim, we must consider the behavior of v_i^t and v_i^c in three separate regions: $x \in [X_i^c, X_i^t]$, $x \in [X_i^t, x_i^c]$, and $x \in [x_i^c, x_i^t]$.

Since we have bounds on the characteristic speeds, it is straightforward to use (4.7) to express explicit bounds on the widening of the characteristics in the regions $x \in [X_i^c, X_i^t]$ and $x \in [x_i^c, x_i^t]$:

$$(X_i^t - X_i^c)(s_4 - s_1) \geq v_i^c(X_i^t) - v_i^t(X_i^t), \tag{4.55}$$

$$(x_i^t - x_i^c)(s_3 - s_2) \leq v_i^c(x_i^c) - v_i^t(x_i^c). \tag{4.56}$$

Now we consider the range $x \in [X_i^t, x_i^c]$. We can apply the method used to obtain (4.43) in Lemma 4.6 to establish an analogous result for the backward characteristic v_i^t in this region:

$$v_i^t(x_i^c) - v_i^t(X_i^t) \geq \frac{1}{w_0^2 z_0}(x_i^c - X_i^t) - K\varepsilon_1^2(w_1 - w_0). \tag{4.57}$$

The forward characteristic $v_i^c(x)$ can propagate with either classical or shock speed. Either way, we have from (2.13b) and (3.6) that

$$v_i^c(x_i^c) - v_i^c(X_i^t) = \int_{X_i^t}^{x_i^c} \frac{dx}{z(v_i^c(x)\pm, x)w(v_i^c(x)-, x)w(v_i^c(x)+, x)}. \tag{4.58}$$

Now we use Corollary 4.1 to obtain the bound

$$v_i^c(x_i^c) - v_i^c(X_i^t) \le \int_{X_i^t}^{x_i^c} \frac{dx}{z_0 w_0 (w_0 + \varepsilon_1 (w_1 - w_0)\exp(-3H\alpha_2))}. \tag{4.59}$$

Using the reference invariant region, the bound in (4.59) can be manipulated into the following form:

$$v_i^c(x_i^c) - v_i^c(X_i^t) \le \frac{1}{w_0^2 z_0}(x_i^c - X_i^t) - \frac{\varepsilon_1 (w_1 - w_0)\exp(-3H\alpha_2)}{W_1^3 Z_1}(x_i^c - X_i^t). \tag{4.60}$$

We claim that $x_i^c - X_i^t$ cannot be too small. This comes from the requirement that τ_M and τ_m be at least one period apart from each other. Since $v_i^c(x_i^c) - v_i^c(X_i^c) \ge L$, we know that $x_i^c - X_i^c \ge L/s_4$. From Lemma 4.3 and (4.32), we see that $X_i^c - X_i^t \to 0$ as $\varepsilon_M \to 0$; therefore, $x_i^c - X_i^t \ge L/(2s_4)$ and $x_i^c - X_i^t$ cannot be too small for sufficiently small ε_M. This allows us to express (4.60) as

$$v_i^c(x_i^c) - v_i^c(X_i^t) \le \frac{1}{w_0^2 z_0}(x_i^c - X_i^t) - K\varepsilon_1 (w_1 - w_0). \tag{4.61}$$

We combine (4.57) with (4.61) to obtain

$$v_i^c(x_i^c) - v_i^t(x_i^c) \le v_i^c(X_i^t) - v_i^t(X_i^t) - (w_1 - w_0)(K\varepsilon_1 - K\varepsilon_1^2). \tag{4.62}$$

Therefore, for ε_1 sufficiently small, we have

$$v_i^c(x_i^c) - v_i^t(x_i^c) \le v_i^c(X_i^t) - v_i^t(X_i^t). \tag{4.63}$$

(4.63) can now be combined with the bounds in the other regions, (4.55) and (4.56), to yield

$$x_i^t - x_i^c \le \frac{s_4 - s_1}{s_3 - s_2}(X_i^t - X_i^c), \tag{4.64}$$

which, by defining $C \equiv (s_4 - s_1)/(s_3 - s_2)$, establishes the claim of the lemma.

We are now ready to show proportional shrinkage of the invariant region. We begin by showing that the range of z values in the solution must shrink in proportion to $z_1 - z_0$ as the solution progresses through a spatial interval of length $7H$.

THEOREM 4.2. *There is a constant $\varepsilon' > 0$ that depends* only *on the period L, the bounds on the characteristic speeds s_1, s_2, s_3, and s_4, and the reference invariant region Z_0, Z_1, W_0, and W_1 such that*

$$\frac{z_M - z_m}{z_1 - z_0} \le 1 - \varepsilon'. \tag{4.65}$$

Proof. From Lemmas 4.6 and 4.7, we know that there exists some sufficiently small $\bar{\varepsilon} > 0$ that depends only on the period, the bounds on the characteristic speeds, and the reference invariant region such that for all $\varepsilon_m \le \bar{\varepsilon}$ and $\varepsilon_M \le \bar{\varepsilon}$, we have that

$x_i^t - x_i^c \leq C(X_i^t - X_i^c)$. We sum this relation over all the i-blocks and use (4.32) to obtain

$$m(A_{\varepsilon_m}) \leq \sum_i (x_i^t - x_i^c) \leq C \sum_i (X_i^t - X_i^c) \leq Cm(A_{\varepsilon_M}). \tag{4.66}$$

Therefore, the measure of the set of x values on τ_M where $w(\tau_M(x)+, x)$ is bounded away from w_1 bounds the measure of the set of x values on τ_m where $w(\tau_M(x)+, x)$ is close to w_0.

Now we set $\varepsilon_m \equiv \bar{\varepsilon}$. From Lemma 4.3 and (4.66), it is clear that $\varepsilon_M \geq Km(A_{\bar{\varepsilon}})$. Therefore, there is a constant $\delta \in (0, \bar{\varepsilon}]$ that depends only on the period, the bounds on the characteristic speeds, and the reference invariant region such that $(\varepsilon_M)^2 \leq \delta$ implies $m(A_{\bar{\varepsilon}}) \leq H/2$. This allows us to use Lemma 4.4 to conclude that $(\varepsilon_{\bar{m}})^2 \geq K\bar{\varepsilon}$ whenever $(\varepsilon_M)^2 \leq \delta$. Therefore, $(\varepsilon_M)^2 + (\varepsilon_{\bar{m}})^2 \geq \varepsilon'$, where $\varepsilon' \equiv \min[\delta, K\bar{\varepsilon}]$. However, by the definitions of ε_M and $\varepsilon_{\bar{m}}$, we know that

$$\frac{z_M - z_m}{z_1 - z_0} = 1 - (\varepsilon_M)^2 - (\varepsilon_{\bar{m}})^2, \tag{4.67}$$

which immediately implies (4.65).

Now we look at the shrinkage in the range of values of w as we proceed through an additional interval of length $\ln 2/\alpha_2$.

THEOREM 4.3. *There is a constant $\varepsilon > 0$ that depends only on the period, the bounds on the characteristic speeds, and the reference invariant region and there are also constants w_M and w_m such that*

$$\frac{w_M - w_m}{w_1 - w_0} \leq 1 - \varepsilon, \tag{4.68}$$

$$w_m \leq w(t, 3H + (\ln 2)/\alpha_2) \leq w_M \quad \forall t \in (-\infty, \infty), \tag{4.69}$$

$$\text{and } z_e(w_m) \leq z(t, 3H + \ln 2/\alpha_2) \leq z_e(w_M) \quad \forall t \in (-\infty, \infty). \tag{4.70}$$

Proof. From Theorem 4.2, we know that there is an $r \in [0, 1]$ such that $z_1 - z_M \geq r\varepsilon'(z_1 - z_0)$ and $z_m - z_0 \geq (1 - r)\varepsilon'(z_1 - z_0)$. From Theorem 4.5, we know that if $x \in [3H, 3H + \ln 2/\alpha_2]$, then $z(t, x) \leq z_M + (r\varepsilon'/2)(z_1 - z_0)$ and therefore $z(t, x) \leq z_1 - (r\varepsilon'/2)(z_1 - z_0)$. The method of the proof of Theorem 4.5 can also be used to show that since z is bounded away from z_1 in the region $x \in [3H, 3H + \ln 2/\alpha_2]$, w must be drawn away from w_1 as the solution progresses through this region. Specifically,

$$\sup_t w(t, 3H + (\ln 2)/\alpha_2) \leq w_M, \tag{4.71}$$

where

$$w_M \equiv w_e\left(z_1 - \frac{r\varepsilon'}{2}(z_1 - z_0)\right) + \left[w_1 - w_e\left(z_1 - \frac{r\varepsilon'}{2}(z_1 - z_0)\right)\right] e^{-\frac{\alpha_1}{\alpha_2}\ln 2}. \tag{4.72}$$

The definition in (4.72) can be transformed using the bounds on the derivative of $w_e(z)$ to yield

$$\frac{w_1 - w_M}{w_1 - w_0} \geq \frac{r\varepsilon' E}{2}\left(1 - 2^{-\frac{\alpha_1}{\alpha_2}}\right). \tag{4.73}$$

Further, if we take z_e of both sides of (4.72), we see that

$$\sup_t z(t, 3H + \ln 2/\alpha_2) \leq z_e(w_M). \tag{4.74}$$

Similarly, we also have that $z(t,x) \geq z_m - ((1-r)\varepsilon'/2)(z_1 - z_0) \geq z_0 + ((1-r)\varepsilon'/2)(z_1 - z_0)$ for $x \in [3H, 3H + \ln 2/\alpha_2]$, and this leads to

$$\inf_t w(t, 3H + (\ln 2)/\alpha_2) \geq w_m, \tag{4.75}$$

where

$$w_m \equiv w_e\left(z_0 + \frac{(1-r)\varepsilon'}{2}(z_1 - z_0)\right) + \left[w_1 - w_e\left(z_0 + \frac{(1-r)\varepsilon'}{2}(z_1 - z_0)\right)\right] e^{-\frac{\alpha_1}{\alpha_2}\ln 2}. \tag{4.76}$$

We also have that

$$\frac{w_m - w_0}{w_1 - w_0} \geq \frac{(1-r)\varepsilon' E}{2}\left(1 - 2^{-\frac{\alpha_1}{\alpha_2}}\right) \tag{4.77}$$

and

$$\inf_t z(t, 3H + \ln 2/\alpha_2) \geq z_e(w_m). \tag{4.78}$$

Adding (4.73) to (4.77) and defining $\varepsilon \equiv (\varepsilon' E/2)(1 - 2^{-\alpha_1/\alpha_2})$ establishes the claim of the lemma.

Theorem 4.3 allows us to ascertain the asymptotic behavior of the periodic case.

COROLLARY 4.2. *The invariant set containing the solution shrinks exponentially to a single point on the equilibrium curve as $x \to \infty$.*

Proof. First, we redefine our coordinate system so that $x = 0$ corresponds to a location where the solution lies within the reference invariant region. By repeatedly applying Theorem 4.3, we have that for any nonnegative integer n,

$$\sup_{(t,x)\in X_n} w(t,x) - \inf_{(t,x)\in X_n} w(t,x) \leq (W_1 - W_0)\exp[n \ln(1-\varepsilon)] \tag{4.79}$$

and

$$w_e\left(\sup_{(t,x)\in X_n} z(t,x)\right) - w_e\left(\inf_{(t,x)\in X_n} z(t,x)\right) \leq (W_1 - W_0)\exp[n\ln(1-\varepsilon)], \tag{4.80}$$

where $X_n \equiv \{(t,x) : n(7H + \ln 2/\alpha_2) \leq x < (n+1)(7H + \ln 2/\alpha_2)\}$. Therefore, the solution lies in invariant rectangles that shrink exponentially in size to a single point as $x \to \infty$. Further, since the lower left-hand corner and the upper right-hand corner of the rectangle are on the equilibrium curve, we conclude that the point to which the invariant regions shrink must also be on the equilibrium curve.

Finally, we determine which point on the equilibrium curve is approached as $x \to \infty$ by using the form of the mass balance for the two chemical components given in (2.3b).

THEOREM 4.4. *Under periodic conditions, the solution exponentially shrinks to the point (Z, W) on the equilibrium curve defined by*

$$Z \equiv \frac{M}{2} + \sqrt{\frac{M^2}{4} - \frac{bM - c}{a}} \quad \textit{and} \quad W \equiv w_e(Z), \tag{4.81}$$

where a, b, and c are defined in (2.3), $M \equiv (1/L)\int_{t_0}^{t_0+L} z(t,0) + w(t,0)dt$, *and* t_0 *is any arbitrary number.*

Proof. Combining the mass balance in (2.3b) with the fact that $v = z + w$ and $u = zw$ yields

$$\frac{\partial}{\partial x}\left(z(t,x) + w(t,x)\right) - \frac{\partial}{\partial t}\left(\frac{1}{z(t,x)w(t,x)}\right) = 0. \tag{4.82}$$

We integrate (4.82) over the region $\{(t,x) : t_0 \le t \le t_0 + L,\ 0 \le x \le x_1\}$. Since the solution is periodic, the integral of the term in (4.82) with the partial derivative with respect to t equals zero. If we let $x_1 \to \infty$ and apply Corollary 4.2, we obtain from the integrated form of (4.82) that

$$\int_{t_0}^{t_0+L} [z(t,0) + w(t,0)]dt = L(Z + W), \tag{4.83}$$

where $W = w_e(Z)$. After substituting the definition of $w_e(z)$ given in (2.9) into (4.83), we explicitly solve for Z, which yields the result in (4.81).

Acknowledgments. This paper was originally part of the author's Ph.D. dissertation at Brown University, which was supervised by Professor Constantine Dafermos, whom I wish to thank for supplying invaluable advice throughout this paper's development. I would also like to thank the anonymous referee of this paper for providing so many helpful comments and suggestions.

REFERENCES

[1] C. M. Dafermos, *Generalized characteristics in hyperbolic systems of conservation laws*, Arch. Rational Mech. Anal., 107 (1989), pp. 127–155.

[2] C. M. Dafermos and X. Geng, *Generalized characteristics, uniqueness, and regularity of solutions in a hyperbolic system of conservation laws*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 8 (1991), pp. 231–269.

[3] A. F. Filippov, *Differential equations with discontinuous right-hand side*, Mat. Sb., 51 (1960), pp. 99–128; English translation: Amer. Math. Soc. Transl. Ser. 2, 42 (1960), pp. 199–231.

[4] J. Hale, *Ordinary differential equations*, Wiley–Interscience, New York, 1969, p. 36.

[5] P. D. Lax, *Shock waves and entropy*, Contributions to Functional Analysis, E. A. Zaranteonello, ed., Academic Press, New York, 1971, pp. 603–634.

[6] D. N. Ostrov, *Hyperbolic conservation laws arising in chromatography*, Ph.D. thesis, Brown University, Providence, RI, 1994.

[7] H.-K. Rhee, R. Aris, and N. R. Amundson, *First-Order Partial Differential Equations*, vol. II, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[8] D. Serre, *Solutions à variations bornées pour certains systèmes hyperboliques de Lois de conservation*, J. Differential Equations, 68 (1987), pp. 137–168.

[9] B. Temple, *Systems of conservation laws with invariant submanifolds*, Trans. Amer. Math. Soc., 280 (1983), pp. 781–795.

SIAM J. MATH. ANAL.
Vol. 27, No. 6, pp. 1597–1630, November 1996

005

ON THE SOLUTION OF TIME-HARMONIC SCATTERING PROBLEMS FOR MAXWELL'S EQUATIONS*

CHRISTOPHE HAZARD† AND MARC LENOIR†

Abstract. This paper deals with the scattering of a monochromatic electromagnetic wave by a perfect conductor surrounded by a locally inhomogeneous medium. The direct numerical solution of this problem by a finite-element method requires special edge elements. The aim of the present paper is to give an equivalent formulation of the problem well suited for both easy theoretical investigation and numerical implementation. Following a well-known idea, this formulation is obtained by adding a regularizing term such as "grad div" in the time-harmonic Maxwell equations, which leads us to solve an elliptic problem similar to the vector Helmholtz equation instead of Maxwell's equation. The numerical treatment of this new formulation requires only standard Lagrange finite elements.

A unified approach, which is valid for the equations satisfied by either the electric or the magnetic field, is presented. It applies for a conductor with a Lipschitz-continuous boundary surrounded by a dissipative or nondissipative medium whose electromagnetic coefficients (permittivity and permeability) may be irregular. A family of scattering problems is defined, that is, the classical problem (which follows from Maxwell's equations) and the so-called "regularized problem" obtained by adding a regularizing term in Maxwell's equations. These problems are shown to be well posed and to have the same solution. An integral representation technique is described.

Key words. Maxwell's equations, scattering by obstacles, integral representation

AMS subject classifications. 35C15, 35J55, 78A45

1. Introduction.

1.1. Motivation.

Consider a perfect conductor surrounded by a medium whose electromagnetic coefficients (permittivity and permeability) are assumed constant outside a bounded domain. We are concerned with the scattering of a time-harmonic electromagnetic wave by this inhomogeneity of the space. (Let us mention that the presence of the perfect conductor is not essential; the same holds for the scattering by a dielectric obstacle.) The direct numerical solution of such a problem has been extensively studied. In particular, the use of Nédélec's curl-conforming finite elements [20] seems to be widely held (see, e.g., Levillain [16]), although there remain some open questions regarding the proof of the convergence properties of the numerical schemes involving high-order Nédélec finite elements (see Kikuchi [12]). Our aim is to propose an alternative approach that allows the use of standard Lagrange finite elements. It consists of replacing the classical Maxwell equations by an elliptic problem that has the same solution.

The main idea to construct this "regularized problem" is not new and may be found, for instance, in the papers from Werner [25], Leis [15], Knauff and Kress [13], or, more recently, Bamberger and Bonnet [3] and Mayergoyz and D'Angelo [17]. It is based on the simple fact that a three-dimensional field U that is a solution to

$$\operatorname{curl}(\operatorname{curl} U) - k^2 U = 0, \quad k \neq 0, \tag{1.1}$$

in an open set of $\mathbb{R}^3$ satisfies $\operatorname{div} U = 0$. As a consequence, it is also a solution to the vector Helmholtz equation

$$-\Delta U - k^2 U = 0 \tag{1.2}$$

* Received by the editors July 18, 1994; accepted for publication July 26, 1995.

† Laboratoire de Simulation et Modélisation des Phénomènes de Propagation (Associé au Centre National de la Recherche Scientifique et à l'Université Pierre et Marie Curie), École Nationale Supérieure de Techniques Avancées, Chemin de la Hunière, 91120 Palaiseau, France.

by virtue of the relation $\operatorname{curl}(\operatorname{curl} U) - \operatorname{grad}(\operatorname{div} U) = -\Delta U$. The converse is of course wrong if no additional condition is given, but it becomes true in the context of the scattering of electromagnetic waves by an obstacle with additional boundary conditions on the obstacle and at infinity. This may be easily understood as follows. Let U be a solution to (1.2) and assume that U satisfies the well-known Sommerfeld radiation condition at infinity. It is readily seen that its divergence $\varphi = \operatorname{div} U$ satisfies the scalar Helmholtz equation $-\Delta\varphi - k^2\varphi = 0$ as well as the scalar Sommerfeld radiation condition. Consequently, if we add a condition such as $\varphi = 0$ on the boundary of the obstacle (which has to be assumed regular), we deduce from classical results (Rellich [21]) that φ vanishes everywhere outside the obstacle, which shows that U is a solution to (1.1).

Our purpose is to extend this result to the general case of variable (discontinuous) electromagnetic coefficients and irregular boundaries for which this simple proof does not apply anymore. In this situation, (1.1) is replaced by the time-harmonic Maxwell equation

$$\operatorname{curl}(\zeta^{-1} \operatorname{curl} U) - \omega^2 \xi\, U = 0. \tag{1.3}$$

Noticing that every solution to this equation satisfies $\operatorname{div} \xi U = 0$, we will see that instead of solving (1.3) we can equivalently solve its regularized form

$$\operatorname{curl}(\zeta^{-1} \operatorname{curl} U) - \bar{\xi} \operatorname{grad}(\tau^{-1} \operatorname{div} \xi U) - \omega^2 \xi\, U = 0, \tag{1.4}$$

which has the advantage of involving an elliptic second-order differential operator. To prove the equivalence between the classical and regularized scattering problems (related, respectively, to (1.3) and (1.4)), we cannot study the equation satisfied by $\operatorname{div} \xi U$ as described above. The method we propose consists of proving the well-posedness of both problems and noticing that the solution to the classical problem solves the regularized equations since it is divergence-free.

From a numerical point of view, the regularized problem has a twofold interest compared with the classical problem. On one hand, its associated fundamental solution (Green tensor) has a weaker singularity (order 1 instead of 3), which is essential as far as integral representations are concerned. On the other hand, we will see that the regularized problem comes within the classical framework of approximation of Fredholm operators, which allows us in particular to use a standard discretization by Lagrange finite elements.

1.2. Outline of the paper. In §2.1, we present the classical equations that model the scattering of a time-harmonic wave by an inhomogeneous medium. We deal simultaneously with the equations satisfied by the electric or magnetic fields in the cases of dissipative or nondissipative media. Then, in §2.2, we introduce the regularizing term $\bar{\xi} \operatorname{grad}(\tau^{-1} \operatorname{div} \xi U)$ mentioned above. By suitable modifications of the boundary conditions as well as the radiation condition at infinity, we thus define a "regularized problem." Section 2.3 is devoted to the description of the functional framework associated with this problem. We finally state in §2.4 the main result of this paper, which consists of the equivalence between the classical and regularized formulations; the introduction of the regularizing term (which depends on the choice of function τ) does not affect the solution to the problem. The proof of this statement is the object of the remainder of the paper, where we develop a method that is well adapted for the numerical solution to the regularized problem by means of standard finite elements.

Section 3 consists of the proof of the uniqueness of the solution to the scattering problems (classical and regularized). The method we use is rather classical. The treatment of infinity (which is assumed homogeneous) is based on the well-known asymptotic behavior of the solutions to the scalar Helmholtz equation (Rellich [21]). Then, the treatment of the remainder of the space (containing all the inhomogeneities) follows by a unique continuation technique; here appear some restrictions concerning the regularity of the electromagnetic coefficients. Indeed, we will see that these coefficients have to be assumed piecewise Lipschitz continuous in the part of the domain where the medium is not dissipative.

Section 4 is devoted to the proof of the existence of a solution for the classical or regularized problem. The technique we present is similar to integral equation techniques [5]; it actually is an adaptation of the so-called "method of coupling between variational formulation and integral representation" introduced by Jami and Lenoir [9] in linear hydrodynamics. This method consists of replacing the initial problem by an equivalent problem (set in a bounded domain) that is obtained by means of an integral representation formula. We show that the existence and the uniqueness of the solution to this latter problem are a matter for Fredholm alternative; by virtue of the uniqueness proved in §3, the existence property for the initial problem follows. Let us point out that for the regularized problem, the same results may be obtained by a standard integral equation approach. Our method simply has the advantage of leading to an integral operator that involves a nonsingular kernel.

Finally, we have postponed until Appendices A and B some technical but essential results that concern, respectively, the statement of integral representation formulas and compactness properties.

1.3. Notation. In this section, we recall some usual notation concerning the function spaces that are used in the context of Maxwell's equations; we refer, for instance, to Girault and Raviart [8] for a detailed study of these spaces.

Let $\mathcal{O}$ be a bounded open set in $\mathbb{R}^3$. We define $\mathcal{D}(\mathcal{O})$ to be the space of infinitely differentiable functions with compact support in $\mathcal{O}$, $\mathcal{D}(\overline{\mathcal{O}}) = \{\varphi_{|\mathcal{O}} \mid \varphi \in \mathcal{D}(\mathbb{R}^3)\}$, and $\mathcal{D}'(\mathcal{O})$ is the dual space of $\mathcal{D}(\mathcal{O})$ (space of distributions). We use the notation $H^s(\mathcal{O})$, $s \in \mathbb{R}$, for the classical Sobolev spaces and $H_0^s(\mathcal{O})$ for the completion of $\mathcal{D}(\mathcal{O})$ in $H^s(\mathcal{O})$. We denote by $H(\mathrm{curl};\mathcal{O})$ and $H(\mathrm{div};\mathcal{O})$ the Hilbert spaces

$$H(\mathrm{curl};\mathcal{O}) = \{V \in L^2(\mathcal{O})^3 \mid \mathrm{curl}\, V \in L^2(\mathcal{O})^3\},$$
$$H(\mathrm{div};\mathcal{O}) = \{V \in L^2(\mathcal{O})^3 \mid \mathrm{div}\, V \in L^2(\mathcal{O})\}.$$

If $\mathcal{O}$ has a Lipschitz-continuous boundary $\partial\mathcal{O}$, then for every function $V \in H(\mathrm{curl};\mathcal{O})$ (respectively, $V \in H(\mathrm{div};\mathcal{O})$), the quantity $(V \wedge n)_{|\partial\mathcal{O}}$ (respectively, $(V \cdot n)_{|\partial\mathcal{O}}$) is defined by Green's formula as an element of $H^{-1/2}(\partial\mathcal{O})^3$ (respectively, $H^{-1/2}(\partial\mathcal{O})$). Let $H_0(\mathrm{curl};\mathcal{O})$ and $H_0(\mathrm{div};\mathcal{O})$ be the completions of $\mathcal{D}(\mathcal{O})^3$ in $H(\mathrm{curl};\mathcal{O})$ and $H(\mathrm{div};\mathcal{O})$. We have

$$H_0(\mathrm{curl};\mathcal{O}) = \{V \in H(\mathrm{curl};\mathcal{O}) \mid V \wedge n = 0 \text{ on } \partial\mathcal{O}\},$$
$$H_0(\mathrm{div};\mathcal{O}) = \{V \in H(\mathrm{div};\mathcal{O}) \mid V \cdot n = 0 \text{ on } \partial\mathcal{O}\}.$$

For an unbounded domain $\Omega \subset \mathbb{R}^3$, $H_{\mathrm{loc}}(\mathrm{curl};\Omega)$ (respectively, $H_{\mathrm{loc}}(\mathrm{div};\Omega)$) denotes the Frechet space of functions $V \in L^2_{\mathrm{loc}}(\Omega)$ such that $V_{|\mathcal{O}} \in H(\mathrm{curl};\mathcal{O})$ (respectively, $V_{|\mathcal{O}} \in H(\mathrm{div};\mathcal{O})$) for every bounded set $\mathcal{O} \subset \Omega$.

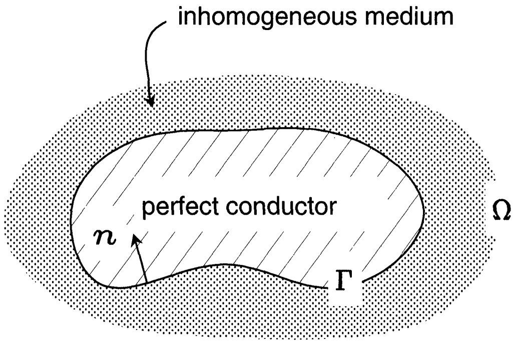


FIG. 2.1. *Scattering by a perfect conductor.*

2. The classical and regularized scattering problems.

2.1. Classical formulation. Let $\Omega \subset \mathbb{R}^3$ be an unbounded connected domain (the complement of a compact set of $\mathbb{R}^3$) with a Lipschitz-continuous boundary Γ. We denote by n the unit outward normal on Γ. The domain Ω is filled by an isotropic medium that may be inhomogeneous in a bounded part of Ω (see Fig. 2.1). Let us mention that the assumption of isotropy is not essential; the same study may be easily extended to anisotropic media.

2.1.1. Maxwell's equations. For a given frequency $\omega > 0$, the electric and magnetic fields E and H satisfy in Ω the time-harmonic Maxwell equations

$$-i\omega \begin{pmatrix} E \\ H \end{pmatrix} = \begin{pmatrix} 0 & \varepsilon^{-1}\operatorname{curl} \\ -\mu^{-1}\operatorname{curl} & 0 \end{pmatrix} \begin{pmatrix} E \\ H \end{pmatrix},$$

where the functions ε and μ are, respectively, the electric permittivity and the magnetic permeability of the medium. By eliminating E or H, Maxwell's equations turn into the second-order equation

$$\operatorname{curl}(\zeta^{-1}\operatorname{curl} U) - \omega^2 \xi\, U = 0 \quad \text{in } \Omega, \tag{2.1}$$

where U denotes either the electric or the magnetic field, and

$$\begin{aligned} \zeta = \mu \quad &\text{and} \quad \xi = \varepsilon \quad \text{if } U = E, \\ \zeta = \varepsilon \quad &\text{and} \quad \xi = \mu \quad \text{if } U = H. \end{aligned} \tag{2.2}$$

In the case of a nondissipative medium, these functions are real and positive. The effect of dissipation is taken into account by adding to ε or μ a positive imaginary part that corresponds to the electric or magnetic conductivity of the medium (see, e.g., [10]). In the present paper, we assume that ζ and ξ are $L^\infty(\Omega)$ complex-valued functions such that

$$\begin{aligned} \mathfrak{Re}\,\zeta(x) \geq \alpha \quad &\text{and} \quad \mathfrak{Im}\,\zeta(x) \geq 0, \\ \mathfrak{Re}\,\xi(x) \geq \alpha \quad &\text{and} \quad \mathfrak{Im}\,\xi(x) \geq 0 \end{aligned} \tag{2.3}$$

almost everywhere in Ω (α is a real positive constant). Moreover, the medium is assumed homogeneous in a vicinity of infinity; we thus suppose that there exists r_0 such that

$$\zeta(x) = \zeta_0 \quad \text{and} \quad \xi(x) = \xi_0 \quad \text{for } \|x\| \geq r_0 > 0. \tag{2.4}$$

Note that outside such a ball, (2.1) simplifies to

$$\operatorname{curl}(\operatorname{curl} U) - k_s^2 U = 0, \quad \text{where } \; k_s = \omega\sqrt{\zeta_0 \xi_0}. \tag{2.5}$$

We use the notation "k_s" because of the similarity with the so-called S-waves in linear elasticity that are characterized by a divergence-free displacement field (since solutions to (2.5) are infinitely differentiable [18], every solution to (2.5) clearly satisfies $\operatorname{div} U = 0$).

Remark 2.1. Assuming the choice of the principal determination for the square root in (2.5), assumptions (2.3) imply that

$$\Re e\, k_s > 0, \qquad \Im m\, k_s \geq 0, \quad \text{and} \tag{2.6}$$

$$k_s \in \mathbb{R}^+ \Leftrightarrow \big(\zeta_0 \in \mathbb{R}^+ \text{ and } \xi_0 \in \mathbb{R}^+\big). \tag{2.7}$$

2.1.2. Boundary conditions. The behavior of U in the vicinity of the perfect conductor depends on whether U represents the electric or the magnetic field. These boundary conditions read classically as

$$U \wedge n = 0 \quad \text{on } \Gamma \;\; \text{(electric boundary condition)} \quad \text{or} \tag{2.8}$$

$$\operatorname{curl} U \wedge n = 0 \quad \text{on } \Gamma \;\; \text{(magnetic boundary condition)}. \tag{2.9}$$

They will be denoted, respectively, by $\mathcal{B}_\infty^E$ and $\mathcal{B}_\infty^H$ in what follows. (The index ∞ refers to the classical Maxwell equation in this paper.)

Remark 2.2. It is well known that the boundary condition $\mathcal{B}_\infty^E$ implies that the quantity $\operatorname{curl} U \cdot n$ vanishes on Γ. On the other hand, the boundary condition $\mathcal{B}_\infty^H$ together with Maxwell's equation (2.1) implies $U \cdot n = 0$ on Γ.

2.1.3. Radiation condition. Since we are concerned with a scattering problem, the field U actually represents the superposition of a given monochromatic incident wave U_I (i.e., a solution to (2.5) in $\mathbb{R}^3$) and an unknown scattered wave U_S, i.e.,

$$U = U_I + U_S. \tag{2.10}$$

The asymptotic behavior of U_S at infinity is specified by means of a radiation condition, which expresses that the energy associated with the scattered wave radiates toward infinity (see §3.2). The expression of this condition depends on k_s (defined in (2.5)):

$$\begin{aligned} &\lim_{R\to\infty} \int_{\Sigma_R} \|\operatorname{curl} U_S \wedge n - ik_s U_S\|^2 \, d\gamma = 0 && \text{if } \Im m\, k_s = 0, \\ &\lim_{R\to\infty} \int_{\Sigma_R} \Big[\|\operatorname{curl} U_S \wedge n\|^2 + |k_s|^2 \, \|U_S\|^2\Big] \, d\gamma = 0 && \text{if } \Im m\, k_s > 0, \end{aligned} \tag{2.11}$$

where $\Sigma_R = \big\{x \in \mathbb{R}^3 \mid \|x\| = R\big\}$. The first condition is the well-known outgoing Silver–Müller radiation condition (see [10] or [18]); the second one is a decay condition for U_S at infinity. They will be referred to as $\mathcal{R}_\infty$ in what follows.

2.1.4. The scattering problems. For a given incident wave U_I (i.e., a given solution to Maxwell's equation (2.5) in the whole space $\mathbb{R}^3$), we consider the two kinds of scattering problems, denoted, respectively, by $\mathcal{P}_\infty^E$ or $\mathcal{P}_\infty^H$ according to the choice of the boundary condition on Γ:

$$(\mathcal{P}_\infty^{E/H}) \qquad \begin{aligned} &\operatorname{curl}(\zeta^{-1} \operatorname{curl} U) - \omega^2 \xi U = 0 \quad \text{in } \Omega, \\ &U \text{ satisfies the boundary condition } \mathcal{B}_\infty^{E/H} \text{ (i.e., (2.8) or (2.9))}, \\ &U_S = U - U_I \text{ satisfies the radiation condition } \mathcal{R}_\infty, \text{ (i.e., (2.11))}. \end{aligned}$$

The aim of the present paper is to show a practical method for solving these problems.

2.2. The regularized formulation. When the medium is homogeneous, there is a close connection between the equations of linear elasticity and Maxwell's equations. If $\mathcal{U}(x,t) = \Re e\left(U(x)e^{-i\omega t}\right)$ is the time-periodic displacement field in a medium with Lamé coefficients λ and μ and density ρ, it satisfies

$$-\sigma_{ij,j}(U) - \omega^2 \rho U = 0, \quad \text{where } \sigma_{ij} = \lambda(\operatorname{div} U)\delta_{ij} + \mu(U_{i,j} + U_{j,i}),$$

which can be written as

$$\mu \operatorname{curl}\,(\operatorname{curl} U) - (\lambda + 2\mu) \operatorname{grad}(\operatorname{div} U) - \omega^2 \rho U = 0.$$

For $\lambda = -2\mu$, this is actually Maxwell's equation (2.5), but even when $\lambda \neq 2\mu$, it is satisfied by any solution to Maxwell's equation since it is divergence-free. For $\mu > 0$ and $\lambda \geq 0$, the above equation is known to be strongly elliptic, which makes it attractive for discretization instead of Maxwell's equations provided that its solution actually is divergence-free.

2.2.1. Regularized equation. The generalization of this remark to nonhomogeneous media leads to

$$\operatorname{curl}(\zeta^{-1} \operatorname{curl} U) - \bar{\xi} \operatorname{grad}(\tau^{-1} \operatorname{div} \xi U) - \omega^2 \xi U = 0, \tag{2.12}$$

where τ is a given function (which has no particular physical meaning as far as electromagnetism is concerned). If τ is assumed to be infinite in the whole domain Ω (i.e., $\tau^{-1} = 0$), this equation reduces to Maxwell's equation (2.1). If τ and τ^{-1} are bounded in Ω, it will be referred to as the "regularized Maxwell equation" in what follows. In the case of a real-valued function ξ, it can be considered as obtained by eliminating E or H in the first-order system

$$-i\omega \begin{pmatrix} E \\ H \\ \varphi \\ \psi \end{pmatrix} = \begin{pmatrix} 0 & \varepsilon^{-1} \operatorname{curl} & \operatorname{grad} \lambda^{-1} & 0 \\ -\mu^{-1} \operatorname{curl} & 0 & 0 & \operatorname{grad} \nu^{-1} \\ \operatorname{div} \varepsilon & 0 & 0 & 0 \\ 0 & \operatorname{div} \mu & 0 & 0 \end{pmatrix} \begin{pmatrix} E \\ H \\ \varphi \\ \psi \end{pmatrix},$$

where $\tau = \lambda$ if $U = E$ and $\tau = \nu$ if $U = H$.

We choose for τ to satisfy conditions similar to (2.3)–(2.4), i.e., $\tau \in L^\infty(\Omega)$ and

$$\Re e\, \tau(x) \geq \alpha > 0 \quad \text{and} \quad \Im m\, \tau(x) \geq 0 \quad \text{a.e. in } \Omega, \quad \text{with} \tag{2.13}$$

$$\tau(x) = \tau_0 \quad \text{for } \|x\| \geq r_0 > 0. \tag{2.14}$$

For sufficiently large $\|x\|$, (2.12) thus simplifies to

$$\operatorname{curl}(\operatorname{curl} U) - t^{-1} \operatorname{grad}(\operatorname{div} U) - k_s^2 U = 0 \quad \text{with } t = \tau_0 \zeta_0^{-1} |\xi_0|^{-2}. \tag{2.15}$$

In what follows, we will assume that either τ satisfies the conditions stated above or $\tau^{-1} = 0$ in Ω (in other words, $\tau \equiv \infty$). In this latter case, (2.12) corresponds to the classical Maxwell equation, and the simplified form (2.15) is merely (2.5) (since $t = \infty$ for $\tau_0 = \infty$).

Just as the notation k_s was related to S-waves in linear elasticity, we denote

$$k_p = k_s \sqrt{t} = \omega \sqrt{\tau_0 / \overline{\xi_0}} \quad \text{if } \tau_0 \neq \infty, \tag{2.16}$$

which corresponds to the wave number of P-waves (i.e., the irrotational displacement fields).

Remark 2.3. As in Remark 2.1, if τ_0 is finite, the assumptions concerning τ and ξ yield

$$\mathfrak{Re}\, k_p > 0, \qquad \mathfrak{Im}\, k_p \geq 0, \quad \text{and} \tag{2.17}$$

$$k_p \in \mathbb{R}^+ \Leftrightarrow \left(\xi_0 \in \mathbb{R}^+ \text{ and } \tau_0 \in \mathbb{R}^+\right). \tag{2.18}$$

2.2.2. Extended boundary conditions. Taking into account the regularizing term in (2.12) leads to an extended expression of the boundary conditions near the perfect conductor. The fact that the regularized equation does not imply the divergence-free condition requires us to complete the classical conditions as follows:

$$U \wedge n = 0, \quad \tau^{-1} \operatorname{div} \xi U = 0 \quad \text{on } \Gamma \quad \text{(electric boundary condition)} \quad \text{or} \tag{2.19}$$

$$\operatorname{curl} U \wedge n = 0, \quad \tau^{-1} U \cdot n = 0 \quad \text{on } \Gamma \quad \text{(magnetic boundary condition)}, \tag{2.20}$$

where the notation $\tau^{-1} \operatorname{div} \xi U = 0$ simply means that we add the condition $\operatorname{div} \xi U = 0$ to the classical electric boundary condition only when the regularized equation is considered (and the same holds for the boundary condition $\tau^{-1} U \cdot n = 0$). These new conditions are denoted in what follows by $\mathcal{B}_\tau^E$ and $\mathcal{B}_\tau^H$, respectively.

2.2.3. Extended radiation conditions. Similar to elasticity, the radiation condition associated with (2.15) (for $t \neq \infty$) involves two relations, respectively, concerning the behavior at infinity of the transverse and radial components of the scattered wave. Their expressions, which depend on k_s and k_p, are

$$\begin{aligned} &\lim_{R\to\infty} \int_{\Sigma_R} \left\| \operatorname{curl} U_S \wedge n - i k_s n \wedge (U_S \wedge n) \right\|^2 d\gamma = 0 && \text{if } \mathfrak{Im}\, k_s = 0, \\ &\lim_{R\to\infty} \int_{\Sigma_R} \left[\left\| \operatorname{curl} U_S \wedge n \right\|^2 + |k_s|^2 \left\| n \wedge (U_S \wedge n) \right\|^2 \right] d\gamma = 0 && \text{if } \mathfrak{Im}\, k_s > 0 \end{aligned} \tag{2.21}$$

and

$$\begin{aligned} &\lim_{R\to\infty} \int_{\Sigma_R} \left| \sqrt{t^{-1}} \operatorname{div} U_S - i k_s\, U_S \cdot n \right|^2 d\gamma = 0 && \text{if } \mathfrak{Im}\, k_p = 0, \\ &\lim_{R\to\infty} \int_{\Sigma_R} \left[\left| t^{-1} \right| \left| \operatorname{div} U_S \right|^2 + |k_s|^2 \left| U_S \cdot n \right| \right] d\gamma = 0 && \text{if } \mathfrak{Im}\, k_p > 0. \end{aligned} \tag{2.22}$$

The conjunction of these two conditions will be denoted by $\mathcal{R}_t$ in what follows.

Remark 2.4. Note that for $t \neq \infty$, the radial radiation condition (2.22) may be written as

$$\begin{aligned} &\lim_{R\to\infty} \int_{\Sigma_R} \left| \operatorname{div} U_S - i k_p\, U_S \cdot n \right|^2 d\gamma = 0 && \text{if } \mathfrak{Im}\, k_p = 0, \\ &\lim_{R\to\infty} \int_{\Sigma_R} \left[\left| \operatorname{div} U_S \right|^2 + |k_p|^2 \left| U_S \cdot n \right|^2 \right] d\gamma = 0 && \text{if } \mathfrak{Im}\, k_p > 0, \end{aligned} \tag{2.23}$$

and for $t = \infty$, it reduces to

$$\lim_{R\to\infty} \int_{\Sigma_R} \left| U_S \cdot n \right|^2 d\gamma = 0. \tag{2.24}$$

By virtue of the relation $\|V\|^2 = \|V \wedge n\|^2 + |V \cdot n|^2$, this latter condition together with the transverse radiation condition (2.21) clearly amounts to (2.11). The formulation (2.21)–(2.22) of $\mathcal{R}_t$ for $t = \infty$ is thus consistent with the classical radiation condition $\mathcal{R}_\infty$.

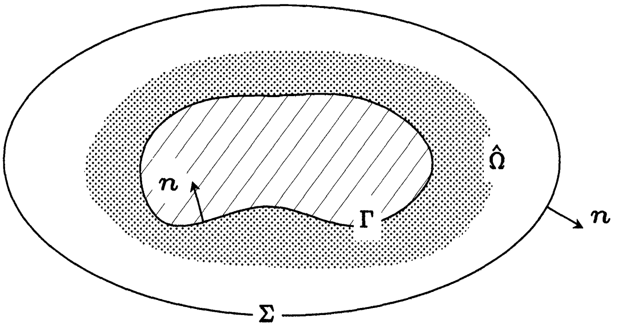


FIG. 2.2.

2.2.4. The new scattering problems. For every function τ that satisfies the assumptions stated above, we thus consider the following problem instead of problem $\mathcal{P}_\infty^{E/H}$:

$$(\mathcal{P}_\tau^{E/H})\quad \begin{array}{c} \operatorname{curl}(\zeta^{-1}\operatorname{curl} U) - \bar{\xi}\operatorname{grad}(\tau^{-1}\operatorname{div}\xi U) - \omega^2\xi U = 0 \quad \text{in } \Omega, \\ U \text{ satisfies the boundary condition } \mathcal{B}_\tau^{E/H} \text{ (i.e., (2.19) or (2.20))}, \\ U_S = U - U_I \text{ satisfies the radiation condition } \mathcal{R}_t \text{ (i.e., (2.21) and (2.22))}, \end{array}$$

where we recall that the incident wave U_I is a solution to Maxwell's equation (2.5) in the whole space $\mathbb{R}^3$. The two cases $\tau \equiv \infty$ and $\tau \in L^\infty(\Omega)$ correspond, respectively, to the classical and regularized equations.

2.3. Functional framework. In this paragraph, we make the exact significance of (2.12) precise as well as the associated functional framework in the classical or regularized situation. The interpretation of (2.12) used in the present paper consists of a weak formulation that involves a function space containing a divergence-free condition. This formulation is unusual as far as the classical Maxwell equation is concerned; we will show that it actually is consistent with the usual interpretation (in the sense of distributions). The main advantage of this formulation will appear in §4. It allows us to express problem $\mathcal{P}_\tau^{E/H}$ as a Fredholm equation.

2.3.1. Weak formulation. Consider the Frechet spaces

$$(2.25)\quad \begin{array}{l} \mathcal{F}_\tau^E(\Omega) = \{V \in H_{\text{loc}}(\text{curl};\Omega) \mid \operatorname{div}\xi V \in L^2_{\text{loc}}(\Omega), \\ \qquad\qquad \operatorname{div}\xi V = 0 \text{ in } \Omega \text{ if } \tau \equiv \infty, \text{ and } V \wedge n = 0 \text{ on } \Gamma\}, \\ \mathcal{F}_\tau^H(\Omega) = \{V \in H_{\text{loc}}(\text{curl};\Omega) \mid \operatorname{div}\xi V \in L^2_{\text{loc}}(\Omega), \\ \qquad\qquad \operatorname{div}\xi V = 0 \text{ in } \Omega \text{ if } \tau \equiv \infty, \text{ and } \xi V \cdot n = 0 \text{ on } \Gamma\}. \end{array}$$

When there is no ambiguity about the domain, we will simply denote these spaces by $\mathcal{F}_\tau^{E/H}$ (the case of a subdomain $\hat{\Omega}$ of Ω is used in Definition 2.5 below). In these spaces, the interpretation of the classical ($\tau \equiv \infty$) or regularized ($\tau \in L^\infty(\Omega)$) Maxwell equation is given by the following statement.

DEFINITION 2.5. *Let Σ be a closed regular surface surrounding the inhomogeneous medium and delimiting a bounded domain $\hat{\Omega} \subset \Omega$ (see Fig. 2.2). A function $U \in \mathcal{F}_\tau^E(\Omega)$ (respectively, $U \in \mathcal{F}_\tau^H(\Omega)$) is said to satisfy* (2.12) *and the electric boundary condition $\mathcal{B}_\tau^E$, i.e.,* (2.19) *(respectively, the magnetic boundary condition $\mathcal{B}_\tau^H$, i.e.,* (2.20)) *if the following conditions hold.*

(i) *In the region where the coefficients ξ, ζ, and τ are constant (in particular outside Σ), U satisfies in the sense of distributions the simplified equation* (2.15).

(ii) *For every field* $V \in \mathcal{F}_\tau^{E/H}(\hat{\Omega})$, *we have*

$$\begin{aligned}(2.26)\qquad &\int_{\hat\Omega} \zeta^{-1} \operatorname{curl} U \cdot \overline{\operatorname{curl} V} + \int_{\hat\Omega} \tau^{-1} \operatorname{div} \xi U \, \overline{\operatorname{div} \xi V} - \omega^2 \int_{\hat\Omega} \xi U \cdot \overline{V} \\ &+ \zeta_0^{-1} \int_\Sigma \operatorname{curl} U \cdot \overline{(V \wedge n)}\, d\gamma - t^{-1} \zeta_0^{-1} \int_\Sigma \operatorname{div} U \, \overline{(V \cdot n)}\, d\gamma = 0.\end{aligned}$$

This definition requires some comments in particular about the boundary integrals in (2.26). Let us first recall the following property about the regularity of U.

PROPOSITION 2.6. *If* $U \in \mathcal{F}_\tau^{E/H}(\Omega)$ *satisfies* (i), *then* U *is infinitely differentiable in the region where the coefficients* ξ, ζ, *and* τ *are constant and satisfies in the strong sense the simplified equation* (2.15).

Proof. This follows from standard results of interior regularity for elliptic equations. □

Remark 2.7. In (2.26), the integrals on Σ should be written as duality products between $H^{1/2}(\Sigma)$ and $H^{-1/2}(\Sigma)$ since $V \wedge n \in H^{-1/2}(\Sigma)^3$, $V \cdot n \in H^{-1/2}(\Sigma)$, and U is regular in the vicinity of Σ (by virtue of Proposition 2.6). For the sake of simplicity, we keep the integral notation.

At first glance, Definition 2.5 may seem intricate. As we will see below, the regularized equation cannot be interpreted in the sense of distributions when ξ is not regular enough (simply because $\mathcal{D}(\Omega)$ is not contained in $\mathcal{F}_\tau^{E/H}$). This explains why we need a weak formulation such as (2.26). The remainder of this section shows that Definition 2.5 agrees with the other possible interpretations of the problem.

2.3.2. Usual formulation in the classical case. The classical Maxwell equation (2.1) is usually interpreted in the sense of distributions. Indeed, for every function $U \in H_{\mathrm{loc}}(\mathrm{curl}; \Omega)$, we can define the distribution $\operatorname{curl}(\zeta^{-1} \operatorname{curl} U) \in \mathcal{D}'(\Omega)^3$ by setting

$$\langle \operatorname{curl}(\zeta^{-1} \operatorname{curl} U), V \rangle = \int_\Omega \zeta^{-1} \operatorname{curl} U \cdot \operatorname{curl} V \quad \forall V \in \mathcal{D}(\Omega)^3.$$

As a consequence, by considering the Frechet spaces

$$\begin{aligned}\mathcal{F}^E(\Omega) &= \{V \in H_{\mathrm{loc}}(\mathrm{curl}; \Omega) \mid V \wedge n = 0 \text{ on } \Gamma\}, \\ \mathcal{F}^H(\Omega) &= H_{\mathrm{loc}}(\mathrm{curl}; \Omega)\end{aligned}$$

that contain $\mathcal{F}_\infty^E(\Omega)$ and $\mathcal{F}_\infty^H(\Omega)$, respectively, defined by (2.25), we are led to a definition that is different from Definition 2.5.

DEFINITION 2.8. *A function* $U \in \mathcal{F}^E(\Omega)$ (*respectively,* $U \in \mathcal{F}^H(\Omega)$) *is said to satisfy the classical Maxwell equation* (2.1) *and the electric boundary condition* $\mathcal{B}_\infty^E$ (*respectively, the magnetic boundary condition* $\mathcal{B}_\infty^H$) *if*

$$(2.27)\qquad \int_\Omega \zeta^{-1} \operatorname{curl} U \cdot \overline{\operatorname{curl} V} - \omega^2 \int_\Omega \xi U \cdot \overline{V} = 0$$

for every field $V \in \mathcal{F}^E(\Omega)$ (*respectively,* $V \in \mathcal{F}^H(\Omega)$) *with compact support.*

Remark 2.9. A function $U \in \mathcal{F}^{E/H}(\Omega)$ that satisfies (2.27) obviously satisfies the divergence-free condition $\operatorname{div} \xi U = 0$ in the sense of distributions. Indeed, taking $V = \operatorname{grad} \overline{\varphi}$ for $\varphi \in \mathcal{D}(\Omega)$ in (2.27) yields

$$(2.28)\qquad \int_\Omega \xi U \cdot \operatorname{grad} \varphi = 0 \quad \forall \varphi \in \mathcal{D}(\Omega)$$

(which can be extended by density to every function $\varphi \in H_0^1(\Omega)$). Note that in the case of the magnetic field, this relation is also valid for every $\varphi \in \mathcal{D}(\overline{\Omega})$ (since $\operatorname{grad}\varphi \in \mathcal{F}^H(\Omega)$). By Green's formula, this implies that $\xi U \cdot n = 0$ on Γ. In short, a function $U \in \mathcal{F}^{E/H}(\Omega)$ that satisfies (2.27) belongs to $\mathcal{F}_\infty^{E/H}(\Omega)$.

Let $U \in \mathcal{F}^{E/H}(\Omega)$ satisfy (2.27). As in Proposition 2.6, we know from interior regularity results that outside the inhomogeneous medium, U is infinitely differentiable. Equation (2.5) is thus satisfied in the strong sense. Consider then the subdomain $\hat{\Omega}$ of Ω defined in Definition 2.5. Using Green's formula outside Σ in (2.27) shows that relation (2.26) (with $\tau \equiv \infty$) is valid for every field $V \in \mathcal{F}^{E/H}(\hat{\Omega})$.

We can now show in the classical case the equivalence between Definitions 2.5 and 2.8, which is a straightforward consequence of the following proposition.

PROPOSITION 2.10. *If $\tau \equiv \infty$, the following statements are equivalent:*

(i) $U \in \mathcal{F}^{E/H}(\Omega)$ *satisfies* (2.26) *for every* $V \in \mathcal{F}^{E/H}(\hat{\Omega})$;

(ii) $U \in \mathcal{F}_\infty^{E/H}(\Omega)$ *satisfies* (2.26) *for every* $V \in \mathcal{F}_\infty^{E/H}(\hat{\Omega})$.

Proof. The fact that (i) $\Rightarrow$ (ii) is obvious. If (i) is satisfied, then U clearly belongs to $\mathcal{F}_\infty^{E/H}(\Omega)$ (see Remark 2.9). The proof of the converse is based on a decomposition of vector fields given in Appendix B. Suppose that $U \in \mathcal{F}_\infty^{E/H}(\Omega)$ satisfies (2.26) for every function $V' \in \mathcal{F}_\infty^{E/H}(\hat{\Omega})$. Let $V \in \mathcal{F}^{E/H}(\hat{\Omega})$; by Lemma B.6 (taking $\mathcal{O} = \hat{\Omega}$ and $F_0 = \Gamma \cup \Sigma$, $F_1 = \emptyset$ if $V \in \mathcal{F}^E(\hat{\Omega})$, or $F_0 = \Sigma$ and $F_1 = \Gamma$ if $V \in \mathcal{F}^H(\hat{\Omega})$), we have

$$V = \operatorname{grad}\varphi + V'$$

where

$$\varphi \in \mathcal{H}^{E/H} \quad \text{with} \quad \begin{cases} \mathcal{H}^E = H_0^1(\hat{\Omega}), \\ \mathcal{H}^H = \left\{\psi \in H^1(\hat{\Omega}) \mid \psi_{|\Sigma} = 0\right\} \end{cases}$$

and $V' \in \mathcal{F}_\infty^{E/H}(\hat{\Omega})$. Since relation (2.26) is assumed to be satisfied for V', we deduce

$$\begin{aligned} \int_{\hat{\Omega}} \zeta^{-1} \operatorname{curl} U \cdot \overline{\operatorname{curl} V} - \omega^2 \int_{\hat{\Omega}} \xi U \cdot \overline{V} + \zeta_0^{-1} \int_\Sigma \operatorname{curl} U \cdot \overline{(V \wedge n)}\, d\gamma \\ = -\omega^2 \int_{\hat{\Omega}} \xi U \cdot \overline{\operatorname{grad}\varphi} + \zeta_0^{-1} \int_\Sigma \operatorname{curl} U \cdot \overline{(\operatorname{grad}\varphi \wedge n)}\, d\gamma. \end{aligned}$$

In the right-hand side of this equality, the integral on Σ vanishes since $\operatorname{grad}\varphi \wedge n = 0$ on Σ (see Lemma B.6). The integral on $\hat{\Omega}$ also vanishes; indeed, by Green's formula,

$$\int_{\hat{\Omega}} \xi U \cdot \overline{\operatorname{grad}\varphi} = -\int_{\hat{\Omega}} (\operatorname{div}\xi U)\, \overline{\varphi} + \int_{\Gamma \cup \Sigma} (\xi U \cdot n)\, \overline{\varphi}\, d\gamma,$$

where $\operatorname{div}\xi U = 0$. Moreover, in the case of the electric (respectively, magnetic) field, we have $\varphi = 0$ on $\Gamma \cup \Sigma$ (respectively, $\xi U \cdot n = 0$ on Γ and $\varphi = 0$ on Σ). This completes the proof. □

2.3.3. Some remarks about the regularized situation. Unlike the quantity $\operatorname{curl}(\zeta^{-1} \operatorname{curl} U)$, the term $\bar{\xi} \operatorname{grad}(\tau^{-1} \operatorname{div}\xi U)$ cannot be interpreted in the sense of distributions unless the coefficients are regular enough. In particular, if ξ is assumed Lipschitz continuous in Ω, then for every function $U \in H_{\mathrm{loc}}(\operatorname{div};\Omega)$, we can define the distribution $\bar{\xi} \operatorname{grad}(\tau^{-1} \operatorname{div}\xi U) \in \mathcal{D}'(\Omega)^3$ by the relation

$$\langle \bar{\xi} \operatorname{grad}(\tau^{-1} \operatorname{div}\xi U), V \rangle = -\int_\Omega \tau^{-1} \operatorname{div}\xi U \operatorname{div}\bar{\xi} V \quad \forall V \in \mathcal{D}(\Omega)^3.$$

In this case, Definition 2.5 agrees with this interpretation: since every field $V \in \mathcal{D}(\hat{\Omega})^3$ belongs to $\mathcal{F}_\tau^{E/H}(\hat{\Omega})$, relation (2.26) implies that U satisfies the regularized Maxwell equation (2.12) in Ω in the sense of distributions.

Remark 2.11. Definition 2.5 implies some transmission conditions on a surface of discontinuity of the coefficients. Indeed, the quantities $U \wedge n$ and $\xi U \cdot n$ must be continuous across this boundary for $\operatorname{curl} U \in L^2_{\text{loc}}(\Omega)$ and $\operatorname{div} \xi U \in L^2_{\text{loc}}(\Omega)$. Moreover, if this boundary is $C^{1,1}$ and if ξ is Lipschitz continuous on both sides, it is readily seen that the quantity $\zeta^{-1} \operatorname{curl} U \wedge n + n \tau^{-1} \operatorname{div} \xi U$ (defined in $H^{-1/2}$ by means of Green's formula) is also continuous.

2.4. Equivalence between the classical and regularized problems.

Let $U \in \mathcal{F}_\infty^{E/H}$ be a solution of the classical scattering problem $\mathcal{P}_\infty^{E/H}$. For every function τ that satisfies the assumptions stated in §2.2, U belongs to $\mathcal{F}_\tau^{E/H}$ and satisfies in the sense of Definition 2.5 the regularized Maxwell equation as well as the boundary conditions $\mathcal{B}_\tau^{E/H}$ (see Proposition 2.10). Moreover, the associated scattered wave $U_S = U - U_I$ clearly satisfies the transverse and radial radiation conditions $\mathcal{R}_t$ since $\operatorname{div} U_I = 0$. This shows that every solution to the classical scattering problem $\mathcal{P}_\infty^{E/H}$ is also a solution to problem $\mathcal{P}_\tau^{E/H}$ (for the same incident wave).

The aim of the present paper is to prove the converse of this statement and to present a convenient method for its numerical solution. More precisely, by adding some suitable conditions on the coefficients ξ, ζ, and τ (which ensure the uniqueness of the solution to $\mathcal{P}_\tau^{E/H}$ (see §3.1)), we will prove the following result.

THEOREM 2.12. *Let ξ, ζ, and τ be chosen such that $\mathcal{P}_\tau^{E/H}$ admits at most one solution. Let U_I be a given divergence-free incident wave, i.e., a solution to Maxwell equation* (2.5) *in the whole space* $\mathbb{R}^3$.

(i) *The scattering problem $\mathcal{P}_\tau^{E/H}$ has exactly one solution in $\mathcal{F}_\tau^{E/H}$.*

(ii) *This solution does not depend on τ.*

The last statement actually amounts to saying that the solution to $\mathcal{P}_\tau^{E/H}$ satisfies the divergence-free condition $\operatorname{div} \xi U = 0$ in Ω. It is a straightforward consequence of the well-posedness of $\mathcal{P}_\tau^{E/H}$ (i.e., point (i)) and the fact that a solution to $\mathcal{P}_\infty^{E/H}$ is also a solution to $\mathcal{P}_\tau^{E/H}$. The remainder of the paper consists of proving (i).

3. Uniqueness.

The aim of this section is to prove that for every function τ, problem $\mathcal{P}_\tau^{E/H}$ has at most one solution. By linearity, this amounts to proving that if the incident wave vanishes, the only solution to $\mathcal{P}_\tau^{E/H}$ is 0.

3.1. Some restrictions for uniqueness.

Thus far, we did not make any assumption regarding the regularity of the coefficients ξ, ζ, and τ (they were just assumed bounded). In this general framework, the uniqueness of the solution to problem $\mathcal{P}_\tau^{E/H}$ cannot be ensured. We thus have to add some restrictive conditions that allow us to prove uniqueness. These assumptions depend on whether the coefficients are real-valued or not.

More precisely, we will suppose in this section that the domain Ω can be decomposed as shown in Fig. 3.1 into several disjointed (and nonempty) open sets $\{\Omega_n \mid n = \overline{0,N}\}$ with piecewise $C^{1,1}$ boundaries; Ω then appears as the interior of $\bigcup_{n=\overline{0,N}} \overline{\Omega_n}$ (where $\overline{\Omega_n}$ denotes the closure of Ω_n). For the convenience of the presentation (and because of the particular role of infinity), we define one of these sets, say Ω_0, to be the exterior of a ball containing all the inhomogeneities. In addition to the hypotheses stated in §2, we will assume that in every subdomain Ω_n, each of the

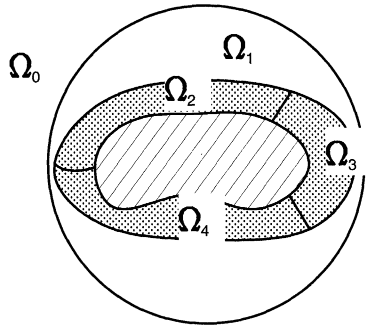


FIG. 3.1. *Decomposition of Ω into subdomains.*

coefficients ξ, ζ, and τ satisfies one of the following conditions.

1. Its restriction to Ω_n is real-valued and defines a Lipschitz-continuous function in Ω_n.

2. Its restriction to Ω_n is a Lipschitz-continuous complex-valued function whose imaginary part is positive almost everywhere in Ω_n.

3. An exceptional condition for τ only: $\tau \equiv \infty$ in the whole domain Ω.

Of course, these conditions can be satisfied independently by each coefficient; in a given subdomain Ω_n, ξ may satisfy 1 whereas ζ satisfies 2 and τ satisfies 3.

Let us mention that the assumption concerning the Lipschitz regularity in the second condition may be removed for the coefficients ζ and τ but not for ξ (since it is involved in the definition of the function space $\mathcal{F}_\tau^{E/H}$).

Our purpose now is to prove the following uniqueness result.

THEOREM 3.1. *Let ξ, ζ, and τ satisfy the hypotheses stated above. If $U \in \mathcal{F}_\tau^{E/H}$ is a solution to $\mathcal{P}_\tau^{E/H}$ with no incident wave (i.e., $U_I \equiv 0$), then $U \equiv 0$.*

The main tools for the proof of this statement are on one hand the energy conservation law presented in §3.2 and on the other hand some results concerning the unique continuation principle for elliptic (or subelliptic) equations. The detailed proof is given in §§3.3 and 3.4.

3.2. Energy flux and radiation conditions. Let $R > 0$ be chosen such that the coefficients ξ, ζ, and τ are constant outside $\Sigma_R = \{x \in \mathbb{R}^3 \mid \|x\| = R\}$. We denote $\hat{\Omega}_R = \{x \in \Omega \mid \|x\| < R\}$. Let $U \in \mathcal{F}_\tau^{E/H}$ satisfy (2.12) and the boundary condition $\mathcal{B}_\tau^{E/H}$ (in the sense of Definition 2.5). By taking the imaginary part of (2.26) with $V = \bar{U}$, $\Sigma = \Sigma_R$, and $\hat{\Omega} = \hat{\Omega}_R$, we obtain

$$\mathcal{F}_R(U) + \mathcal{J}_R(U) = 0, \quad \text{where} \tag{3.1}$$

$$\mathcal{F}_R(U) = \mathcal{F}_R^s(U) + \mathcal{F}_R^p(U) \quad \text{with} \tag{3.2}$$

$$\mathcal{F}_R^s(U) = -\mathfrak{Im}\left\{\zeta_0^{-1} \int_{\Sigma_R} \operatorname{curl} U \cdot \overline{(U \wedge n)}\, d\gamma\right\}, \tag{3.3}$$

$$\mathcal{F}_R^p(U) = \mathfrak{Im}\left\{t^{-1}\zeta_0^{-1} \int_{\Sigma_R} \operatorname{div} U\, \overline{(U \cdot n)}\, d\gamma\right\}, \quad \text{and} \tag{3.4}$$

$$\mathcal{J}_R(U) = \mathfrak{Im} \int_{\hat{\Omega}_R} \left[\omega^2 \xi \|U\|^2 - \zeta^{-1} \|\operatorname{curl} U\|^2 - \tau^{-1} |\operatorname{div} \xi U|^2\right]. \tag{3.5}$$

For the classical Maxwell equation ($\tau \equiv \infty$), $\mathcal{F}_R(U)$ and $\mathcal{J}_R(U)$, respectively, represent (up to a factor $(2\omega)^{-1}$) the mean outgoing energy flux (per time period) through the boundary Σ_R and the energy loss by Joule effect in $\hat{\Omega}_R$. Equation (3.1) stands for energy conservation. We keep the same terminology for the regularized formulation.

The radiation conditions allow us to express the energy flux at infinity as follows.

LEMMA 3.2. *If U satisfies the outgoing radiation conditions* (2.21) *and* (2.22), *then the outgoing energy flux at infinity is given by*

$$\lim_{R\to\infty} \mathcal{F}_R(U) = \mathcal{F}^s_\infty(U) + \mathcal{F}^p_\infty(U),$$

where $\mathcal{F}^s_\infty$ if $\mathfrak{Im}\, k_s = 0$ is defined as

$$\mathcal{F}^s_\infty(U) = \frac{1}{2k_s\zeta_0} \lim_{R\to\infty} \int_{\Sigma_R} \left\{ \|\operatorname{curl} U \wedge n\|^2 + |k_s|^2 \, \|n \wedge (U \wedge n)\|^2 \right\} d\gamma$$

and $\mathcal{F}^s_\infty(U) = 0$ otherwise. Similarly, if $t \neq 0$ and $\mathfrak{Im}\, k_p = 0$, then $\mathcal{F}^p_\infty(U)$ reads

$$\mathcal{F}^p_\infty(U) = \frac{1}{2tk_p\zeta_0} \lim_{R\to\infty} \int_{\Sigma_R} \left\{ |\operatorname{div} U|^2 + |k_p|^2 \, |U \cdot n|^2 \right\} d\gamma;$$

otherwise, $\mathcal{F}^p_\infty(U) = 0$.

Proof. If k_s is real (respectively, $t \neq \infty$ and $k_p \in \mathbb{R}^+$), this follows from the radiation conditions and the formulas

$$\begin{aligned} \|\operatorname{curl} U \wedge n - ik_s\, n \wedge (U \wedge n)\|^2 &= \|\operatorname{curl} U \wedge n\|^2 + |k_s|^2 \, \|n \wedge (U \wedge n)\|^2 \\ &\quad + 2k_s \,\mathfrak{Im}\{\operatorname{curl} U \cdot \overline{(U \wedge n)}\}, \\ |\operatorname{div} U - ik_p\, U \cdot n|^2 &= |\operatorname{div} U|^2 + |k_p|^2 \, |U \cdot n|^2 - 2k_p \,\mathfrak{Im}\{\operatorname{div} U\, \overline{(U \cdot n)}\}. \end{aligned}$$

If $\mathfrak{Im}\, k_s \neq 0$ (or $\mathfrak{Im}\, k_p \neq 0$, or $t = \infty$), this result readily follows from the radiation conditions using the Schwarz inequality. □

PROPOSITION 3.3. *If $U \in \mathcal{F}^{E/H}_\tau$ is a solution to problem $\mathcal{P}^{E/H}_\tau$ with $U_I \equiv 0$, then*

$$\lim_{R\to\infty} \mathcal{F}_R(U) = \lim_{R\to\infty} \mathcal{J}_R(U) = 0. \tag{3.6}$$

Proof. First notice that $\mathcal{J}_R(U)$ is a nonnegative and nondecreasing function of R. Moreover, by Lemma 3.2, $\lim_{R\to\infty} \mathcal{F}_R(U)$ is also nonnegative. The conclusion follows from relation (3.1). □

A straightforward. consequence of this result and of the definition (3.5) of $\mathcal{J}_R$ follows.

COROLLARY 3.4. *For every open subset $\mathcal{O} \subset \Omega$, we have the properties*

$$\mathfrak{Im}\,\xi \geq 0 \quad \text{a.e.} \quad \text{in } \mathcal{O} \quad \Longrightarrow \quad U_{|\mathcal{O}} = 0, \tag{3.7}$$

$$\mathfrak{Im}\,\zeta \geq 0 \quad \text{a.e.} \quad \text{in } \mathcal{O} \quad \Longrightarrow \quad (\operatorname{curl} U)_{|\mathcal{O}} = 0, \tag{3.8}$$

$$\mathfrak{Im}\,\tau \geq 0 \quad \text{a.e.} \quad \text{in } \mathcal{O} \quad \Longrightarrow \quad (\operatorname{div} \xi U)_{|\mathcal{O}} = 0. \tag{3.9}$$

3.3. First step of the proof of Theorem 3.1: The particular case of Ω_0. We prove in this paragraph that *if $U \in \mathcal{F}^{E/H}_\tau$ is a solution to $\mathcal{P}^{E/H}_\tau$ with $U_I \equiv 0$, then U vanishes in Ω_0.*

Recall that Ω_0 is the exterior of a ball containing all the inhomogeneities. In this subdomain, the field U satisfies (see Proposition 2.6)

$$\operatorname{curl}(\operatorname{curl} U) - t^{-1} \operatorname{grad}(\operatorname{div} U) - k_s^2 U = 0. \tag{3.10}$$

First, notice that if $\mathfrak{Im}\,\xi_0 \neq 0$, the result follows from property (3.7). In what follows, we thus assume that ξ_0 is real.

3.3.1. A scalar result. Let us prove that $\operatorname{div} U = 0$ in Ω_0. This is obvious if $\tau_0 = \infty$ (since the condition $\operatorname{div} \xi U = 0$ in Ω is contained in the definition of the space $\mathcal{F}_\tau^{E/H}$). Moreover, if $\mathfrak{Im}\, \tau_0 \neq 0$, this results from property (3.9). Consequently, we suppose below that $\tau_0 \in \mathbb{R}^+$.

Setting $\varphi = \operatorname{div} U$ and taking the divergence of (3.10) yields

$$\Delta\varphi + k_p^2\varphi = 0, \quad \text{where } k_p = \omega\sqrt{\tau_0/\xi_0} \in \mathbb{R}^+.$$

The fact that φ vanishes is based on the following lemma (due to Rellich [21]), which concerns the asymptotic behavior at infinity of the solutions to the Helmholtz equation.

LEMMA 3.5. *Let φ be a solution to equation $\Delta\varphi + k^2\varphi = 0$ outside a sphere Σ_{R_0} of radius R_0 (where k is a positive constant). If φ satisfies the condition*

$$\lim_{R\to\infty} \int_{\Sigma_R} |\varphi|^2 d\gamma = 0, \tag{3.11}$$

then $\varphi \equiv 0$ outside Σ_{R_0}.

From (3.6) and the expression of $\lim_{R\to\infty} \mathcal{F}_R$ given in Lemma 3.2, we see that condition (3.11) is satisfied. As a consequence, $\varphi = \operatorname{div} U$ vanishes in Ω_0.

3.3.2. Back to the vector problem. If $\mathfrak{Im}\, \zeta_0 \neq 0$, property (3.8) implies moreover that $\operatorname{curl} U = 0$ in Ω_0. By (3.10), we deduce that U vanishes. On the other hand, if $\zeta_0 \in \mathbb{R}^+$, the fact that $\operatorname{div} U = 0$ allows us to replace t^{-1} by any other value in (3.10). In particular, by choosing $t = 1$, we see that U satisfies the vector Helmholtz equation

$$\Delta U + k_s^2 U = 0, \quad \text{where } k_s = \omega\sqrt{\zeta_0\xi_0} \in \mathbb{R}^+.$$

And here again, using (3.6) and the expression of $\lim_{R\to\infty} \mathcal{F}_R$ given in Lemma 3.2, we have

$$\lim_{R\to\infty} \int_{\Sigma_R} \|U\|^2 d\gamma = 0$$

by virtue of the relation $\|U\|^2 = \|n \wedge (U \wedge n)\|^2 + |U \cdot n|^2$. (Note that if $\tau_0 = \infty$, the fact that $\lim_{R\to\infty} \int_{\Sigma_R} |U \cdot n|^2 d\gamma = 0$ is contained in the radiation conditions.) We can thus apply Lemma 3.5 to each component of U. The conclusion follows as $U = 0$ in Ω_0.

3.4. Second step of the proof: The other subdomains. To prove that U vanishes in every subdomain Ω_n, $n > 0$, we will use either the results of Corollary 3.4 or a unique continuation technique, which is the object of Proposition 3.6 below, depending on whether the coefficients ξ, ζ, and τ are real-valued or not.

3.4.1. The unique continuation principle. Consider a given subdomain Ω_n $(n > 0)$ of Ω, and suppose that in Ω_n the coefficients ξ, ζ, and τ all satisfy condition 1 given in §3.1. These are Lipschitz-continuous real-valued functions. In this case, we have seen in §2.3 that a solution $U \in \mathcal{F}_\tau^{E/H}$ to $\mathcal{P}_\tau^{E/H}$ satisfies

$$\operatorname{curl}(\zeta^{-1} \operatorname{curl} U) - \xi \operatorname{grad}(\tau^{-1} \operatorname{div} \xi U) - \omega^2 \xi U = 0 \quad \text{in } \Omega_n \tag{3.12}$$

in the sense of distributions. Our aim is to prove the following statement.

PROPOSITION 3.6. *Let ξ, ζ, and τ all satisfy condition* 1 *of* §3.1. *If a solution $U \in \mathcal{F}_\tau^{E/H}$ to* (3.12) *vanishes in a (arbitrary small) ball $B \subset \Omega_n$, then $U = 0$ in the whole domain Ω_n.*

Proof. We first prove that $\operatorname{div} \xi U = 0$ in Ω_n, from which we will be able to deduce that $U \equiv 0$.

(i) Consider the scalar function $\varphi = \tau^{-1} \operatorname{div} \xi U \in L^2(\Omega_n)$. By taking the divergence of (3.12), we see that φ satisfies

$$\operatorname{div}(\xi \operatorname{grad} \varphi) + \omega^2 \tau \varphi = 0 \quad \text{in } \Omega_n$$

in the sense of distributions. From interior regularity for (very weak) solutions to elliptic equations (see, e.g., Nečas [19]), we infer that in every open set $\mathcal{O}$ such that $\overline{\mathcal{O}} \subset \Omega_n$, we have $\varphi_{|\mathcal{O}} \in H^2(\mathcal{O})$. The classical results of unique continuation for second-order elliptic scalar equations (see, e.g., Kenig [11]) then show that if $\varphi = 0$ in a ball $B \subset \Omega_n$, then φ vanishes everywhere in Ω_n. We have thus proved that $\operatorname{div} \xi U \equiv 0$.

(ii) This result implies in particular that U is a solution to the classical Maxwell equation

$$\operatorname{curl}(\zeta^{-1} \operatorname{curl} U) - \omega^2 \xi U = 0 \quad \text{in } \Omega_n. \tag{3.13}$$

To see that U vanishes, we use the strong unique continuation principle proved by Vogelsang [23], who deals with the original first-order system of Maxwell's equations instead of (3.13). By setting, for instance,

$$H = U \quad \text{and} \quad E = \frac{-1}{i\omega\zeta} \operatorname{curl} U,$$

we clearly have

$$\operatorname{curl} E = i\omega\xi H \quad \text{and} \quad \operatorname{curl} H = -i\omega\zeta E,$$

where both fields E and H obviously belong to $H(\operatorname{curl}; \Omega_n)$. To apply [23], we simply have to check that E and H actually belong to $H^1(\mathcal{O})^3$ for every open set $\mathcal{O}$ such that $\overline{\mathcal{O}} \subset \Omega_n$. From [23], we know in this case that if E and H vanish in a ball $B \subset \mathcal{O}$, then they vanish in the whole domain $\mathcal{O}$. To prove that $(E, H) \in H^1(\mathcal{O})^3$, note that $\operatorname{div} \zeta E = 0$ and $\operatorname{div} \xi H = 0$, which shows that E and H also belong to $H(\operatorname{div}; \Omega_n)$ since ζ and ξ are Lipschitz continuous. The conclusion follows from the H^1 interior regularity of functions of $H(\operatorname{curl}) \cap H(\operatorname{div})$ (see [8]). □

3.4.2. End of the proof of Theorem 3.1. We now come back to the general situation where ξ, ζ, and τ satisfy one of the three conditions stated in §3.1. Let Ω_n ($n > 0$) be a given subdomain of Ω. We prove below that *if a solution U to $\mathcal{P}_\tau^{E/H}$ vanishes in some subdomain $\Omega_{n'}$ adjacent to Ω_n, then it also vanishes in Ω_n.* The statement of Theorem 3.1 follows, for we already know that $U = 0$ in Ω_0 (§3.3) and Ω is assumed connected (every subdomain Ω_n is linked to Ω_0 by a path contained in Ω).

(i) First, note that if ξ satisfies condition 2 of §3.1 in Ω_n (i.e., its imaginary part is positive almost everywhere), the result obviously follows from property (3.7). We assume now that ξ is a real-valued and Lipschitz-continuous function (condition 1).

(ii) If τ satisfies condition 2 in Ω_n, we know that $\operatorname{div}\xi U = 0$ in Ω_n (property (3.9)) and, consequently, U satisfies the regularized Maxwell equation (3.12) in Ω_n for any other function τ. The same holds if τ is infinite (condition 3). Therefore, it is enough to deal with the case of condition 1 for τ.

(iii) Similarly, if ζ satisfies condition 2, we know that $\operatorname{curl} U = 0$ in Ω_n, which shows that we can modify (3.12) by replacing ζ by any other function; U will always be a solution to the modified equation. Here again, we can suppose that ζ satisfies condition 1.

(iv) We are finally in the context of Proposition 3.6. Assume that in one of the subdomains adjacent to Ω_n, say $\Omega_{n'}$, the field U vanishes. To apply Proposition 3.6, consider the domain $\tilde{\Omega}_n = \Omega_n \cup B$, where B is a small ball centered at a point of a regular part (i.e., $C^{1,1}$) of $\partial\Omega_n \cap \partial\Omega_{n'}$. Let $\tilde{\xi}$, $\tilde{\zeta}$, and $\tilde{\tau}$ denote, respectively, Lipschitz-continuous (and real-valued) extensions of $\xi_{|\Omega_n}$, $\zeta_{|\Omega_n}$, and $\tau_{|\Omega_n}$ in $\tilde{\Omega}_n$ (these extensions are built for instance by the standard reflection technique). It is then clear that if we replace in $\tilde{\Omega}_n$ these three functions by $\tilde{\xi}$, $\tilde{\zeta}$, and $\tilde{\tau}$, the field U still remains a solution to (3.12) in $\tilde{\Omega}_n$ in the sense of distributions (see Remark 2.11). The statement of Proposition 3.6 then applies in the new subdomain $\tilde{\Omega}_n$. This completes the proof.

4. Existence of a solution.

4.1. The method of coupling by integral representation.

We present in this paragraph the so-called "method of coupling between variational formulation and integral representation," which consists of reducing problem $\mathcal{P}_\tau^{E/H}$ to an equivalent problem set in a bounded domain. This will allow us to use the Fredholm alternative to prove the existence result. The cases of classical or regularized equation are dealt with simultaneously.

Let F be a closed regular surface surrounding the perfect conductor and located in the region where the coefficients ξ, ζ, and τ are constant (see Fig. 4.1). If U is a solution to $\mathcal{P}_\tau^{E/H}$, we know (see Proposition 2.6) that outside F, the scattered wave $U_S = U - U_I$ satisfies the simplified equation

$$\operatorname{curl}(\operatorname{curl} U_S) - t^{-1}\operatorname{grad}(\operatorname{div} U_S) - k_s^2 U_S = 0 \tag{4.1}$$

as well as the radiation condition $\mathcal{R}_t$, i.e., (2.21) and (2.22). We prove in Appendix A (Proposition A.7) that U_S consequently satisfies the following integral representation outside F:

$$\begin{aligned} U_S(x) = &\int_F \mathbb{G}_t(x-y)\left\{\operatorname{curl} U_S(y) \wedge n_y + t^{-1} n_y \operatorname{div} U_S(y)\right\} d\gamma_y \\ &- \int_F (\operatorname{curl}_y \mathbb{G}_t(x-y))\left\{U_S(y) \wedge n_y\right\} d\gamma_y \\ &- t^{-1}\int_F (\operatorname{div}_y \mathbb{G}_t(x-y))^T\left\{U_S(y)\cdot n_y\right\} d\gamma_y, \end{aligned} \tag{4.2}$$

where $\mathbb{G}_t$ is the outgoing Green tensor associated with (4.1) (see Proposition A.1). In what follows, this formula will be written in the form

$$U_S = \mathcal{I}_t[F; U_S] \quad \text{outside } F. \tag{4.3}$$

Notice that since U_I is a solution to

$$\operatorname{curl}(\operatorname{curl} U_I) - k_s^2 U_I = 0 \quad \text{in } \mathbb{R}^3,$$

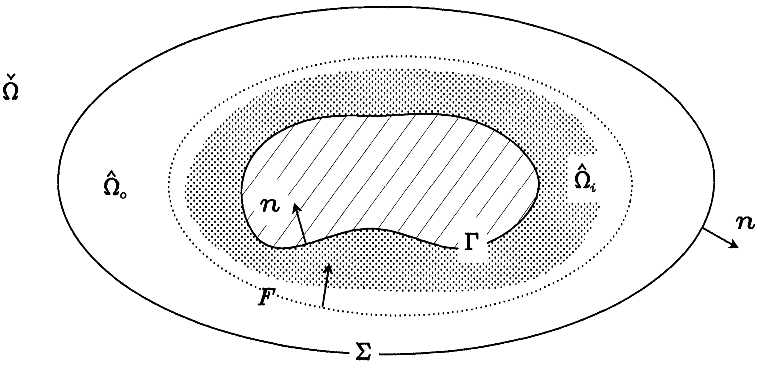


FIG. 4.1. *Reduction to a bounded domain.*

it satisfies in particular (4.1) inside F. We deduce from the integral representation formula in a bounded domain (Proposition A.5) that

$$0 = \mathcal{I}_t[F; U_I] \quad \text{outside } F. \tag{4.4}$$

As a consequence, the integral representation (4.3) can be written in terms of the total field U as

$$U = U_I + \mathcal{I}_t[F; U] \quad \text{outside } F. \tag{4.5}$$

Consider then a closed regular surface Σ surrounding F (and which has no point in common with F). We denote by $\hat{\Omega}$ the bounded part of Ω delimited by Σ, $\hat{\Omega}_i$ and $\hat{\Omega}_o$ the parts of $\hat{\Omega}$ located, respectively, inside and outside F and finally $\check{\Omega}$ the exterior of Σ. The orientation of the unit normal on the surfaces Γ, F, and Σ is shown in Fig. 4.1. Let us define two boundary operators on Σ that involve, respectively, the tangential and normal components of the field as

$$\begin{aligned} T_\lambda U &= \{\operatorname{curl} U \wedge n + \lambda\, n \wedge (U \wedge n)\}_{|\Sigma} \quad \text{and} \\ N_\nu U &= \{\operatorname{div} U + \nu\, U \cdot n\}_{|\Sigma}, \end{aligned} \tag{4.6}$$

where λ and ν are complex parameters. If U is a solution to $\mathcal{P}_\tau^{E/H}$, it clearly satisfies $T_\lambda U = T_\lambda\,(U_I + \mathcal{I}_t[F; U])$ and $N_\nu U = N_\nu\,(U_I + \mathcal{I}_t[F; U])$ by virtue of (4.5). It follows that the restriction $\hat{U}$ of U to $\hat{\Omega}$ is a solution to the problem (set in the bounded domain $\hat{\Omega}$)

$$\begin{gathered} \operatorname{curl}(\zeta^{-1} \operatorname{curl} \hat{U}) - \bar{\xi} \operatorname{grad}(\tau^{-1} \operatorname{div} \xi \hat{U}) - \omega^2 \xi \hat{U} = 0 \quad \text{in } \hat{\Omega}, \\ \hat{U} \text{ satisfies } \mathcal{B}_\tau^{E/H} \quad \text{on } \Gamma, \\ T_\lambda(\hat{U} - \mathcal{I}_t[F; \hat{U}]) = T_\lambda U_I \quad \text{and} \quad t^{-1} N_\nu(\hat{U} - \mathcal{I}_t[F; \hat{U}]) = t^{-1} N_\nu U_I \quad \text{on } \Sigma. \end{gathered} \tag{4.7}$$

Note that in the condition on Σ, called the "coupling condition," the relation that involves N_ν is taken into account only if $t \neq \infty$.

The natural function space associated with this problem is obtained by taking the functions of $\mathcal{F}_\tau^{E/H}(\hat{\Omega})$ that are regular enough in the vicinity of Σ (because of the coupling condition on this surface). We thus define the Hilbert spaces

$$\begin{aligned} \mathcal{H}_\tau^E = \{ & V \in H(\operatorname{curl}; \hat{\Omega}) \mid \operatorname{div} \xi V \in L^2(\hat{\Omega}),\ \operatorname{div} \xi V = 0 \text{ in } \hat{\Omega} \text{ if } \tau \equiv \infty, \\ & (V \wedge n)_{|\Gamma} = 0,\ (V \wedge n)_{|\Sigma} \in L^2(\Sigma)^2 \text{ and } t^{-1}(V \cdot n)_{|\Sigma} \in L^2(\Sigma)\}, \\ \mathcal{H}_\tau^H = \{ & V \in H(\operatorname{curl}; \hat{\Omega}) \mid \operatorname{div} \xi V \in L^2(\hat{\Omega}),\ \operatorname{div} \xi V = 0 \text{ in } \hat{\Omega} \text{ if } \tau \equiv \infty, \\ & (\xi V \cdot n)_{|\Gamma} = 0,\ (V \wedge n)_{|\Sigma} \in L^2(\Sigma)^2 \text{ and } t^{-1}(V \cdot n)_{|\Sigma} \in L^2(\Sigma)\}. \end{aligned} \tag{4.8}$$

To write a variational formulation of problem (4.7) in these spaces, we have to modify the coupling condition on Σ and, more precisely, the expression of the integral representation $\mathcal{I}_t[F;\hat{U}]$. Indeed, the quantity $\mathcal{I}_t[F;V]$ is not defined for every function $V \in \mathcal{H}_\tau^{E/H}$ because of the term $(\operatorname{curl} V \wedge n + t^{-1} n \operatorname{div} V)_{|F}$ whose definition requires an additional condition such as $\operatorname{curl}(\operatorname{curl} V) - t^{-1}\operatorname{grad}(\operatorname{div} V) \in L^2(\hat{\Omega}_o)$ (see Remark A.6). To remove this term, consider a regular right inverse R of the trace operator from $\hat{\Omega}_o$ to F, i.e., a linear operator that maps every regular function φ defined on F onto a regular function $R\varphi$ defined in $\hat{\Omega}_o$ that satisfies $(R\varphi)_{|F} = \varphi$. Suppose in addition that R vanishes on Σ (i.e., $(R\varphi)_{|\Sigma} = 0$ for every φ). Integrating by parts the embarrassing term in (4.2) yields

$$\begin{aligned}\int_F \mathbb{G}_t(x,\cdot)\left\{\operatorname{curl} U_S \wedge n + t^{-1} n \operatorname{div} U_S\right\} d\gamma \\ = -\int_{\hat{\Omega}_o} R\mathbb{G}_t(x,\cdot)\left\{\operatorname{curl}\operatorname{curl} U_S - t^{-1}\operatorname{grad}\operatorname{div} U_S\right\} \\ + \int_{\hat{\Omega}_o}(\operatorname{curl} R\mathbb{G}_t(x,\cdot))\operatorname{curl} U_S + t^{-1}\int_{\hat{\Omega}_o}(\operatorname{div} R\mathbb{G}_t(x,\cdot))^T \operatorname{div} U_S\end{aligned}$$

for x located outside F, where we have denoted $\mathbb{G}_t(x,y) = \mathbb{G}_t(x-y)$. Using the fact that U_S satisfies (4.1) in $\hat{\Omega}_o$, the integral representation (4.2) can be written equivalently as $U_S = \mathcal{I}_t^R[F;U_S]$ outside F, where

$$\begin{aligned}\mathcal{I}_t^R[F;U_S](x) = -k_s^2 \int_{\hat{\Omega}_o} R\mathbb{G}_t(x,\cdot)\, U_S \\ + \int_{\hat{\Omega}_o}(\operatorname{curl} R\mathbb{G}_t(x,\cdot))\operatorname{curl} U_S + t^{-1}\int_{\hat{\Omega}_o}(\operatorname{div} R\mathbb{G}_t(x,\cdot))^T \operatorname{div} U_S \\ - \int_F(\operatorname{curl}\mathbb{G}_t(x,\cdot))\left\{U_S \wedge n\right\} d\gamma - t^{-1}\int_F(\operatorname{div}\mathbb{G}_t(x,\cdot))^T\left\{U_S\cdot n\right\} d\gamma.\end{aligned} \tag{4.9}$$

More generally, for every field U that satisfies (4.1) in $\hat{\Omega}_o$, we clearly have

$$\mathcal{I}_t[F;U] = \mathcal{I}_t^R[F;U] \quad \text{outside } F. \tag{4.10}$$

Substituting this relation in the coupling condition on Σ in (4.7), we are now able to define the reduced problem, denoted by $\hat{\mathcal{P}}_\tau^{E/H}$ in what follows, as

$$(\hat{\mathcal{P}}_\tau^{E/H}) \quad \begin{array}{c}\text{Find } \hat{U} \in \mathcal{H}_\tau^{E/H} \text{ such that} \\ \operatorname{curl}(\zeta^{-1}\operatorname{curl}\hat{U}) - \bar{\xi}\operatorname{grad}(\tau^{-1}\operatorname{div}\xi\hat{U}) - \omega^2\xi\hat{U} = 0 \quad \text{in } \hat{\Omega}, \\ \hat{U} \text{ satisfies } \mathcal{B}_\tau^{E/H} \quad \text{on } \Gamma, \\ T_\lambda(\hat{U} - \mathcal{I}_t^R[F;\hat{U}]) = T_\lambda U_I \quad \text{and} \quad t^{-1}N_\nu(\hat{U} - \mathcal{I}_t^R[F;\hat{U}]) = t^{-1}N_\nu U_I \quad \text{on } \Sigma,\end{array}$$

where we recall that the classical or regularized equation in $\hat{\Omega}$ has to be understood in the sense of Definition 2.5.

4.2. Equivalence between the initial and the reduced problems. The link between the initial problem $\mathcal{P}_\tau^{E/H}$ and the new one $\hat{\mathcal{P}}_\tau^{E/H}$ may be stated as follows.

PROPOSITION 4.1. *Let λ and ν be two complex parameters chosen such that $\mathfrak{Im}(\lambda k_s^{-2}) < 0$ and $\mathfrak{Im}(\nu k_p^{-2}) < 0$ (if $t \neq \infty$). Problem $\mathcal{P}_\tau^{E/H}$ admits at least (respectively, at most) one solution if and only if the same holds for $\hat{\mathcal{P}}_\tau^{E/H}$. Moreover,*

(i) *if U is a solution to $\mathcal{P}_\tau^{E/H}$, then $\hat{U} = U_{|\hat{\Omega}}$ is a solution to $\hat{\mathcal{P}}_\tau^{E/H}$;*

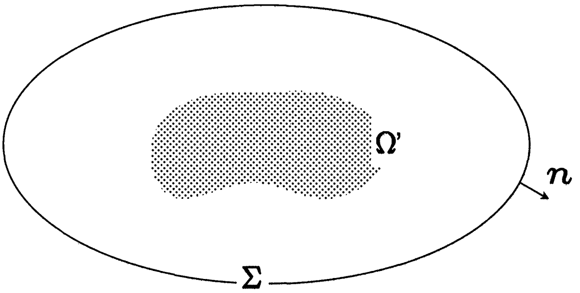


FIG. 4.2.

(ii) *if $\hat{U}$ is a solution to $\hat{\mathcal{P}}_T^{E/H}$, then the field U defined by*

$$U = \hat{U} \quad \textit{in } \hat{\Omega} \quad \textit{and} \quad U = U_I + \mathcal{I}_t[F;\hat{U}] \quad \textit{in } \check{\Omega} \tag{4.11}$$

is a solution to $\mathcal{P}_T^{E/H}$.

Proof. First, note that it is enough to prove the equivalence of $\mathcal{P}_T^{E/H}$ and $\hat{\mathcal{P}}_T^{E/H}$ for the existence of a solution. The equivalence for the uniqueness readily follows using the uniqueness of the continuation of U outside F. (A solution to $\mathcal{P}_T^{E/H}$ that vanishes in a vicinity of F must vanish everywhere outside F by virtue of the integral representation (4.2).)

(i) If U is a solution to $\mathcal{P}_T^{E/H}$, it satisfies by construction the coupling condition on Σ, and consequently $U_{|\hat{\Omega}}$ is a solution to $\hat{\mathcal{P}}_T^{E/H}$ (for every pair $(\lambda,\nu) \in \mathbb{C}^2$).

(ii) Conversely, let $\hat{U}$ be a solution to $\hat{\mathcal{P}}_T^{E/H}$. If $\hat{U}$ matches analytically the function $U_I + \mathcal{I}_t[F;\hat{U}]$ defined in $\check{\Omega}$, then the field U given by (4.11) is clearly a solution to $\mathcal{P}_T^{E/H}$. To prove this analytical matching, we show below that

$$\hat{U} = U_I + \mathcal{I}_t[F;\hat{U}] \quad \text{in } \hat{\Omega}_o.$$

The integral representation of $\hat{U}$ in $\hat{\Omega}_o$ reads (see Proposition A.5)

$$\hat{U} = \mathcal{I}_t[F;\hat{U}] + V \quad \text{in } \hat{\Omega}_o, \quad \text{where } V = \mathcal{I}_t[\Sigma;\hat{U}];$$

we thus have to verify that $V = U_I$ in $\hat{\Omega}_o$. The expression $\mathcal{I}_t[\Sigma;\hat{U}]$ (which is nothing but $\mathcal{I}_t^R[\Sigma;\hat{U}]$ by (4.10)) actually defines a field in the whole domain Ω' located inside Σ (see Fig. 4.2). This field is obviously a solution to

$$\begin{array}{ll} \operatorname{curl}(\operatorname{curl} V) - t^{-1}\operatorname{grad}(\operatorname{div} V) - k_s^2 V = 0 & \text{in } \Omega', \\ T_\lambda V = T_\lambda U_I \quad \text{and} \quad t^{-1}N_\nu V = t^{-1}N_\nu U_I & \text{on } \Sigma. \end{array}$$

To see that U_I is the only solution to this problem, note that its solution for $U_I = 0$ satisfies

$$\int_{\Omega'} \|\operatorname{curl} V\|^2 + t^{-1}\int_{\Omega'} |\operatorname{div} V|^2 - k_s^2 \int_{\Omega'} \|V\|^2 \\ + \lambda \int_\Sigma \|V \wedge n\|^2 d\gamma + t^{-1}\nu \int_\Sigma |V \cdot n|^2 d\gamma = 0,$$

from which we deduce

$$\Im m\, k_s^{-2} \int_{\Omega'} \|\operatorname{curl} V\|^2 + \Im m\, k_p^{-2} \int_{\Omega'} |\operatorname{div} V|^2 \\ + \Im m(\lambda k_s^{-2}) \int_\Sigma \|V \wedge n\|^2 d\gamma + \Im m(\nu k_p^{-2}) \int_\Sigma |V \cdot n|^2 d\gamma = 0,$$

where the terms that involve k_p have to be removed if $t = \infty$. We know that $\mathfrak{Im}\, k_s^{-2} \le 0$ and $\mathfrak{Im}\, k_p^{-2} \le 0$ (see §2). Consequently, if λ and ν are chosen such that $\mathfrak{Im}(\lambda k_s^{-2}) < 0$ and $\mathfrak{Im}(\nu k_p^{-2}) < 0$, we have $V \wedge n = 0$ and $t^{-1} V \cdot n = 0$ on Σ, and thus $\operatorname{curl} V \wedge n = 0$ and $t^{-1} \operatorname{div} V = 0$ on Σ (for $T_\lambda V = 0$ and $t^{-1} N_\nu V = 0$). It follows that V vanishes in Ω' (since $V = \mathcal{I}_t[\Sigma; V]$), which completes the proof. □

4.3. Fredholm alternative: An existence result. Our aim now is to show that problem $\hat{\mathcal{P}}_\tau^{E/H}$ can be written as a Fredholm equation, from which we will be able to deduce that it is well posed. Let us first write a variational formulation of $\hat{\mathcal{P}}_\tau^{E/H}$. By Definition 2.5, it may be easily seen that this problem is equivalent to

$$(4.12) \qquad \begin{array}{c} \text{Find } \hat{U} \in \mathcal{H}_\tau^{E/H} \text{ such that} \\ a_\tau(\hat{U}, V) = l_\tau(V) \quad \forall V \in \mathcal{H}_\tau^{E/H}, \end{array}$$

where $a_\tau(\cdot\,,\cdot)$ is the sesquilinear form defined on $\mathcal{H}_\tau^{E/H} \times \mathcal{H}_\tau^{E/H}$ by

$$\begin{aligned} a_\tau(\hat{U}, V) = & \int_{\hat{\Omega}} \zeta^{-1} \operatorname{curl} \hat{U} \cdot \overline{\operatorname{curl} V} + \int_{\hat{\Omega}} \tau^{-1} \operatorname{div} \xi \hat{U}\, \overline{\operatorname{div} \xi V} \\ & - \omega^2 \int_{\hat{\Omega}} \xi \hat{U} \cdot \overline{V} + \zeta_0^{-1} \int_\Sigma \left\{ \lambda (\hat{U} \wedge n) \cdot \overline{(V \wedge n)} + t^{-1} \nu (\hat{U} \cdot n)\, \overline{(V \cdot n)} \right\} d\gamma \\ & - \zeta_0^{-1} \int_\Sigma \left\{ T_\lambda \mathcal{I}_t^R[F; \hat{U}] \cdot \overline{V} + t^{-1} N_\nu \mathcal{I}_t^R[F; \hat{U}]\, \overline{V \cdot n} \right\} d\gamma, \end{aligned}$$

and $l_\tau(\cdot)$ is the semilinear form given by

$$l_\tau(V) = \zeta_0^{-1} \int_\Sigma \left\{ T_\lambda U_I \cdot \overline{V} + t^{-1} N_\nu U_I\, \overline{V \cdot n} \right\} d\gamma.$$

Consider then the operators $\mathbb{J}_\tau$ and $\mathbb{K}_\tau$ defined on $\mathcal{H}_\tau^{E/H}$ by

$$(4.13) \qquad \begin{aligned} (\mathbb{J}_\tau \hat{U}, V)_{\mathcal{H}_\tau^{E/H}} = & \int_{\hat{\Omega}} \zeta^{-1} \operatorname{curl} \hat{U} \cdot \overline{\operatorname{curl} V} + \int_{\hat{\Omega}} \tau^{-1} \operatorname{div} \xi \hat{U}\, \overline{\operatorname{div} \xi V} \\ & + \omega^2 \xi_0 \int_{\hat{\Omega}} \hat{U} \cdot \overline{V} + \zeta_0^{-1} \int_\Sigma \left\{ \lambda (\hat{U} \wedge n) \cdot \overline{(V \wedge n)} + t^{-1} \nu (\hat{U} \cdot n)\, \overline{(V \cdot n)} \right\} d\gamma \end{aligned}$$

and

$$(4.14) \qquad \begin{aligned} (\mathbb{K}_\tau \hat{U}, V)_{\mathcal{H}_\tau^{E/H}} = & -\omega^2 \int_{\hat{\Omega}} (\xi + \xi_0)\, \hat{U} \cdot \overline{V} \\ & - \zeta_0^{-1} \int_\Sigma \left\{ T_\lambda \mathcal{I}_t^R[F; \hat{U}] \cdot \overline{V} + t^{-1} N_\nu \mathcal{I}_t^R[F; \hat{U}]\, \overline{V \cdot n} \right\} d\gamma, \end{aligned}$$

and let L_τ be the vector of $\mathcal{H}_\tau^{E/H}$ associated with the semilinear form $l_\tau(\cdot)$ by the relation

$$(4.15) \qquad (L_\tau, V)_{\mathcal{H}_\tau^{E/H}} = l_\tau(V) \quad \forall V \in \mathcal{H}_\tau^{E/H}.$$

In these definitions, $(\cdot\,,\cdot)_{\mathcal{H}_\tau^{E/H}}$ denotes the natural scalar product in $\mathcal{H}_\tau^{E/H}$ as

$$(4.16) \qquad \begin{aligned} (U, V)_{\mathcal{H}_\tau^{E/H}} = & \int_{\hat{\Omega}} \{ \operatorname{curl} U \cdot \overline{\operatorname{curl} V} + U \cdot \overline{V} \} + |t^{-1}| \int_{\hat{\Omega}} \operatorname{div} \xi U\, \overline{\operatorname{div} \xi V} \\ & + \int_\Sigma (\hat{U} \wedge n) \cdot \overline{(\hat{V} \wedge n)}\, d\gamma + |t^{-1}| \int_\Sigma (\hat{U} \cdot n)\, \overline{(\hat{V} \cdot n)}\, d\gamma. \end{aligned}$$

The variational formulation (4.12) clearly amounts to the equation

$$(\mathbb{J}_\tau + \mathbb{K}_\tau)\hat{U} = L_\tau \quad \text{in } \mathcal{H}_\tau^{E/H}. \tag{4.17}$$

We are now able to state the main result of this section.

THEOREM 4.2. *Suppose that the coefficients ξ, ζ, and τ satisfy the conditions given in* §3.1. *Let $(\lambda,\nu) \in \mathbb{C}^2$ be chosen such that $\Im m(\lambda k_s^{-2}) < 0$ and $\Im m(\nu k_p^{-2}) < 0$ (if $t \neq \infty$). The reduced problem $\hat{\mathcal{P}}_\tau^{E/H}$ has a unique solution in $\mathcal{H}_\tau^{E/H}$.*

Proof. We prove in §4.4 that $\mathbb{J}_\tau$ and $\mathbb{K}_\tau$ are, respectively, an automorphism and a compact operator in $\mathcal{H}_\tau^{E/H}$. The Fredholm alternative shows that if the only solution to (4.17) with $L_\tau = 0$ is the trivial solution $\hat{U} = 0$, then (4.17) has exactly one solution for every $L_\tau \in \mathcal{H}_\tau^{E/H}$. And the required uniqueness property simply follows from the uniqueness of the solution to problem $\mathcal{P}_\tau^{E/H}$ (Theorem 3.1) and the equivalence between this latter problem and $\hat{\mathcal{P}}_\tau^{E/H}$ for this uniqueness property (Proposition 4.1). □

Remark 4.3. This theorem together with the equivalence between $\mathcal{P}_\tau^{E/H}$ and $\hat{\mathcal{P}}_\tau^{E/H}$ for the existence property (Proposition 4.1) complete the proof of (i) in Theorem 2.12. Problem $\mathcal{P}_\tau^{E/H}$ is well-posed, which implies that its solution is independent of τ (since $\operatorname{div} \xi U = 0$, see §2.4). This shows in particular that the same holds for the solution to $\hat{\mathcal{P}}_\tau^{E/H}$.

4.4. Two technical lemmas. We prove below that operators $\mathbb{J}_\tau$ and $\mathbb{K}_\tau$ given by (4.13) and (4.14) define, respectively, an automorphism and a compact operator in $\mathcal{H}_\tau^{E/H}$.

LEMMA 4.4. *Let $(\lambda,\nu) \in (\mathbb{C}^*)^2$ be chosen such that $\Im m(\lambda k_s^{-2}) < 0$ and $\Im m(\nu k_p^{-2}) < 0$ (if $t \neq \infty$). Then $\mathbb{J}_\tau$ is a bounded invertible operator in $\mathcal{H}_\tau^{E/H}$ with bounded inverse.*

Proof. By virtue of the Lax–Milgram theorem, it is enough to prove that the sesquilinear form associated with $\mathbb{J}_\tau$ in (4.13) is coercive in $\mathcal{H}_\tau^{E/H}$, i.e.,

$$\left|(\mathbb{J}_\tau V, V)_{\mathcal{H}_\tau^{E/H}}\right| \geq \alpha \|V\|^2_{\mathcal{H}_\tau^{E/H}} \quad \forall V \in \mathcal{H}_\tau^{E/H}$$

for some positive constant α. Intuitively, this results from the fact that the functions ζ^{-1}, τ^{-1}, $\omega^2\xi_0$, $\lambda\zeta_0^{-1}$, and $t^{-1}\nu\zeta_0^{-1}$ involved in the different integrals in (4.13) take their values in a sector of the complex plane (with vertex 0) whose opening angle is less than π. More precisely, by dividing these functions by $\omega^2\xi_0$, we infer that there exists $\theta \in]-\pi, 0[$ such that

$$\begin{gathered} \theta \leq \arg(\omega^2\xi_0\zeta)^{-1} \leq 0 \quad \text{a.e.} \quad \text{in } \hat{\Omega}, \\ \theta \leq \arg(\omega^2\xi_0\tau)^{-1} \leq 0 \quad \text{a.e.} \quad \text{in } \hat{\Omega}, \\ \theta \leq \arg(\lambda k_s^{-2}) \leq 0 \quad \text{and} \quad \theta \leq \arg(\nu k_p^{-2}) \leq 0 \quad (\text{if } t \neq \infty). \end{gathered}$$

The two former statements follow from hypotheses (2.3) and (2.13), and the latter from the choice of the parameters λ and ν. By noticing moreover that functions ζ^{-1} and τ^{-1} are bounded from below, we deduce that there exists a positive constant C such that

$$\begin{gathered} \Re e\left\{(\omega^2\xi_0\zeta)^{-1} e^{-i\theta/2}\right\} \geq C \quad \text{a.e.} \quad \text{in } \hat{\Omega}, \\ \Re e\left\{(\omega^2\xi_0\tau)^{-1} e^{-i\theta/2}\right\} \geq C \quad \text{a.e.} \quad \text{in } \hat{\Omega}, \\ \Re e\left\{\lambda k_s^{-2}\, e^{-i\theta/2}\right\} \geq C \quad \text{and} \quad \Re e\left\{\nu k_p^{-2}\, e^{-i\theta/2}\right\} \geq C \quad (\text{if } t \neq \infty). \end{gathered}$$

Hence we have

$$\Re e\left\{(\omega^2\xi_0)^{-1}\, e^{-i\theta/2}\, (\mathbb{J}_\tau V, V)_{\mathcal{H}_\tau^{E/H}}\right\} \ge C\int_{\hat\Omega} \|\operatorname{curl} V\|^2 + C\left|t^{-1}\right| \int_{\hat\Omega} |\operatorname{div} \xi V|^2$$
$$+\cos\theta/2 \int_{\hat\Omega} \|V\|^2 + C\int_\Sigma \|V\wedge n\|^2 d\gamma + C\left|t^{-1}\right|\int_\Sigma |V\cdot n|^2 d\gamma$$

and consequently

$$\Re e\left\{(\omega^2\xi_0)^{-1}\, e^{-i\theta/2} (\mathbb{J}_\tau V, V)_{\mathcal{H}_\tau^{E/H}}\right\} \ge \min(C, \cos\theta/2)\|V\|^2_{\mathcal{H}_\tau^{E/H}}$$

for every $V \in \mathcal{H}_\tau^{E/H}$. The coerciveness of the form follows. □

LEMMA 4.5. *$\mathbb{K}_\tau$ is a compact operator in $\mathcal{H}_\tau^{E/H}$.*

Proof. We study separately the two terms that define operator $\mathbb{K}_\tau$ in (4.14) by writing it as $\mathbb{K}_\tau = -\omega^2\mathbb{K}^{\hat\Omega} - \zeta_0^{-1}\mathbb{K}_\tau^\Sigma$, where

$$(\mathbb{K}^{\hat\Omega} U, V)_{\mathcal{H}_\tau^{E/H}} = \int_{\hat\Omega} (\xi + \xi_0)\; U\cdot \overline{V},$$
$$(\mathbb{K}_\tau^{\Sigma} U, V)_{\mathcal{H}_\tau^{E/H}} = \int_\Sigma \left\{T_\lambda \mathcal{I}_t^R[F;U]\cdot\overline{V} + t^{-1} N_\nu \mathcal{I}_t^R[F;U]\;\overline{V\cdot n}\right\} d\gamma.$$

(i) To see that $\mathbb{K}^{\hat\Omega}$ is compact, first notice from the Schwarz inequality that

$$\left|(\mathbb{K}^{\hat\Omega} U, V)_{\mathcal{H}_\tau^{E/H}}\right| \le C\|U\|_{L^2(\hat\Omega)}\|V\|_{L^2(\hat\Omega)}$$

since $\xi \in L^\infty(\Omega)$. As a consequence, we have

$$\left\|\mathbb{K}^{\hat\Omega} U\right\|_{\mathcal{H}_\tau^{E/H}} = \sup_{V\in\mathcal{H}_\tau^{E/H},\ V\neq 0} \frac{\left|(\mathbb{K}^{\hat\Omega} U, V)_{\mathcal{H}_\tau^{E/H}}\right|}{\|V\|_{\mathcal{H}_\tau^{E/H}}} \le C\|U\|_{L^2(\hat\Omega)}$$

for every $U \in \mathcal{H}_\tau^{E/H}$. Thus $\mathbb{K}^{\hat\Omega}$ appears as a continuous operator from $L^2(\hat\Omega)$ into $\mathcal{H}_\tau^{E/H}$. The conclusion follows from the compactness of the canonical injection from $\mathcal{H}_\tau^{E/H}$ into $L^2(\hat\Omega)$ (see Corollary B.5).

(ii) Let us prove now that $\mathbb{K}_\tau^\Sigma$ is compact. From the definition (4.9) of $\mathcal{I}_t^R[F;U]$, it is easy to see that this function is infinitely differentiable in a vicinity of Σ, and if D denotes any derivative operator on Σ (of any order), we have

$$\left|D\mathcal{I}_t^R[F;U](x)\right| \le C_D\Big(\|\operatorname{curl} U\|_{L^2(\hat\Omega_o)} + \left|t^{-1}\right| \|\operatorname{div} U\|_{L^2(\hat\Omega_o)}$$
$$+\ \|U\|_{L^2(\hat\Omega_o)} + \|U\wedge n\|_{H^{-1/2}(F)} + \left|t^{-1}\right| \|U\cdot n\|_{H^{-1/2}(F)}\Big)$$

at every point $x \in \Sigma$ (C_D is a positive constant which depends on D). We thus have

$$\left|D\mathcal{I}_t^R[F;U](x)\right| \le C_D\|U\|_{\mathcal{H}_\tau^{E/H}} \quad \forall x\in\Sigma.$$

This implies in particular that for every $s\ge 0$,

$$\left\|T_\lambda \mathcal{I}_t^R[F;U]\right\|_{H^s(\Sigma)} \le C_s\|U\|_{\mathcal{H}_\tau^{E/H}},$$

as well as the same inequality for $N_\nu \mathcal{I}_t^R[F;U]$. We deduce from these properties that

$$\left|(\mathbb{K}_\tau^\Sigma U, V)_{\mathcal{H}_\tau^{E/H}}\right| \le C_s\ \|U\|_{\mathcal{H}_\tau^{E/H}}\ \left\|n\wedge(V\wedge n) + t^{-1} n(V\cdot n)\right\|_{H^{-s}(\Sigma)^3}\cdot$$

Hence the adjoint $(\mathbb{K}_\tau^\Sigma)^*$ of operator $\mathbb{K}_\tau^\Sigma$ satisfies

$$\begin{aligned}\left\|(\mathbb{K}_\tau^\Sigma)^* V\right\|_{\mathcal{H}_\tau^{E/H}} &= \sup_{U\in\mathcal{H}_\tau^{E/H},\ U\neq 0} \frac{\left|(\mathbb{K}_\tau^\Sigma U, V)_{\mathcal{H}_\tau^{E/H}}\right|}{\|U\|_{\mathcal{H}_\tau^{E/H}}} \\ &\leq C_s \ \left\|n\wedge(V\wedge n)+t^{-1}n(V\cdot n)\right\|_{H^{-s}(\Sigma)^3}\end{aligned}$$

for every $s\geq 0$. As a consequence, $(\mathbb{K}_\tau^\Sigma)^*$ appears as a continuous operator from $H^{-s}(\Sigma)^3$ into $\mathcal{H}_\tau^{E/H}$. The trace operator $V\to\big(n\wedge(V\wedge n)+t^{-1}n(V\cdot n)\big)_{|\Sigma}$ is obviously continuous from $\mathcal{H}_\tau^{E/H}$ into $L^2(\Sigma)^3$; the compactness of $(\mathbb{K}_\tau^\Sigma)^*$ (and consequently the compactness of $\mathbb{K}_\tau^\Sigma$) then results from the compactness of the canonical injection from $L^2(\Sigma)$ into $H^{-s}(\Sigma)$ (with $s>0$). □

5. Conclusion. Let us first notice that from a theoretical point of view, the method of "coupling between variational formulation and integral representation" described above provides a direct proof for the existence of a solution to the classical scattering problem. Indeed, the separation between the boundary F (on which the integral representation is written) and the boundary Σ (where this integral representation is used) leads to the compactness of the corresponding operator and, consequently, allows us to write the problem as a Fredholm equation. This differs from usual integral techniques (see, e.g., [5], [13]) that deduce the existence result for the classical problem from the existence of a solution to the regularized problem by studying the equations satisfied by the divergence of the field (this requires a sufficient regularity of the electromagnetic coefficients as well as of the boundary of the obstacle). Let us mention two other methods for proving the existence of a solution, which both consist of introducing a sequence of well-posed scattering problems, and studying the limit of their solutions. First, the so-called "limiting absorption principle" applies in the nondissipative case, which actually is considered as the limit of dissipative problems. (Bendali [4] presents this method for the vector Helmholtz equation outside a regular boundary, but it is probably easy to extend it to the case of Maxwell's equations and for irregular boundaries.) Second, the original proof of Abboud [1] is based on a sequence of regular boundaries, which tends to an irregular one.

To a certain extent, our method of regularization of Maxwell equations can be compared with the work of Abboud and Nédélec [2], who obtained an H^1-formulation of the problem (using a series expansion of the field by means of spherical harmonics instead of an integral representation). Their "regularized problem" (which also comes within the context of the Fredholm alternative) is, however, different; in particular, its variational formulation involves a boundary sesquilinear form on every surface of discontinuity of the electromagnetic coefficients (this form depends on the curvature tensor of the surface, which requires regular enough boundaries).

From a numerical point of view, let us recall the two main advantages of the regularized problem compared with the classical one.

(i) On one hand, the integral representations involve a less singular kernel (order 1 instead of 3); our approach agrees with the integral equation technique developed in the case of a regular boundary by Bendali [4]. (The singularity of the kernel is also of order 1 in its integral equation.)

(ii) On the other hand, the approximation of the function spaces appearing in the regularized formulation is a matter for a standard discretization by Lagrange finite elements. Note, however, that one must be careful if the coefficients of the medium or the boundary of the obstacle are irregular. Indeed the spaces to be discretized contain transmission conditions on the surfaces of discontinuity of the coefficients (following from the condition $\operatorname{div}\xi U \in L^2$), which leads us to "split" the degrees of freedom located on these surfaces. Moreover, the solution to the scattering problem may be singular in the vicinity of the singularities of the boundary (for instance, the corners and edges for a piecewise regular boundary). More precisely, this solution can be decomposed (see Costabel [7]) as the sum of a "regular" part (which belongs to H^1 and can thus be approximated by Lagrange finite elements) and a "singular" part (which is orthogonal to H^1 and must be taken into account explicitly).

The numerical implementation of our method is in progress and will be presented in the near future.

Appendix A. Integral representation. We have pointed out in §2.2 the analogy between the regularized Maxwell equation and linear elasticity. Actually, part of the results given in this appendix comes within the classical framework of potential theory in elasticity (see, e.g., Kupradze [14]). However, for the sake of consistency, we present below a complete and concise method of working out the integral representation formulas that are used in the present paper. In §A.1, we exhibit the expression of the Green tensors associated with the classical and regularized equations. Then, in §A.2, we prove an integral representation formula in a bounded domain, using elementary properties of distributions. We finally generalize this formula in §A.3 to the case of an exterior domain.

A.1. The Green tensors. A 3×3 matrix of distributions $\mathbb{G}_t$ is said to be a Green tensor (or a tensor of fundamental solutions) of the classical ($t = \infty$) or the regularized ($0 < t < \infty$) Maxwell equation if

$$\text{(A.1)} \qquad \operatorname{curl}(\operatorname{curl}\mathbb{G}_t) - t^{-1}\operatorname{grad}(\operatorname{div}\mathbb{G}_t) - k_s^2\mathbb{G}_t = \delta\mathbb{I} \quad \text{in } \mathbb{R}^3,$$

where δ denotes the Dirac measure at point $x = 0$ and $\mathbb{I}$ is the identity matrix. In other words, if we denote by $\mathbb{G}_t^{(i)}$ for $i = 1, \dots, 3$ the column vectors of $\mathbb{G}_t$, this amounts to the relations (to be understood in the sense of distributions)

$$\text{(A.2)} \qquad \operatorname{curl}(\operatorname{curl}\mathbb{G}_t^{(i)}) - t^{-1}\operatorname{grad}(\operatorname{div}\mathbb{G}_t^{(i)}) - k_s^2\mathbb{G}_t^{(i)} = \delta x^{(i)}, \quad i = 1, \dots, 3,$$

where $(x^{(1)}, x^{(2)}, x^{(3)})$ is the canonical basis of $\mathbb{R}^3$. Such a Green tensor $\mathbb{G}_t$ is said to be outgoing if each column vector $\mathbb{G}_t^{(i)}$ satisfies the outgoing transverse and radial radiation conditions:

$$\text{(A.3)} \qquad \begin{aligned} &\lim_{R\to\infty} \int_{\Sigma_R} \left\| \operatorname{curl}\mathbb{G}_t^{(i)} \wedge n - ik_s n \wedge (\mathbb{G}_t^{(i)} \wedge n) \right\|^2 d\gamma = 0 && \text{if } \Im m\, k_s = 0, \\ &\lim_{R\to\infty} \int_{\Sigma_R} \left[\left\| \operatorname{curl}\mathbb{G}_t^{(i)} \wedge n \right\|^2 + |k_s|^2 \left\| n \wedge (\mathbb{G}_t^{(i)} \wedge n) \right\|^2 \right] d\gamma = 0 && \text{if } \Im m\, k_s > 0, \end{aligned}$$

$$\text{(A.4)} \qquad \begin{aligned} &\lim_{R\to\infty} \int_{\Sigma_R} \left| \sqrt{t^{-1}} \operatorname{div}\mathbb{G}_t^{(i)} - ik_s\, \mathbb{G}_t^{(i)} \cdot n \right|^2 d\gamma = 0 && \text{if } \Im m\, k_p = 0, \\ &\lim_{R\to\infty} \int_{\Sigma_R} \left[\left|t^{-1}\right| \left| \operatorname{div}\mathbb{G}_t^{(i)} \right|^2 + |k_s|^2 \left| \mathbb{G}_t^{(i)} \cdot n \right| \right] d\gamma = 0 && \text{if } \Im m\, k_p > 0, \end{aligned}$$

where we recall that $k_p = k_s\sqrt{t}$ if $t \neq \infty$.

PROPOSITION A.1. *The outgoing Green tensor $\mathbb{G}_t$ associated with the classical* (*respectively, regularized*) *Maxwell equation is uniquely defined by*

$$\text{(A.5)} \qquad \begin{aligned} \mathbb{G}_\infty &= g_{k_s}\mathbb{I} + k_s^{-2}\operatorname{Hess} g_{k_s}, \\ \mathbb{G}_t &= g_{k_s}\mathbb{I} + k_s^{-2}\operatorname{Hess}(g_{k_s} - g_{k_p}) \quad \textit{if } t \neq \infty, \end{aligned}$$

where $\mathbb{I}$ *is the identity matrix,* Hess *stands for the Hessian operator, and* g_k *denotes the function defined on* $\mathbb{R}^3$ *by*

$$g_k(x) = \frac{e^{ik\|x\|}}{4\pi\|x\|}. \tag{A.6}$$

Remark A.2. For every k such that $\Re e\, k > 0$ and $\Im m\, k \geq 0$, function g_k is nothing but the outgoing Green function of operator $-\triangle - k^2$ (see, e.g., Wilcox [26]), i.e., the only function that satisfies

$$-\triangle g_k - k^2 g_k = \delta \;\text{ in } \mathbb{R}^3, \tag{A.7}$$

and the well-known Sommerfeld outgoing radiation condition

$$\begin{aligned} &\lim_{R\to\infty}\int_{\Sigma_R} |\partial_n g_k - ikg_k|^2\, d\gamma = 0 && \text{if } \Im m\, k = 0,\\ &\lim_{R\to\infty}\int_{\Sigma_R} \left[|\partial_n g_k|^2 + |k|^2\, \|g_k\|^2\right] d\gamma = 0 && \text{if } \Im m\, k > 0. \end{aligned} \tag{A.8}$$

Remark A.3. By (A.7) and noticing that $\operatorname{Hess} g_k = \operatorname{grad}\operatorname{div}(g_k\mathbb{I})$, we see that (A.5) can be equivalently expressed as

$$\begin{aligned} \mathbb{G}_\infty &= k_s^{-2}\left(\operatorname{curl}\operatorname{curl}(g_{k_s}\mathbb{I}) - \delta\mathbb{I}\right),\\ \mathbb{G}_t &= k_s^{-2}\left(\operatorname{curl}\operatorname{curl}(g_{k_s}\mathbb{I}) - \operatorname{grad}\operatorname{div}(g_{k_p}\mathbb{I}) - \delta\mathbb{I}\right) \quad \text{if } t \neq \infty\\ &= k_s^{-2}\operatorname{curl}\operatorname{curl}((g_{k_s} - g_{k_p})\mathbb{I}) + tg_{k_p}\mathbb{I}), \end{aligned} \tag{A.9}$$

where the two first formulas may be seen as the decompositions of $\mathbb{G}_\infty$ and $\mathbb{G}_t$ into "S" and "P" waves.

Remark A.4. The singularity of $\mathbb{G}_\infty$ in the vicinity of $x = 0$ is of order 3 (i.e., it has an asymptotic behavior such as $\|x\|^{-3}$). The singularity of $\mathbb{G}_t$ is only of order 1; this is obvious in the particular case $t = 1$ since $\mathbb{G}_1$ is simply the diagonal operator $g_{k_s}\mathbb{I}$).

Proof. We first show that (A.1) has at most one outgoing solution and then that the tensor $\mathbb{G}_t$ given by (A.5) is this solution.

(i) The uniqueness of $\mathbb{G}_t$ follows from the same arguments as those developed in §3.3 for the proof of Theorem 3.1. Indeed, if U is a solution to the homogeneous equation

$$\operatorname{curl}(\operatorname{curl} U) - t^{-1}\operatorname{grad}(\operatorname{div} U) - k_s^2 U = 0$$

in the whole space $\mathbb{R}^3$ and satisfies the radiation conditions (A.3) and (A.4), it may be readily verified that the energy flux at infinity vanishes, which implies $U = 0$. Note that the uniqueness of $\mathbb{G}_\infty$ results from that of $\mathbb{G}_t$ ($t \neq \infty$).

(ii) It may be easily seen that

$$\begin{aligned} \operatorname{curl}\mathbb{G}_\infty = \operatorname{curl}(g_{k_s}\mathbb{I}) \quad &(\text{and} \quad \operatorname{div}\mathbb{G}_\infty = -k_s^{-2}\operatorname{div}(\delta\mathbb{I})),\\ \operatorname{curl}\mathbb{G}_t = \operatorname{curl}(g_{k_s}\mathbb{I}) \quad &\text{and} \quad \operatorname{div}\mathbb{G}_t = t\operatorname{div}(g_{k_p}\mathbb{I}). \end{aligned} \tag{A.10}$$

These relations together with the expressions (A.9) show that $\mathbb{G}_\infty$ and $\mathbb{G}_t$ are solution to (A.1).

(iii) It remains to verify that the column vectors of $\mathbb{G}_\infty$ and $\mathbb{G}_t$ satisfy the radiation conditions. Let $\Sigma_R = \left\{x \in \mathbb{R}^3 \mid \|x\| = R\right\}$, and let n denote the outer unit normal on Σ_R. From (A.10), we deduce that on Σ_R we have

$$\begin{aligned} \operatorname{curl} \mathbb{G}_\infty^{(i)} \wedge n &= d_r g_{k_s}\; n \wedge (x^{(i)} \wedge n) \quad (\text{and } \operatorname{div} \mathbb{G}_\infty^{(i)} = 0), \\ \operatorname{curl} \mathbb{G}_t^{(i)} \wedge n &= d_r g_{k_s}\; n \wedge (x^{(i)} \wedge n) \quad \text{and} \quad \operatorname{div} \mathbb{G}_t^{(i)} = t\, d_r g_{k_p}\; x^{(i)} \cdot n. \end{aligned} \tag{A.11}$$

Moreover, by noticing that for large R

$$d_r g_k(R) = O(R^{-1}) \quad \text{and} \quad d_r^2 g_k(R) = -k^2 g_k(R) + O(R^{-2}),$$

we see that

$$\operatorname{grad} \operatorname{div}(g_k x^{(i)}) = -k^2 g_k\; (x^{(i)} \cdot n) n + O(R^{-2}).$$

Hence, we infer from (A.5) that

$$\begin{aligned} \mathbb{G}_\infty^{(i)} &= g_{k_s}\; n \wedge (x^{(i)} \wedge n) + O(R^{-2}), \\ \mathbb{G}_t^{(i)} &= g_{k_s}\; n \wedge (x^{(i)} \wedge n) + t\, g_{k_p}\; (x^{(i)} \cdot n) n + O(R^{-2}). \end{aligned} \tag{A.12}$$

The conclusion follows from these formulas together with (A.11) and the fact that g_k satisfies the radiation condition (A.8). For instance, if $k_s \in \mathbb{R}^+$, we have

$$\operatorname{curl} \mathbb{G}_t^{(i)} \wedge n - i k_s\; n \wedge (\mathbb{G}_t^{(i)} \wedge n) = (d_r g_{k_s} - i k_s g_{k_s})\; n \wedge (x^{(i)} \wedge n) + O(R^{-2})$$

for $t = \infty$ or $t \neq \infty$, which shows that each $\mathbb{G}_t^{(i)}$ satisfies the transverse radiation condition (A.3). □

A.2. Integral representation in a bounded domain. Let $\mathcal{O}$ be a bounded open set of $\mathbb{R}^3$ with Lipschitz-continuous boundary F. We denote by n the unit outward normal on F and by $\mathcal{O}'$ the exterior of $\mathcal{O}$, i.e., $\mathcal{O}' = \mathbb{R}^3 \setminus \overline{\mathcal{O}}$. Let $U \in H(\text{curl}; \mathcal{O}) \cap H(\text{div}; \mathcal{O})$ be a function that satisfies (in the sense of distributions) the classical ($t = \infty$) or regularized ($t \neq \infty$) Maxwell equation

$$\operatorname{curl}(\operatorname{curl} U) - t^{-1} \operatorname{grad}(\operatorname{div} U) - k_s^2 U = 0 \tag{A.13}$$

in $\mathcal{O}$. Consider the function $\mathcal{U} \in L^2(\mathbb{R}^3)$ defined by $\mathcal{U} = U$ in $\mathcal{O}$ and $\mathcal{U} = 0$ in $\mathcal{O}'$.

PROPOSITION A.5. *In $\mathcal{O} \cup \mathcal{O}'$, the function $\mathcal{U}$ satisfies the integral representation*

$$\begin{aligned} \mathcal{U}(x) = &\int_F \mathbb{G}_t(x-y) \left\{\operatorname{curl} U(y) \wedge n_y + t^{-1} n_y \operatorname{div} U(y)\right\} d\gamma_y \\ &- \int_F (\operatorname{curl}_y \mathbb{G}_t(x-y)) \left\{U(y) \wedge n_y\right\} d\gamma_y \\ &- t^{-1} \int_F (\operatorname{div}_y \mathbb{G}_t(x-y))^T \left\{U(y) \cdot n_y\right\} d\gamma_y \end{aligned} \tag{A.14}$$

(*where the exponent T denotes the transposed of a matrix*).

Note that by virtue of (A.10), this integral representation can be expressed in the form

$$\begin{aligned} \mathcal{U}(x) = &\int_F \mathbb{G}_t(x-y) \left\{\operatorname{curl} U(y) \wedge n_y + t^{-1} n_y \operatorname{div} U(y)\right\} d\gamma_y \\ &- \int_F \operatorname{grad}_y g_{k_s}(x-y) \wedge \left\{U(y) \wedge n_y\right\} d\gamma_y \\ &- \int_F \operatorname{grad}_y g_{k_p}(x-y) \left\{U(y) \cdot n_y\right\} d\gamma_y, \end{aligned} \tag{A.15}$$

where the last integral has to be removed in the case $t = \infty$.

Remark A.6. In these formulas, all the integrals should be written as a duality product between $H^{-1/2}(F)$ and $H^{1/2}(F)$. Indeed, we know from the hypothesis $U \in H(\mathrm{curl}; \mathcal{O}) \cap H(\mathrm{div}; \mathcal{O})$ that the quantities $(U \wedge n)_{|F}$ and $(U \cdot n)_{|F}$ are defined as elements of $H^{-1/2}(F)$ (see, e.g., [8]). Moreover, with the additional condition $\mathrm{curl}(\mathrm{curl}\, U) - t^{-1}\,\mathrm{grad}(\mathrm{div}\, U) \in L^2(\mathcal{O})$ that follows from (A.13), the quantity $(\mathrm{curl}\, U \wedge n + t^{-1} n\, \mathrm{div}\, U)_{|F}$ also appears as a distribution of $H^{-1/2}(\Sigma)$. However, for the sake of simplicity, we retain the integral notation.

Proof. We show below that the integral representation (A.14) is nothing but the convolution between the Green tensor $\mathbb{G}_t$ and the distribution

$$S = \mathrm{curl}(\mathrm{curl}\mathcal{U}) - t^{-1}\,\mathrm{grad}(\mathrm{div}\mathcal{U}) - k_s^2 \mathcal{U} \tag{A.16}$$

whose restriction to $\mathcal{O}$ and $\mathcal{O}'$ is obviously 0 by construction of $\mathcal{U}$; the support of S is thus contained in F.

(i) Let us first prove that

$$\mathcal{U} = \mathbb{G}_t * S, \tag{A.17}$$

where the convolution is defined similarly to the corresponding matrix product, i.e.,

$$(\mathbb{G}_t * S)^{(i)} = \sum_{j=1}^{3} \mathbb{G}_t^{(i,j)} * S^{(j)}.$$

With this definition, it may be easily seen that the operators curl, grad, and div can be expressed as

$$\mathrm{curl}\, V = (\mathrm{curl}\, \delta \mathbb{I}) * V, \quad \mathrm{grad}\, v = (\mathrm{grad}\, \delta) * v, \quad \text{and} \quad \mathrm{div}\, V = (\mathrm{grad}\, \delta)^T * V,$$

and we have the transposition relation $(A * B)^T = B^T * A^T$ (note that $(\mathrm{curl}\, \delta \mathbb{I})^T = -\,\mathrm{curl}\, \delta \mathbb{I}$). We deduce

$$\begin{aligned} &\mathbb{G}_t^T * \left\{\mathrm{curl}(\mathrm{curl}\mathcal{U}) - t^{-1}\,\mathrm{grad}(\mathrm{div}\mathcal{U}) - k_s^2 \mathcal{U}\right\} \\ &\qquad = \left\{\mathrm{curl}(\mathrm{curl}\, \mathbb{G}_t) - t^{-1}\,\mathrm{grad}(\mathrm{div}\, \mathbb{G}_t) - k_s^2 \mathbb{G}_t\right\}^T * \mathcal{U}, \end{aligned}$$

where the right-hand side is nothing but $(\delta \mathbb{I}) * \mathcal{U} = \mathcal{U}$ by virtue of (A.1). Property (A.17) follows.

(ii) The second step consists of calculating the distribution S. For every $V \in \mathcal{D}(\mathbb{R}^3)^3$, we have

$$\langle S, V \rangle = \int_{\mathcal{O}} U \cdot \left\{\mathrm{curl}(\mathrm{curl}\, V) - t^{-1}\,\mathrm{grad}(\mathrm{div}\, V) - k_s^2 V\right\}.$$

Integrating by parts and using the fact that U satisfies Maxwell's equation in $\mathcal{O}$ yields

$$\begin{aligned} \langle S, V \rangle = &\int_F \left\{\mathrm{curl}\, U \wedge n + t^{-1} n\, \mathrm{div}\, U\right\} \cdot V\, d\gamma \\ &+ \int_F \{U \wedge n\} \cdot \mathrm{curl}\, V\, d\gamma - t^{-1} \int_F \{U \cdot n\}\, \mathrm{div}\, V\, d\gamma, \end{aligned}$$

which can be expressed in the form

$$S = \left\{\mathrm{curl}\, U \wedge n + t^{-1} n\, \mathrm{div}\, U\right\} \delta_F + \mathrm{curl}\left\{(U \wedge n)\delta_F\right\} + t^{-1}\,\mathrm{grad}\left\{(U \cdot n)\delta_F\right\}, \tag{A.18}$$

where we denote by $f\delta_F$ the distribution of $\mathcal{D}'(\mathbb{R}^3)$ given by

$$\langle f\delta_F, \varphi \rangle = \int_F f \varphi\, d\gamma \quad \forall \varphi \in \mathcal{D}(\mathbb{R}^3).$$

(iii) Substituting the expression (A.18) of S in (A.17), we finally have

$$\begin{aligned}\mathcal{U} = \mathbb{G}_t * \left\{\operatorname{curl} U \wedge n + t^{-1} n \operatorname{div} U\right\} \delta_F \\ + (\operatorname{curl} \mathbb{G}_t) * \{(U \wedge n)\delta_F\} + t^{-1}(\operatorname{div} \mathbb{G}_t)^T * \{(U \cdot n)\delta_F\}.\end{aligned}$$

Each term of the right-hand side is a convolution between a function that is infinitely differentiable outside 0 and a measure with compact support (on F). Outside the support of this measure, it thus defines an infinitely differentiable function that can be expressed in an integral form (see, e.g., Schwartz [22]); this form is simply the integral representation formula (A.14). □

A.3. The case of an exterior domain. Let $\mathcal{O}$, $\mathcal{O}'$, and F be defined as in the previous section, and let n denote now the unit normal on F oriented toward $\mathcal{O}$. Consider a function $U \in H_{\text{loc}}(\text{curl}; \mathcal{O}') \cap H_{\text{loc}}(\text{div}; \mathcal{O}')$ that satisfies (in the sense of distributions) the classical or regularized Maxwell equation (A.13) in $\mathcal{O}'$. Define the function $\mathcal{U} \in L^2(\mathbb{R}^3)$ by $\mathcal{U} = 0$ in $\mathcal{O}$ and $\mathcal{U} = U$ in $\mathcal{O}'$.

PROPOSITION A.7. *If U satisfies the outgoing radiation conditions* (A.3) *and* (A.4), *then the statement of Proposition* A.5 *holds*; *formulas* (A.14) *or* (A.15) *are still valid.*

The main ingredients of the proof are the integral representation in a bounded domain (Proposition A.5), the uniqueness of the Green tensor (Proposition A.1), and the following lemma.

LEMMA A.8. *Let A, B, and c be three functions defined on the boundary F of $\mathcal{O}$* (*where A and B are vector-valued and c is scalar-valued*). *Then the function $\mathcal{U}$ given by*

$$\begin{aligned}\mathcal{U}(x) = \int_F \mathbb{G}_t(x-y)\, A(y)\, d\gamma_y \\ - \int_F (\operatorname{curl}_y \mathbb{G}_t(x-y))\, B(y)\, d\gamma_y - t^{-1} \int_F (\operatorname{div}_y \mathbb{G}_t(x-y))^T c(y)\, d\gamma_y,\end{aligned} \tag{A.19}$$

satisfies Maxwell's equation (A.13) *separately in $\mathcal{O}$ and $\mathcal{O}'$ as well as the radiation conditions* (A.3) *and* (A.4).

Proof. To see that Maxwell's equation (A.13) is satisfied, just notice that as in the proof of Proposition A.5, the definition (A.19) of $\mathcal{U}$ can be expressed equivalently as

$$\mathcal{U} = \mathbb{G}_t * (A\delta_F) + (\operatorname{curl} \mathbb{G}_t) * (B\delta_F) + t^{-1}(\operatorname{div} \mathbb{G}_t)^T * (c\delta_F),$$

from which we deduce

$$\begin{aligned}&\operatorname{curl}(\operatorname{curl} \mathcal{U}) - t^{-1} \operatorname{grad}(\operatorname{div} \mathcal{U}) - k_s^2 \mathcal{U} \\ &\quad = (\delta\mathbb{I}) * (A\delta_F) + (\operatorname{curl} \delta\mathbb{I}) * (B\delta_F) + t^{-1}(\operatorname{div} \delta\mathbb{I})^T * (c\delta_F) \\ &\quad = A\delta_F + \operatorname{curl}(B\delta_F) + t^{-1} \operatorname{grad}(c\delta_F).\end{aligned}$$

This distribution vanishes outside F.

To see that $\mathcal{U}$ satisfies the radiation conditions, we simply have to exhibit the asymptotic behavior of (A.19) when $\|x\|$ tends to infinity. The conclusion follows from the fact that the columns of $\mathbb{G}_t$ and $\operatorname{curl} \mathbb{G}_t$, as well as $(\operatorname{div} \mathbb{G}_t)^T$ if $t \neq \infty$, satisfy the radiation conditions (the proof of this result, which does not raise any particular difficulty, is left to the reader). □

Proof of Proposition A.7. Let Σ be the boundary of an open ball B whose radius is chosen large enough so that it contains F. Using the simplified notation $\mathcal{U} = \mathcal{I}_t[F; U]$

for the integral representation (A.14), we deduce from Proposition A.5 (applied in the bounded domain $\mathcal{O}' \cap B$) that

$$\mathcal{I}_t[F;U] + \mathcal{I}_t[\Sigma;U] = \begin{cases} U & \text{in } \mathcal{O}' \cap B, \\ 0 & \text{in } \mathcal{O} \cup (\mathbb{R}^3 \setminus \overline{B}). \end{cases} \tag{A.20}$$

We thus have to prove that $\mathcal{I}_t[\Sigma;U] = 0$ in $\mathcal{O}' \cap B$. First notice that from Lemma A.8, $\mathcal{I}_t[\Sigma;U]$ defines a function that satisfies Maxwell's equation separately in B and $\mathbb{R}^3 \setminus \overline{B}$ as well as the radiation conditions. Consider then the function V given by

$$V = \mathcal{I}_t[\Sigma;U] \text{ in } B, \quad \text{and} \quad V = U - \mathcal{I}_t[F;U] \text{ in } \mathbb{R}^3 \setminus \overline{B}.$$

Noticing that in $\mathcal{O}' \cap B$ we have $V = U - \mathcal{I}_t[F;U]$ by virtue of (A.20), we infer that V is regular in the vicinity of Σ. Hence V satisfies Maxwell's equation in the whole space $\mathbb{R}^3$ and the radiation conditions. The uniqueness of $\mathbb{G}_t$ (Proposition A.1) ensures that $V \equiv 0$. □

Appendix B. Compactness results. For the sake of completeness, we sum up in this appendix the compactness results that are available to our knowledge regarding the embedding of

$$H(\operatorname{curl}, \operatorname{div}\xi; \mathcal{O}) = \left\{U \in L^2(\mathcal{O})^3 \;\middle|\; \operatorname{curl} U \in L^2(\mathcal{O})^3 \text{ and } \operatorname{div} \xi U \in L^2(\mathcal{O})\right\}$$

into $L^2(\mathcal{O})$ for a bounded open set $\mathcal{O} \subset \mathbb{R}^3$ and a bounded function ξ.

B.1. Statement of the results and consequences. Let us first point out that the embedding of $H(\operatorname{curl}, \operatorname{div}\xi; \mathcal{O})$ into $L^2(\mathcal{O})$ is not compact; we have to add some boundary condition to ensure compactness. In this paragraph, we assume that $\mathcal{O}$ is a bounded simply connected open set in $\mathbb{R}^3$ with Lipschitz-continuous boundary $\partial\mathcal{O}$. The following result may be found in Costabel [6]; it is related to the case of a constant function ξ.

PROPOSITION B.1. *The embedding of each of the following spaces into $L^2(\mathcal{O})$ is compact:*

$$\left\{U \in H(\operatorname{curl}, \operatorname{div}; \mathcal{O}) \;\middle|\; U \wedge n \in L^2(\partial\mathcal{O})^3\right\}, \tag{B.1}$$

$$\left\{U \in H(\operatorname{curl}, \operatorname{div}; \mathcal{O}) \;\middle|\; U \cdot n \in L^2(\partial\mathcal{O})\right\}. \tag{B.2}$$

These properties can be readily extended to the case of a bounded function ξ with bounded gradient. Indeed, we have in this case $H(\operatorname{curl}, \operatorname{div}\xi; \mathcal{O}) = H(\operatorname{curl}, \operatorname{div}; \mathcal{O})$. The case of a function ξ that is only assumed bounded has been dealt with by Weber [24] for homogeneous boundary conditions, which may be summarized as follows.

PROPOSITION B.2. *Let $\xi \in L^\infty(\mathcal{O})$ such that $\Re e\, \xi > \alpha$ a.e. in $\mathcal{O}$ ($\alpha > 0$ fixed). The embedding of each of the following spaces into $L^2(\mathcal{O})$ is compact:*

$$\{U \in H(\operatorname{curl}, \operatorname{div}\xi; \mathcal{O}) \mid U \wedge n = 0 \text{ on } \partial\mathcal{O}\}, \tag{B.3}$$

$$\{U \in H(\operatorname{curl}, \operatorname{div}\xi; \mathcal{O}) \mid \xi U \cdot n = 0 \text{ on } \partial\mathcal{O}\}. \tag{B.4}$$

In view of these propositions, it seems natural to wonder whether these latter compactness properties hold if we replace the homogeneous boundary conditions by $U \wedge n \in L^2(\partial\mathcal{O})^3$ or $\xi U \cdot n \in L^2(\partial\mathcal{O})$. The answer is given in the following proposition, which is proved in the next paragraph.

PROPOSITION B.3. *Let $\xi \in L^\infty(\mathcal{O})$ such that $\Re e\, \xi > \alpha$ a.e. in $\mathcal{O}$ ($\alpha > 0$ fixed).*

(i) *The embedding of the following space into* $L^2(\mathcal{O})$ *is compact:*

$$\{U \in H(\text{curl}, \text{div}\,\xi; \mathcal{O}) \mid \xi U \cdot n \in L^2(\partial\mathcal{O})\}. \tag{B.5}$$

(ii) *Suppose that* $\partial\mathcal{O}$ *is regular enough so that the following* ε*-regularity condition is satisfied for some* $\varepsilon > 0$. *The solution* φ *to*

$$\begin{aligned} \Delta\varphi &= 0 \quad \text{in } \mathcal{O}, \\ \partial_n\varphi &= f \quad \text{on } \partial\mathcal{O} \end{aligned}$$

belongs to $H^{3/2+\varepsilon}(\mathcal{O})/\mathbb{R}$ *for every* $f \in H^{1/2}(\mathcal{O})$ *such that* $\int_{\partial\mathcal{O}} f\,d\gamma = 0$. *Then the embedding of*

$$\{U \in H(\text{curl}, \text{div}\,\xi; \mathcal{O}) \mid U \wedge n \in L^2(\partial\mathcal{O})^3\} \tag{B.6}$$

into $L^2(\mathcal{O})$ *is compact.*

Remark B.4. In case (ii), the boundary $\partial\mathcal{O}$ must be more regular than Lipschitz continuous. For instance, we can choose a surface of class $C^{1,1}$ since a solution to the Laplace equation with a $H^{1/2}(\partial\mathcal{O})$ normal derivative belongs in this case to $H^2(\mathcal{O})$.

As we will see in the proof of Corollary B.5, we can manage without these latter properties for the problem studied in the present paper. However, they become necessary if the boundary condition on the perfect conductor is replaced by an impedance condition.

COROLLARY B.5. *Let* $\mathcal{H}_T^{E/H}$ *be the Hilbert spaces defined in* (4.8), *where the domain* $\hat{\Omega}$ *is assumed to have a Lipschitz-continuous boundary* $\Gamma \cup \Sigma$, *and* $\xi \in L^\infty(\hat{\Omega})$ *satisfies assumptions* (2.3) *and* (2.4). *The embedding of* $\mathcal{H}_T^{E/H}$ *into* $L^2(\hat{\Omega})$ *is compact.*

Proof. Let $\{\mathcal{S}^{(i)} \mid 1 \le i \le I\}$ be a finite family of regular open sets that covers $\overline{\hat{\Omega}}$ (i.e., $\overline{\hat{\Omega}} \subset \bigcup_{1\le i\le I} \mathcal{S}^{(i)}$) such that each $\mathcal{O}^{(i)} = \mathcal{S}^{(i)} \cap \hat{\Omega}$ is a simply connected domain with a Lipschitz-continuous boundary. Suppose that this family is chosen so that no $\mathcal{O}^{(i)}$ contains simultaneously a part of the boundary Σ and a part of the inhomogeneous medium (which is possible because of the assumptions on Σ; see Fig. 4.1). Consider then a partition of unity $\{\alpha^{(i)} \mid 1 \le i \le I\}$ associated with this family, i.e., a family of functions $\alpha^{(i)}$ that satisfy

$$\alpha^{(i)} \in \mathcal{D}(\mathcal{S}^{(i)}), \quad 0 \le \alpha^{(i)} \le 1, \quad \text{and} \quad \sum_{i=1}^{I} \alpha^{(i)} = 1 \quad \text{in } \hat{\Omega}.$$

Let $\{U_k \mid k \in \mathbb{N}\}$ be a bounded sequence of $\mathcal{H}_T^{E/H}$; let us prove that there exists a subsequence of U_k that converges in $L^2(\hat{\Omega})$.

We clearly have $U_k = \sum_{1\le i\le I} \alpha^{(i)} U_k$. Let $U_k^{(i)}$ denote the restriction of $\alpha^{(i)} U_k$ to $\mathcal{O}^{(i)}$. By virtue of the assumptions stated above, it is readily seen that each $\{U_k^{(i)} k \in \mathbb{N}\}$ defines a bounded sequence in one of the spaces given in Propositions B.1 or B.2. Starting for instance with $U_k^{(1)}$, we can extract a subsequence (still denoted by $U_k^{(1)}$) that converges in $L^2(\mathcal{O}^{(1)})$. Going on with the associated subsequence of $U_k^{(2)}$, we can extract from it a subsequence (still denoted by $U_k^{(2)}$) that converges in $L^2(\mathcal{O}^{(2)})$, and so on, in a finite number of steps; the resulting subsequences $U_k^{(i)}$, respectively, converge in $L^2(\mathcal{O}^{(i)})$, which shows that U_k converges in $L^2(\hat{\Omega})$. □

B.2. Proof of Proposition B.3. The presentation of the proof follows the lines of [24]; however, it is shorter since we use some classical decomposition results (given, for instance, in [8]).

We denote by U_k a bounded sequence in (B.5) or (B.6). Our aim is to prove that some subsequence of U_k converges in $L^2(\mathcal{O})$.

B.2.1. Normal boundary conditions. First consider the case of (B.5), that is, $\xi U_k \cdot n$ is bounded in $L^2(\partial\mathcal{O})$. From Lemma B.6 (with $F_0 = \emptyset$ and $F_1 = \partial\mathcal{O}$), we see that U_k has the decomposition $U_k = \operatorname{grad}\varphi_k + W_k$ where $\varphi_k \in H^1(\mathcal{O})/\mathbb{R}$ is defined by

$$\int_{\mathcal{O}} \xi \operatorname{grad}\varphi_k \cdot \overline{\operatorname{grad}\psi} = \int_{\mathcal{O}} \xi U_k \cdot \overline{\operatorname{grad}\psi} \quad \forall \psi \in H^1(\mathcal{O})/\mathbb{R},$$

and W_k is such that

$$\operatorname{curl} W_k = \operatorname{curl} U_k, \quad \operatorname{div} \xi W_k = 0, \quad \text{and} \quad (\xi W_k \cdot n)_{|\partial\mathcal{O}} = 0.$$

Let us prove separately that some subsequences of $\operatorname{grad}\varphi_k$ and W_k converge in $L^2(\mathcal{O})$.

(i) Set $U_{lk} = U_l - U_k$ and $\varphi_{lk} = \varphi_l - \varphi_k$. From the definition of φ_k, we readily have (integrating by parts)

$$\int_{\mathcal{O}} \xi \left\|\operatorname{grad}\varphi_{lk}\right\|^2 = \int_{\mathcal{O}} \xi U_{lk} \cdot \overline{\operatorname{grad}\varphi_{lk}} = -\int_{\mathcal{O}} \operatorname{div}(\xi U_{lk})\, \overline{\varphi_{lk}} + \int_{\partial\mathcal{O}} (\xi U_{lk} \cdot n) \overline{\varphi_{lk}}\, d\gamma.$$

As φ_k is bounded in $H^1(\mathcal{O})/\mathbb{R}$, we can extract a subsequence (still denoted φ_k) that converges in $L^2(\mathcal{O})/\mathbb{R}$ and whose trace on $\partial\mathcal{O}$ converges in $L^2(\partial\mathcal{O})/\mathbb{R}$. The convergence of $\operatorname{grad}\varphi_k$ in $L^2(\mathcal{O})$ follows. (Note that we use here the fact that $\xi U_k \cdot n$ is bounded in $L^2(\partial\mathcal{O})$.)

(ii) To deal with W_k, we apply [8, Thm. 3.6, p. 48], which shows the existence of Ψ_k in $H(\operatorname{curl};\mathcal{O})$ such that

$$\operatorname{curl}\Psi_k = \xi W_k, \quad \operatorname{div}\Psi_k = 0, \quad \text{and} \quad (\Psi_k \wedge n)_{|\partial\mathcal{O}} = 0.$$

From Costabel [6, Thm. 2], we deduce that $\Psi_k \in H^{1/2}(\mathcal{O})^3$ and $\left\|\Psi_k\right\|_{H^{1/2}(\mathcal{O})^3} \leq C \left\|\xi W_k\right\|_{L^2(\mathcal{O})^3}$ and thus deduce the existence of a subsequence, still denoted Ψ_k, converging in $L^2(\mathcal{O})^3$. Since

$$\int_{\mathcal{O}} \xi \left\|W_{lk}\right\|^2 = \int_{\mathcal{O}} \operatorname{curl}\Psi_{lk} \cdot \overline{W_{lk}} = \int_{\mathcal{O}} \Psi_{lk} \cdot \overline{\operatorname{curl} W_{lk}}$$

(where $W_{lk} = W_l - W_k$ and $\Psi_{lk} = \Psi_l - \Psi_k$), W_k converges in $L^2(\mathcal{O})^3$.

B.2.2. Tangential boundary conditions. Consider now a sequence U_k bounded in (B.6) ($U_k \wedge n$ is bounded in $L^2(\partial\mathcal{O})$). From Lemma B.6 (with $F_0 = \partial\mathcal{O}$ and $F_1 = \emptyset$), we have the decomposition $U_k = \operatorname{grad}\varphi_k + W_k$, where $\varphi_k \in H_0^1(\mathcal{O})$ is now defined by

$$\int_{\mathcal{O}} \xi \operatorname{grad}\varphi_k \cdot \overline{\operatorname{grad}\psi} = \int_{\mathcal{O}} \xi U_k \cdot \overline{\operatorname{grad}\psi} \quad \forall \psi \in H_0^1(\mathcal{O}),$$

and W_k is such that

$$\operatorname{curl} W_k = \operatorname{curl} U_k, \quad \operatorname{div} \xi W_k = 0, \quad \text{and} \quad (W_k \wedge n)_{|\partial\mathcal{O}} = (U_k \wedge n)_{|\partial\mathcal{O}}.$$

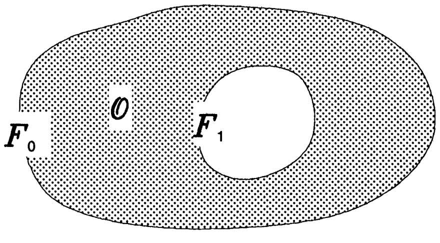


FIG. B.1.

(i) To see that some subsequence of $\operatorname{grad}\varphi_k$ converges in $L^2(\mathcal{O})$, we proceed exactly as above (no boundary term appears in the integration by parts since $\varphi_k = 0$ on $\partial\mathcal{O}$).

(ii) For W_k, we apply [8, Thm. 3.5, p. 47] which shows the existence of Φ_k in $H(\operatorname{curl};\mathcal{O})$ such that

$$\operatorname{curl}\Phi_k = \xi W_k, \quad \operatorname{div}\Phi_k = 0, \quad \text{and} \quad (\Phi_k \cdot n)_{|\partial\mathcal{O}} = 0.$$

Provided $\partial\mathcal{O}$ satisfies the ε-regularity condition, we deduce from Lemma B.8 that $\Phi_k \in H^{1/2+\varepsilon}(\mathcal{O})^3$ and $\|\Phi_k\|_{H^{1/2+\varepsilon}(\mathcal{O})^3} \leq C\,\|\xi W_k\|_{L^2(\mathcal{O})^3}$. Consequently, there exists a subsequence Φ_k converging in $L^2(\mathcal{O})^3$ such that $(\Phi_k \wedge n)_{|\partial\mathcal{O}}$ converges in $L^2(\partial\mathcal{O})^2$. Since

$$\int_{\mathcal{O}} \xi\,\|W_{lk}\|^2 = \int_{\mathcal{O}} \Phi_{lk}\cdot\overline{\operatorname{curl} W_{lk}} + \int_{\partial\mathcal{O}} \Phi_{lk}\cdot\overline{W_{lk}\wedge n}\,d\gamma,$$

the convergence of W_k in $L^2(\mathcal{O})^3$ follows.

B.3. Auxiliary lemmas. The following lemma concerns a decomposition of vector fields that allows adding the divergence-free condition in the function spaces related to the classical Maxwell equation (see §2.3). It is also used for the compactness results shown above.

LEMMA B.6. *Let $\mathcal{O}$ be an open set of $\mathbb{R}^3$ with a Lipschitz-continuous boundary that consists of two disjointed parts F_0 and F_1 as shown in Fig.* B.1 (*where one of them may be $\emptyset$*). *Let $\xi \in L^\infty(\mathcal{O})$ such that $\Re e\,\xi > \alpha$ a.e. in $\mathcal{O}$ ($\alpha > 0$ fixed). Then every function $V \in L^2(\mathcal{O})^3$ has the decomposition*

$$\text{(B.7)}\qquad \begin{gathered} V = \operatorname{grad}\varphi + V' \quad \textit{where} \\ \varphi \in H^1(\mathcal{O}), \quad \varphi = 0 \textit{ on } F_0, \quad \textit{and} \\ V' \in L^2(\mathcal{O})^3, \quad \operatorname{div}\xi V' = 0 \quad \textit{in } \mathcal{O}, \quad \textit{and} \quad \xi V'\cdot n = 0 \quad \textit{on } F_1. \end{gathered}$$

Moreover, $\operatorname{curl} V' = \operatorname{curl} V$ *and* $\operatorname{grad}\varphi \wedge n = 0$ *on* F_0.

Remark B.7. Note that $\xi V' \cdot n$ and $\operatorname{grad}\varphi \wedge n$ are defined, respectively, in $H^{-1/2}(\partial\mathcal{O})$ and $H^{-1/2}(\partial\mathcal{O})^3$ since $\operatorname{div}\xi V' = 0$ and $\operatorname{curl}(\operatorname{grad}\varphi) = 0$.

Proof. Let $\mathcal{H}$ denote the Hilbert space $\{\psi \in H^1(\mathcal{O}) \mid \psi_{|F_0} = 0\}$ if $F_0 \neq \emptyset$ and $\mathcal{H} = H^1(\mathcal{O})/\mathbb{R}$ if $F_0 = \emptyset$. Let V be a given function of $L^2(\mathcal{O})^3$. First, notice that there exists a unique function $\varphi \in \mathcal{H}$ such that

$$\int_{\mathcal{O}} \xi \operatorname{grad}\varphi\cdot\overline{\operatorname{grad}\psi} = \int_{\mathcal{O}} \xi V\cdot\overline{\operatorname{grad}\psi} \quad \forall\psi \in \mathcal{H}.$$

This follows from the Lax–Milgram theorem (by virtue of the assumptions on ξ, the left-hand side of this relation defines a continuous and coercive form in $\mathcal{H}$). In other words, φ is formally the only solution to

$$\begin{aligned} \operatorname{div}(\xi \operatorname{grad} \varphi) &= \operatorname{div} \xi V \quad \text{in } \mathcal{O}, \\ \varphi &= 0 \quad \text{on } F_0, \\ \partial_n \varphi &= V \cdot n \quad \text{on } F_1. \end{aligned}$$

Consider then the function $V' = (V - \operatorname{grad} \varphi) \in L^2(\mathcal{O})^3$ that satisfies by construction of φ the divergence-free condition $\operatorname{div} \xi V' = 0$ (in the sense of distributions) as well as the boundary condition $\xi V' \cdot n = 0$ on F_1. We readily have $\operatorname{curl} V' = \operatorname{curl} V$. It remains to prove that $\operatorname{grad} \varphi \wedge n = 0$ on F_0, which amounts to showing that

$$\int_{\mathcal{O}} \operatorname{grad} \varphi \cdot \operatorname{curl} W = 0$$

for every $W \in \mathcal{D}(\overline{\mathcal{O}})^3$ whose support is localized in a vicinity of F_0 ($W \equiv 0$ near F_1). This property simply follows from Green's formula:

$$\int_{\mathcal{O}} \operatorname{grad} \varphi \cdot \operatorname{curl} W = -\int_{\mathcal{O}} \varphi \operatorname{div}(\operatorname{curl} W),$$

where the boundary integral vanishes since $\varphi = 0$ on F_0. □

The following regularity lemma is derived from Costabel [6].

LEMMA B.8. *Let $\mathcal{O}$ be a simply connected bounded open set satisfying the ε-regularity condition stated in point* (ii) *of Proposition* B.3. *Then the space*

$$\mathcal{V} = \left\{ U \in H(\operatorname{curl}; \mathcal{O}) \,\middle|\, \operatorname{div} U = 0 \;\; \textit{and} \;\; (U \cdot n)_{|\partial\mathcal{O}} = 0 \right\}$$

is embedded in $H^{1/2+\varepsilon}(\mathcal{O})^3$. Moreover, there exists a constant $C > 0$ such that for every $U \in \mathcal{V}$, we have

$$\|U\|_{H^{1/2+\varepsilon}(\mathcal{O})^3} \leq C \|U\|_{\mathcal{V}}.$$

Proof. We proceed exactly as in [6].

(i) Let $U \in \mathcal{V}$. From [8, Thm. 3.4, p. 45], we deduce the existence of $\Phi \in H^1(\mathcal{O})^3$ such that

$$\operatorname{curl} \Phi = \operatorname{curl} U \quad \text{and} \quad \operatorname{div} \Phi = 0.$$

Let $\Psi = U - \Phi$; we have $\Psi \in L^2(\mathcal{O})^3$ with $\operatorname{curl} \Psi = 0$ and by the Stokes theorem [8, Thm. 2.9, p. 31] the existence of $q \in H^1(\mathcal{O})$ such that $\Psi = \operatorname{grad} q$. It follows that q satisfies

$$\begin{aligned} \Delta q &= 0 \quad \text{in } \mathcal{O}, \\ \partial_n q &= -(\Phi \cdot n)_{|\partial\mathcal{O}} \;\in\; H^{1/2}(\partial\mathcal{O}). \end{aligned}$$

By the ε-regularity hypothesis, $q \in H^{3/2+\varepsilon}(\mathcal{O})$, and consequently $U \in H^{1/2+\varepsilon}(\mathcal{O})^3$.

(ii) The estimation follows from the closed graph theorem. We denote by I the identity in $L^2(\mathcal{O})^3$ and by J its restriction to $\mathcal{V}$ considered as an application between $\mathcal{V}$ and $H^{1/2+\varepsilon}(\mathcal{O})^3$. Let (U_n, JU_n) be a sequence that converges in $\mathcal{V} \times H^{1/2+\varepsilon}(\mathcal{O})^3$, say to (U, Y). Considered as a sequence in $L^2(\mathcal{O}) \times L^2(\mathcal{O})$, $(U_n, JU_n) = (U_n, IU_n)$ converges to (U, IU), from which we deduce that $Y = IU = JU$; that is, the graph of J is closed. As a consequence, the natural embedding $\mathcal{V} \longrightarrow H^{1/2+\varepsilon}(\mathcal{O})^3$ is continuous. □

REFERENCES

[1] T. Abboud, *Etude Mathématique et Numérique de Quelques Problèmes de Diffraction d'Ondes Électromagnétiques*, Thesis, École Polytechnique, Palaiseau, France, 1991.

[2] T. Abboud and J. C. Nédélec, *Electromagnetic waves in an inhomogeneous medium*, J. Math. Anal. Appl., 164 (1992), pp. 40–58.

[3] A. Bamberger and A. S. Bonnet, *Mathematical analysis of the guided modes of an optical fiber*, SIAM J. Math. Anal., 21 (1990), pp. 1487–1510.

[4] A. Bendali, *Approximation par éléments finis de surface de problèmes de diffraction des ondes électromagnétiques*, Thesis, Université Pierre et Marie Curie, Paris, 1984.

[5] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, Krieger, Melbourne, FL, 1992.

[6] M. Costabel, *A remark on the regularity of solutions of Maxwell's equations on Lipschitz domains*, Math. Methods Appl. Sci., 12 (1990), pp. 365–368.

[7] ———, *A coercive bilinear form for Maxwell's equations*, J. Math. Anal. Appl., 157 (1991), pp. 527–541.

[8] V. Girault and P. A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1986.

[9] A. Jami and M. Lenoir, *A variational formulation for exterior problems in linear hydrodynamics*, Comput. Methods Appl. Mech. Engrg., 16 (1978), pp. 341–359.

[10] D. S. Jones, *Acoustic and Electromagnetic Waves*, Oxford University Press, London, 1986.

[11] C. E. Kenig, *Carleman estimates, uniform Sobolev inequalities for second-order differential operators, and unique continuation theorems*, in Proc. International Congress of Mathematicians, vol. 2, International Congress of Mathematicians, Berkeley, CA, 1986, pp. 948–960.

[12] F. Kikuchi, *On a discrete compactness property for the Nedelec finite elements*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 36 (1989), pp. 479–490.

[13] W. Knauff and R. Kress, *On the exterior boundary-value problem for the time-harmonic Maxwell equations*, J. Math. Anal. Appl., 72 (1979), pp. 215–235.

[14] V. D. Kupradze, *Potential Methods in the Theory of Elasticity*, Israel Program for Scientific Translations, Jerusalem, 1965.

[15] R. Leis, *Zur theorie elektromagnetischer schwingungen in anisotropen inhomogenen medien*, Math. Z., 106 (1968), pp. 213–224.

[16] V. Levillain, *Couplage éléments finis-équations intégrales pour la résolution des équations de Maxwell en milieu hétérogène*, Thesis, École Polytechnique, Palaiseau, France, 1991.

[17] I.D. Mayergoyz and J. D'Angelo, *A new point of view on the mathematical structure of Maxwell's equations*, IEEE Trans. Mag., 29 (1993), pp. 1315–1320.

[18] C. Müller, *Foundations of the Mathematical Theory of Electromagnetic Waves*, Springer-Verlag, Berlin, 1969.

[19] J. Nečas, *Les Methodes Directes en Théorie des Equations Elliptiques*, Masson, Paris, 1967.

[20] J. C. Nédélec, *Mixed finite element in R^3*, Numer. Math., 35 (1980), pp. 315–341.

[21] F. Rellich, *Über das asymptotische verhalten der Lösungen von $\Delta u + \lambda u = 0$ in unendlichen gebieten*, Jahresber Deutsch. Math.-Verein., 53 (1943), pp. 57–65.

[22] L. Schwartz, *Théorie des Distributions*, Hermann, Paris, 1966.

[23] V. Vogelsang, *On the strong unique continuation principle for inequalities of Maxwell type*, Math. Ann., 289 (1991), pp. 285–295.

[24] C. Weber, *A local compactness theorem for Maxwell's equations*, Math. Methods Appl. Sci., 2 (1980), pp. 12–25.

[25] P. Werner, *On the exterior boundary value problem of perfect reflection for stationary electromagnetic wave fields*, J. Math. Anal. Appl., 7 (1963), pp. 348–396.

[26] C. H. Wilcox, *Scattering Theory for the d'Alembert Equation in Exterior Domains*, Springer-Verlag, Berlin, 1975.

SIAM J. MATH. ANAL.
Vol. 27, No. 6, pp. 1631–1652, November 1996

006

SPECTRAL ANALYSIS OF A MULTISTRATIFIED ACOUSTIC STRIP PART II: ASYMPTOTIC BEHAVIOR OF SOLUTIONS FOR A SIMPLE STRATIFICATION*

ELISABETH CROC† AND YVES DERMENJIAN†

Abstract. We consider the acoustic propagator $A = -\nabla.c^2\nabla$ in $\Omega = \{(x,z) \in \mathbb{R}^2 \; / \; 0 < z < H\}$. The velocity c, which describes the stratification of the strip Ω, depends only on the variable z: it is assumed to be a function in $L^\infty((0,H))$ bounded from below by $c_m > 0$. Let A be the self-adjoint operator associated with the Neumann or Dirichlet condition at $z = 0$ and $z = H$; let μ be a real number in the spectrum of A; and let u be the solutions of the equation $(A - \mu I)u = f$ locally in the domain of A, which are determined by the limiting absorption principle in [E. Croc and Y. Dermenjian, *SIAM J. Math. Anal.*, 26 (1995), pp. 880–924] and made explicit with trace operators. Thanks to accurate Hölder properties for the trace operators, we control the asymptotic behavior of u with so-called "zero-trace" conditions for f.

Key words. stratified medium, acoustic waves, self-adjoint operator, resolvent, limiting absorption principle, bootstrap theorem

AMS subject classifications. 35L05, 35P, 47A70

1. Introduction.

This article is the continuation of [CD95a]. Let us recall the problem: the modeling of a particular seismic problem leads to a scalar wave equation written as

$$(\partial_t)^2 v + Av = S$$

in the strip

$$\Omega = \{(x,z) \in \mathbb{R}^2 \; / \; 0 < z < H\}.$$

The unknown function $v(t,x,z)$ is the displacement in the medium: it is a real-valued function defined for t in $\mathbb{R}$ and (x,z) in Ω.

The spatial operator is the differential operator $A = -\nabla.c^2\nabla = -\partial_x(c^2(x,z)\partial_x) - \partial_z(c^2(x,z)\partial_z)$ acting in Ω. The coefficient $c(x,z)$ characterizes the medium celerity: it is a measurable function in Ω which satisfies $0 < c_m \leq c(x,z) \leq c_M$ for almost every (x,z) in Ω and with some real numbers $c_m, c_M > 0$. Our goal is to deal with media stratified differently according to whether $x < 0$ or $x > 0$: it means that the celerity function is such that $c(x,z) = c_-(z)$ if $x < 0$ and $c(x,z) = c_+(z)$ if $x > 0$.

The right-hand side, $S(t,x,z)$, is a given source. It is a function in $L^2(\Omega)$, the Hilbert space of complex-valued functions defined on Ω which are Lebesgue measurable and square integrable.

The boundary conditions (BC) for the displacement $v(t,x,z)$ are the Neumann or the Dirichlet conditions in $z = 0$ and $z = H$. For instance, geophysical problems lead to the following boundary conditions:

$$c^2 \partial_z v|_{z=0} = 0 \quad \text{and} \quad v|_{z=H} = 0. \tag{1.1}$$

Choosing a constructive stationary approach, we intend to obtain the spectral and scattering theory for the acoustic propagator, that is, the self-adjoint operator

* Received by the editors July 18, 1994; accepted for publication July 24, 1995.

† U.F.R. Mathématiques Informatique Mécanique, Université de Provence, Case V, 3 Place Victor-Hugo, F-13331 Marseille cedex 3, France (ecroc@gyptis.univ-mrs.fr, dermenji@gyptis.univ-mrs.fr).

$(D(A), A)$ acting in $L^2(\Omega)$ that we associate with the problem. The domain $D(A)$ is included in the Sobolev space

$$H^1(\Omega) = \{u \in L^2(\Omega) \;/\; \partial_x u,\ \partial_z u \in L^2(\Omega)\}$$

and is equal to

$$D(A) = \{u \in H^1(\Omega) \;/\; -\nabla.c^2\nabla u \in L^2(\Omega),\ u \text{ satisfies (BC)}\}. \tag{1.2}$$

We call free or unperturbed problem the case where the strip is only stratified in the z direction. Therefore, the celerity c is a function of the single variable z in $(0, H)$ such that $c_-(z) = c_+(z) = c(z)$. The free differential operator is written as

$$A = -c^2(z)(\partial_x)^2 - \partial_z(c^2(z)\partial_z). \tag{1.3}$$

A particular case that could serve as a typical problem is the case of a two-valued function

$$c(z) = c_1 \quad \text{if } 0 < z < h \quad \text{and} \quad c(z) = c_2 \quad \text{if } h < z < H \tag{1.4}$$

with $0 < h < H$ and $0 < \text{Min}(c_1, c_2)$. More generally, we consider a celerity $c(z)$ which satisfies the assumption

$$c \in L^\infty((0, H)) \quad \text{and} \quad \text{Min}\, c(z) \geq c_m > 0. \tag{H}$$

In [CD95a], we developed an explicit spectral analysis of the free operator (1.2)–(1.3) with boundary conditions (1.1) or (BC), and then we deduced a limiting absorption principle for the resolvent operator $R_A(\mu + i\varepsilon) = (A - (\mu + i\varepsilon)I)^{-1}$ when ε goes to 0 and μ is in the spectrum of A.

In this article, the results concern $u^\pm = R_A^\pm(\mu)f$, the solutions of the equation $(A - \mu I)u = f$ which are obtained with an appropriate f by this limiting absorption principle. The article is organized as follows.

In §2, we fix our notation and recall the results of [CD95a]. Then we present our main results. These are the so-called division or bootstrap theorems (Theorems 2.1 and 2.2) for the free operator (1.2)–(1.3) under assumption (H). Under some assumptions on the function f, these division theorems specify the asymptotic behavior of $u^\pm$ when (x, z) tends to infinity in Ω. Moreover, the stated and used results allow us to again visit the limiting absorption principle at thresholds (Theorem 2.3).

In §3, we look for the basic estimates needed to prove the previous theorems. These estimates are essentially derived from Agmon's results, specifically from Lemma B.2 of [A]. They concern the terms that arise in the explicit formula of the resolvent and rely on a thorough study of trace operators. On one hand, we state Sobolev properties and differentiability (see Propositions 3.2–3.5) of the generalized Fourier coefficients (equation (2.5) in §2) of a function in $L^2(\Omega)$. These coefficients are associated with the spectral representation of A obtained in [CD95a]. On the other hand, we collect some Hölder estimates (see Propositions 3.8 and 3.9) scattered throughout [CD95a] and based on Theorem 3.7. Finally, we use [A, Lem. B.2] and Corollary 3.1 to state new estimates (see Propositions 3.10 and 3.11).

In §4, we complete the proofs of Theorems 2.1 and 2.2. We use the limiting absorption principles from [CD95a], called LAP1 and LAP2.

In §5, we prove Theorem 2.3, also called LAP3. The thresholds are the eigenvalues $\lambda(0, m)$, $m \geq 1$ (see equation (2.2) in §2), of the transverse operator $B = -d_z(c^2(z)d_z)$

on $(0,H)$ with domain $D(B) = \{u \in H^1((0,H)) \ / \ Bu \in L^2((0,H)),\ u$ satisfies (BC)$\}$. The statement of LAP3 is the convergence in the space $L^2_{-s}(\Omega) = \{f \ / \ (1+x^2)^{-s/2} f \in L^2(\Omega)\}$ of $R_A(\zeta)f$ when ζ, $\pm\text{Im}\zeta > 0$, tends to a threshold $\mu = \lambda(0,m)$. Sufficient conditions for such a convergence are as follows: $s > 3/2$ and f is in the space $E_s(\mu) = NL^2_s(m,1)$. This space, appearing in Theorem 2.2 and considered in Proposition 3.6, has codimension 2 in $L^2_s(\Omega)$.

We now call the perturbed problem the case of the strip where the celerity is such that $c(x,z) = c_-(z)$ if $x < 0$ and $c(x,z) = c_+(z)$ if $x > 0$ with two distinct functions c_- and c_+ satisfying the assumption (H). With the limiting absorption principles and division theorems for the both free operators associated with the celerities c_- and c_+, we are in a position to deduce a limiting absorption principle for the associated perturbed operator. There are many examples of such a path from free operators to perturbed operators with short range perturbations (see, for example, [A], [DG], [H], [V], and [We]) or long-range perturbations (see, for example, [T81a] and [T81b]). The particularity of the problem under consideration is the radical difference between the behavior of the celerity when x goes to $-\infty$ and when x goes to $+\infty$. It will be studied in a third paper.

2. Notation, review, and results. The operator $(D(A),A)$ is defined by (1.2) and (1.3). The celerity c is a function of z and satisfies assumption (H).

We now present some notation and useful results from [CD95a].

For any real number s, let $L^2_s(\Omega)$ be the weighted space $L^2_s(\Omega) = \{f \ / \ (1+x^2)^{s/2} f \in L^2(\Omega)\}$ equipped with the Hilbertian norm

$$\|f\|_{L^2_s(\Omega)} = \|(1+x^2)^{s/2} f\|_{L^2(\Omega)} = \left(\int_\Omega (1+x^2)^s |f(x,z)|^2 dxdz\right)^{1/2}.$$

The dual $L^2_s(\Omega)'$ and the space $L^2_{-s}(\Omega)$ are isometric. They are identified through the duality bracket

$$\langle f,g\rangle_{L^2_s(\Omega),\ L^2_{-s}(\Omega)} = \int_\Omega f(x,z)g(x,z)dxdz.$$

Let $u \mapsto \mathcal{F}u(\xi) = (2\pi)^{-1/2}\int_{\mathbb{R}} e^{-i\xi x}u(x)dx$ be the Fourier transform on $L^2(\mathbb{R})$. The partial Fourier transform with respect to the variable x on $L^2(\Omega)$ is also denoted $\mathcal{F}$.

For any real number ξ, let $(V(\xi,n,.))_{n\geq 1}$ be an orthonormal basis of the space $L^2((0,H))$ which satisfies the eigenvalue problem

$$-(c^2V')' + c^2\xi^2 V = \lambda V \text{ on } (0,H), \quad V \text{ satisfies (BC)}. \tag{2.1}$$

This problem is studied in §4.2 of [CD95a]. The eigenfunctions $V(\xi,n,.)$ can be chosen real. The eigenvalues λ are simple and strictly positive. For each real number ξ, they define an unbounded real sequence $(\lambda(\xi,n))_{n\geq 1}$, increasingly ordered with n.

The operator $(D(A),A)$ is self-adjoint in the space $L^2(\Omega)$. The spectrum $\sigma(A)$ of the operator A is the half-line $[\lambda(0,1),+\infty) \subseteq [(\pi^2/4H^2)c^2_m,+\infty)$. The eigenvalues of (2.1) for $\xi = 0$ appear as exceptional values or thresholds, and we set

$$\Gamma(A) = \{\lambda(0,n) \ / \ n \geq 1\}. \tag{2.2}$$

The properties of the dispersion curves

$$\xi \mapsto \lambda(\xi,n), \quad n \geq 1, \tag{2.3}$$

are described in Theorem 4.5 of [CD95a]. They are strictly increasing on $[0,+\infty)$ and even and analytic on $\mathbb{R}$, and there exist functions $\xi \mapsto a_n(\xi)$ such that

$$(2.4)\qquad \lambda(\xi,n) = \lambda(0,n) + \xi^2 a_n(\xi) \quad \text{with } c_m^2 \le a_n(\xi) \le \|c\|^2_{L^\infty((0,H))}.$$

Looking at the restriction of $\xi \mapsto \lambda(\xi,n)$ to the interval $[0,+\infty)$, $\lambda \mapsto \xi(\lambda,n)$ is the monotone inverse function which maps $[\lambda(0,n),+\infty)$ onto $[0,+\infty)$.

For any function f in the space $L^2(\Omega)$, any integer $n \ge 1$, for $j = 1$ or 2, the generalized Fourier coefficients are the complex-valued functions defined for almost every real number ξ and for almost every real number λ in the interval $(\lambda(0,n),+\infty)$ by

$$(2.5)\quad \widetilde{f}(\xi,n) = (2\pi)^{-1/2}\int_\Omega f(x,z)e^{-i\xi x}V(\xi,n,z)dxdz = \int_0^H \mathcal{F}f(\xi,z)V(\xi,n,z)dz,$$

$$(2.6)\qquad f^j(\lambda,n) = (\partial_\lambda \xi(\lambda,n))^{1/2}\,\widetilde{f}((-1)^j\xi(\lambda,n),n).$$

The coefficients (2.5) and (2.6), respectively, are associated with the spectral representation of the self-adjoint operator $(D(A),A)$, which is made explicit in Theorems 3.1 and 3.2, respectively, of [CD95a]. The functions $\widetilde{f}(.,n)$ and $f^j(.,n)$ satisfy

$$(2.7)\qquad \sum_{n\ge1}\int_{-\infty}^{+\infty}|\widetilde{f}(\xi,n)|^2d\xi = \sum_{n\ge1}\sum_{j=1}^{2}\int_{\lambda(0,n)}^{+\infty}|\,f^j(\lambda,n)|^2d\lambda < +\infty.$$

The function f in the space $L^2(\Omega)$ and the scalar product $(f \mid g)_{L^2(\Omega)}$ with a function g in $L^2(\Omega)$ can be written as

$$(2.8)\qquad f(x,z) = \sum_{n\ge1}(2\pi)^{-1/2}\int_{-\infty}^{+\infty}\widetilde{f}(\xi,n)e^{i\xi x}V(\xi,n,z)\,d\xi,$$

$$(2.9)\quad (f \mid g)_{L^2(\Omega)} = \sum_{n\ge1}\int_{-\infty}^{+\infty}\widetilde{f}(\xi,n)\overline{\widetilde{g}(\xi,n)}\,d\xi = \sum_{n\ge1}\sum_{j=1}^{2}\int_{\lambda(0,n)}^{+\infty}f^j(\lambda,n)\overline{g^j(\lambda,n)}\,d\lambda.$$

Let s be a real number such that $s > 1/2$.

For any real number ξ and any integer $n \ge 1$, the trace operator $\widetilde{\tau}_n(\xi)$, defined by

$$(2.10)\quad f \in L^2_s(\Omega) \mapsto \widetilde{\tau}_n(\xi)f = \widetilde{f}(\xi,n) = (2\pi)^{-1/2}\int_\Omega f(x,z)e^{-i\xi x}V(\xi,n,z)dxdz,$$

is a continuous linear form on $L^2_s(\Omega)$ with Hölder properties in the variable ξ stated in Proposition 3.2 of [CD95a]. Specifically, for a real number δ in $[0,1]$, $\delta < s - 1/2$, there exists a function $M(\xi,\xi') = M_n(s,\delta,\xi,\xi')$, continuous with respect to ξ and ξ', such that

$$(2.11)\qquad \forall f \in L^2_s(\Omega),\quad |\widetilde{\tau}_n(\xi)f - \widetilde{\tau}_n(\xi')f| \le M(\xi,\xi')|\xi-\xi'|^\delta\|f\|_{L^2_s(\Omega)}.$$

The space

$$(2.12)\qquad NL^2_s(n) = \{f \in L^2_s(\Omega)\ /\ \widetilde{\tau}_n(0)f = 0\}$$

is a closed subspace of $L^2_s(\Omega)$ with codimension one. Referring to Proposition 3.3 of [CD95a]), the space

$$W(n) = NL^2_s(n) \cap C^\infty_0(\Omega) \tag{2.13}$$

is dense in $NL^2_s(n)$ and therefore not dense in $L^2_s(\Omega)$. However, the space $W(n)$ is dense in $L^2_t(\Omega)$ for $t \leq 1/2$.

For any real number λ in $\sigma(A)$, any integer $n \geq 1$, and $j = 1$ or 2, the trace operator $\tau^j_n(\lambda)$ is then defined on $L^2_s(\Omega)$ by

$$\tau^j_n(\lambda)f \; = 0 \quad \text{if } \lambda = \lambda(0,n), \tag{2.14a}$$

$$\tau^j_n(\lambda)f = f^j(\lambda,n) = (\partial_\lambda \xi(\lambda,n))^{1/2}\, \widetilde{\tau}_n((-1)^j \xi(\lambda,n))f \quad \text{if } \lambda > \lambda(0,n). \tag{2.14b}$$

The function $\lambda \mapsto \tau^j_n(\lambda)$ maps $[\lambda(0,n),+\infty)$ in $L^2_s(\Omega)' = L^2_{-s}(\Omega)$. Its properties are stated in Proposition 3.4 of [CD95a]. Far from $\lambda(0,n)$, the function $\lambda \mapsto (\partial_\lambda \xi(\lambda,n))^{1/2}$ is analytic, and the function $\lambda \mapsto \tau^j_n(\lambda)$ has Hölder properties identical to those of the function $\xi \mapsto \widetilde{\tau}_n(\xi)$ (see (2.11)) and, in particular, the same Hölder order δ. Due to the vanishing of $\partial_\xi \lambda(.,n)$ in $\xi = 0$, the function $\lambda \mapsto \tau^j_n(\lambda)$ fails to be continuous in $\lambda(0,n)$. However, there exists a function $C(\lambda) = C_n(s,\lambda)$, continuous with respect to $\lambda \geq \lambda(0,n)$, such that

$$\forall f \in L^2_s(\Omega), \quad |\tau^j_n(\lambda)f| \leq C(\lambda)|\lambda - \lambda(0,n)|^{-1/4}\|f\|_{L^2_s(\Omega)}. \tag{2.15}$$

In the neighborhood of $\lambda(0,n)$, to get Hölder conditions for $\lambda \mapsto \tau^j_n(\lambda)$, we have to consider the restriction of $\tau^j_n(\lambda)$ to $NL^2_s(n)$ and to assume that $s > 1$. Specifically, for $s > 1$ and for a real number δ in $[0,1/4]$, $\delta < (s-1)/2$, there exists a function $M(\lambda,\lambda') = M_n(s,\delta,\lambda,\lambda')$, continuous with respect to $\lambda, \lambda' \geq \lambda(0,n)$, such that

$$\forall f \in NL^2_s(n), \quad |\tau^j_n(\lambda)f - \tau^j_n(\lambda')f| \leq M(\lambda,\lambda')|\lambda - \lambda'|^\delta \|f\|_{L^2_s(\Omega)}. \tag{2.16}$$

The following useful formulas are derived from (2.4) and are used in the proof of Proposition 3.4 of [CD95a]:

$$\partial_\xi \lambda(\pm\xi,n) = \pm\xi\; b_n(\xi), \tag{2.17a}$$

$$\xi = \xi(\lambda,n) = (\lambda - \lambda(0,n))^{1/2}\; G_n(\xi), \tag{2.17b}$$

$$\partial_\lambda \xi(\lambda,n) = (\lambda - \lambda(0,n))^{-1/2}\; H_n(\xi), \tag{2.17c}$$

$$\tau^j_n(\lambda)f = (\lambda - \lambda(0,n))^{-1/4}\; H_n(\xi)^{1/2}\; \widetilde{\tau}_n((-1)^j\; \xi)f, \tag{2.17d}$$

where $\xi \geq 0$ and the functions b_n, G_n, and H_n are strictly positive and analytic on $[0,+\infty)$.

The limiting absorption principle is valid when the real number μ is in $\sigma(A)$, that is, we can determine

$$R^\pm_A(\mu) = \lim_{\zeta \to \mu,\; \pm\mathrm{Im}\zeta > 0} (A - \zeta I)^{-1}. \tag{2.18}$$

Then the function $R^\pm_A$ defined on $\mathbb{C}^\pm = \{\zeta \in \mathbb{C} \; / \; \pm \mathrm{Im}\zeta > 0\}$ by $\zeta \mapsto R^\pm_A(\zeta) = (A - \zeta I)^{-1}$ can be continued on $\sigma(A)$. Theorem 3.4 of [CD95a] claims the following results, called LAP1 and LAP2.

- LAP1.

$$\text{If } \mu \notin \Gamma(A) = \{\lambda(0,n) \ / \ n \geq 1\} \quad \text{and} \quad s > 1/2, \tag{2.19}$$

then the limits in (2.18) exist in the uniform-operator topology on the space $B(L^2_s(\Omega), L^2_{-s}(\Omega))$. Moreover, each function $R^{\pm}_A$ is locally Hölder continuous on $\overline{\mathbb{C}^{\pm}} \setminus \Gamma(A)$ with order δ_1 in $[0, \mathrm{Min}(1, s - 1/2))$.

- LAP2.

$$\text{If } \mu = \lambda(0,n) \quad \text{and} \quad s > 1, \tag{2.20}$$

then the limits in (2.18) exist in the uniform operator topology on the space $B(NL^2_s(n), NL^2_s(n)')$. Moreover, each function $R^{\pm}_A$ is locally Hölder continuous on $J^{\pm}_n = \{\zeta \in \overline{\mathbb{C}^{\pm}} \ / \ \lambda(0, n-1) < \mathrm{Re}\zeta < \lambda(0, n+1)\}$ with order δ_2 in $[0, \mathrm{Min}(1/4, (s-1)/2))$.
Let us recall that $NL^2_s(n)$ is not dense in $L^2_s(\Omega)$, $s > 1/2$. Therefore, $L^2_{-s}(\Omega)$ cannot be identified with a subspace of $NL^2_s(n)'$.

With the above definitions and notations, we have the following statements.

THEOREM 2.1 (division theorem outside of $\Gamma(A)$). *Let $m \geq 2$ be an integer, I_m be the interval $(\lambda(0, m-1), \lambda(0, m))$, and $\mu \in I_m$. Let $s > 1/2$ be a real number and f be a function in $L^2_s(\Omega)$ such that*

$$\tau^j_n(\mu) f = 0, \quad j = 1 \, or \, 2, \quad 1 \leq n < m. \tag{2.21}$$

Then $R^+_A(\mu)f = R^-_A(\mu)f = u$, which is in $L^2_{-s}(\Omega)$ by LAP1, *belongs to $L^2_{-\tilde{s}}(\Omega)$ with $\tilde{s} = \mathrm{Max}(0, 1-s)$. Moreover, there exists a function $C = C_m(\mu, s)$, continuous with respect to the variable μ on I_m, such that for every f in $L^2_s(\Omega)$ satisfying* (2.21), *we have*

$$\|u\|_{L^2_{-\tilde{s}}(\Omega)} \leq C \|f\|_{L^2_s(\Omega)}. \tag{2.22}$$

THEOREM 2.2 (division theorem at thresholds). *Let $m \geq 1$ be an integer and $\mu = \lambda(0, m)$ be the corresponding threshold of A. Let $s > 3/2$ be a real number and f be a function in $L^2_s(\Omega)$ satisfying* (2.21) *and such that*

$$\int_\Omega f(x,z) V(0,m,z) dx dz = \int_\Omega x f(x,z) V(0,m,z) dx dz = 0. \tag{2.23}$$

Then $R^+_A(\mu)f = R^-_A(\mu)f = u$, which is in $NL^2_s(m)'$ by LAP2, *belongs to $L^2_{-\tilde{s}}(\Omega)$ with $\tilde{s} = \mathrm{Max}(0, 2-s)$. Moreover, there exists a constant $C = C(\mu, s)$ such that for every f in $L^2_s(\Omega)$ satisfying* (2.21) *and* (2.23), *we have*

$$\|u\|_{L^2_{-\tilde{s}}(\Omega)} \leq C \|f\|_{L^2_s(\Omega)}. \tag{2.24}$$

Remark 2.1. The space $L^2_{-\tilde{s}}(\Omega)$ is a strict subspace of $NL^2_s(m)'$. Indeed, the real number $\tilde{s} = \mathrm{Max}(0, 2-s)$, $s > 3/2$, is in the interval $[0, 1/2)$. Therefore, the space $W(m) = NL^2_s(m) \cap C^\infty_0(\Omega)$ is dense in the space $L^2_{\tilde{s}}(\Omega)$, and we have the injective maps

$$NL^2_s(m) \subset L^2_{\tilde{s}}(\Omega) \subseteq L^2(\Omega) \subseteq L^2_{-\tilde{s}}(\Omega) \subset NL^2_s(m)'.$$

For $s > 3/2$, according to condition (2.23), we define the space

$$NL^2_s(m,1) = \left\{ f \in NL^2_s(m) \ / \ \int_\Omega x f(x,z) V(0,m,z) dx dz = 0 \right\}. \tag{2.25}$$

We consider this closed hyperplane of $NL^2_s(m)$ in Proposition 3.6 in §3. There we get some properties of nullity and Hölder continuity in $\lambda = \lambda(0,m)$ for the trace functions $\lambda \mapsto f^j(\lambda, m)$, $j = 1$ or 2. They are sufficient to get the following "strong" limiting absorption principle in $L^2_{-s}(\Omega)$.

THEOREM 2.3 (LAP3). *Let $m \geq 1$ be an integer and $\mu = \lambda(0,m)$ be the corresponding threshold of A. We set $J_m = (\lambda(0,m-1), \lambda(0,m+1))$ with $\lambda(0,0) < \lambda(0,1)$. Let $s > 3/2$ be a real number and $NL^2_s(m,1)$ be the Banach space* (2.25) *equipped with the norm of $L^2_s(\Omega)$. Then the following hold:*

(i) *For μ in J_m, the limits in* (2.18) *exist in the uniform operator topology on the space $B(NL^2_s(m,1), L^2_{-s}(\Omega))$.*

(ii) *Each function $\zeta \mapsto R^{\pm}_A(\zeta)$, defined on $J^{\pm}_m = \{\zeta \in \overline{\mathbb{C}^{\pm}} \ / \ \mathrm{Re}\zeta \in J_m\}$ and valued in $B(NL^2_s(m,1), L^2_{-s}(\Omega))$, is locally Hölder continuous with order δ in $[0,1/2]$, $\delta < (s-3/2)/2$. Specifically, there exists a function $C_m(\zeta,\zeta') = C_m(s,\delta,\zeta,\zeta')$, continuous with respect to ζ and ζ' in $J^{\pm}_m$, such that*

(2.26)
$$\forall f \in L^2_s(m,1), \quad \|R^{\pm}_A(\zeta)f - R^{\pm}_A(\zeta')f\|_{L^2_{-s}(\Omega))} \leq C_m(\zeta,\zeta')|\zeta-\zeta'|^{\delta}\|f\|_{L^2_s(\Omega)}.$$

The same properties hold for the functions $\zeta \mapsto \nabla R^{\pm}_A(\zeta)$.

(iii) *Let f be a function in $NL^2_s(m,1)$ and*

$$U^{\pm} = R^{\pm}_A(\mu)f = \lim_{\zeta \to \mu,\ \pm\mathrm{Im}\zeta\ >0} R_A(\zeta)f.$$

With the notation in (2.10) *and* (2.14), *we have*

$$(2.27) \qquad \langle U^{\pm}, g\rangle_{L^2_{-s}(\Omega),\, L^2_s(\Omega)} = \begin{cases} \displaystyle\sum_{n \geq m} \int_{-\infty}^{+\infty} \frac{\tilde{f}(\xi,n)\overline{\tilde{g}(\xi,n)}}{\lambda(\xi,n)-\mu}d\xi \\ \displaystyle\quad + \sum_{n<m} \mathrm{p.v.} \int_{-\infty}^{+\infty} \frac{\tilde{f}(\xi,n)\overline{\tilde{g}(\xi,n)}}{\lambda(\xi,n)-\mu}d\xi \\ \displaystyle\quad \pm i\pi \sum_{n<m}\sum_{j=1}^{2} f^j(\mu,n)\overline{g^j(\mu,n)}. \end{cases}$$

The distributions $\partial_x U^{\pm}$ and $\partial_z U^{\pm}$ are in $L^2_{-s}(\Omega)$. Each function $U^{\pm} = R^{\pm}_A(\mu)f$ is in $D(A)_{\mathrm{loc}}$ and satisfies the differential equation

$$(2.28) \qquad (A - \mu I)U^{\pm} = f \quad in\ \mathcal{D}'(\Omega).$$

Remark 2.2. The restriction of $U^{\pm}$ to $NL^2_s(m)$ is $u^{\pm} = R^{\pm}_A(\mu)f$, the limit form on $NL^2_s(m)$ defined by LAP2. When $s > 3/2$, $NL^2_s(m)$ is not dense in $L^2_s(\Omega)$. Thus the identification of $U^{\pm}$ and $u^{\pm}$ through $R^{\pm}_A(\mu)f$ is not valid. To be more precise, $U^{\pm}$ is the particular continuation of $u^{\pm}$ on $L^2_s(\Omega)$ given by formula (2.27).

Remark 2.3. Theorem 2.3 improves LAP2 since $R^{\pm}_A(\mu)f$ is a form defined on $L^2_s(\Omega)$, a larger space than $NL^2_s(m)$. However, we have to choose larger s and f in an hyperplane of $NL^2_s(m)$. In a future paper, we will examine the question of whether there exist functions f_μ in the space $NL^2_s(m)$ with $s > 1$ such that

$$(2.29) \qquad \lim_{\zeta \to \mu,\ \zeta \notin \sigma(A)} \|R_A(\zeta)f_\mu\|_{L^2_{-s}(\Omega)} = +\infty.$$

Remark 2.4. The Hölder order δ in $[0, 1/2]$, $\delta < (s - 3/2)/2$, obtained in LAP3, has to be compared with the order δ_2 in $[0, \mathrm{Min}(1/4, (s-1)/2))$, obtained in LAP2. In the case where $s > 3/2$, Corollary 3.4 together with Theorem 3.7 (both in §3) allows us to increase the order δ_2 in LAP2 up to δ (see Remark 3.1, also in §3).

3. Basic estimates. For any real number s, $H^s(\mathbb{R})$ denotes the Sobolev space with exponent s, the space of tempered distributions v such that $\mathcal{F}v$ is in $L^2_s(\mathbb{R})$. It is a Hilbert space with the norm $\|v\|_{H^s(\mathbb{R})} = \|\mathcal{F}v\|_{L^2_s(\mathbb{R})}$.

3.1. Agmon's result. We return to a basic result of Agmon. We use it in dimension one.

LEMMA B.2 FROM [A]. *Let $s > 1/2$ be a real number. Let v be a function in $H^s(\mathbb{R})$ such that $v(0) = 0$. Then the function $V(\xi) = (v(\xi)/\xi)$ is in $H^{s-1}(\mathbb{R}) \cap L^1_{\mathrm{loc}}(\mathbb{R})$, and there exists a constant $C = C(s)$, independent of v in $H^s(\mathbb{R})$, such that*

$$\|V\|_{H^{s-1}(\mathbb{R})} \leq C\|v\|_{H^s(\mathbb{R})}. \tag{3.1}$$

Proof. We refer to [A, pp. 209–211]. □

COROLLARY 3.1. *Let $s > 3/2$ be a real number. Let v be a function in $H^s(\mathbb{R})$ such that $v(0) = v'(0) = 0$. Then $V(\xi) = (v(\xi)/\xi^2)$ is a function in $H^{s-2}(\mathbb{R}) \cap L^1_{\mathrm{loc}}(\mathbb{R})$, and there exists a constant $C = C(s)$, independent of v in $H^s(\mathbb{R})$, such that*

$$\|V\|_{H^{s-2}(\mathbb{R})} \leq C\|v\|_{H^s(\mathbb{R})}. \tag{3.2}$$

Proof. Lemma B.2 gives the function $\xi \longmapsto v_1(\xi) = (v(\xi)/\xi)$ in the space $H^{s-1}(\mathbb{R}) \cap L^1_{\mathrm{loc}}(\mathbb{R})$. Since $v_1(0) = \lim_{\xi\to 0} v_1(\xi) = \lim_{\xi\to 0}(v(\xi)/\xi) = v'(0) = 0$, Lemma B.2 can be applied to v_1. This yields the corollary. □

3.2. Accurate properties of the trace operators. First, we give results about the Sobolev regularity, with respect to the variable ξ, of the generalized Fourier coefficients $\widetilde{u}(\xi, n)$ defined by (2.5) for functions u in $L^2(\Omega)$.

PROPOSITION 3.2. *Let $s \geq 0$ be a real number. Let $n \geq 1$ be an integer, M be a real number, and Φ be a function in $C_0^\infty(\mathbb{R})$ with support in the interval $[-M, M]$. Then the continuous map from $L^2(\Omega)$ in $L^2(\mathbb{R})$ defined by $u \mapsto \Phi\widetilde{u}(., n)$ satisfies the inequality*

$$\|\Phi\widetilde{u}(., n)\|_{H^{-s}(\mathbb{R})} \leq C\|u\|_{L^2_{-s}(\Omega)}, \tag{3.3}$$

where the constant $C = C_n(s, \Phi)$ is independent of u in $L^2(\Omega)$. Moreover, if u is in $L^2_s(\Omega)$, we have

$$\Phi\widetilde{u}(., n) \in H^s(\mathbb{R}) \quad \text{and} \quad \|\Phi\widetilde{u}(., n)\|_{H^s(\mathbb{R})} \leq C\|u\|_{L^2_s(\Omega)}. \tag{3.4}$$

The constant $C = C_n(s, \Phi)$ can be chosen continuously dependent on M and on the norms $\|\Phi^{(l)}\|_{L^\infty(\mathbb{R})}, l \leq s + 1$.

Proof. Using the eigenfunctions $V(\xi, n, z)$ introduced in (2.1), we define

$$w(., z) = \mathcal{F}^{-1}(\Phi V(., n, z)). \tag{3.5}$$

Thanks to the C^∞ regularity of $\xi \mapsto V(\xi, n, .)$ from $\mathbb{R}$ in $L^2((0, H))$, for every integer k, the function w satisfies

$$\int_\Omega |\partial_\xi^k(\mathcal{F}w)(\xi, z)|^2 d\xi dz = \int_\Omega \left|\sum_{l=0}^{k} \binom{k}{l} \Phi^{(l)}(\xi)\partial_\xi^{k-l} V(\xi, n, z)\right|^2 d\xi dz \leq c_n(k, \Phi).$$

Therefore, for every real $r \geq 0$, we have

$$\|w\|_{L^2_r(\Omega)} \leq \widetilde{c}_n(r, \Phi) = \text{Max}\{c_n(k, \Phi) \ / \ k \in \mathbb{N},\ 0 \leq k \leq r+1\}. \tag{3.6}$$

Let u be a function in $L^2(\Omega)$. With the definition (2.5) of $\widetilde{u}$ and with the property of the Fourier transform $\mathcal{F}$ with regard to convolution, we have

$$\Phi(\xi)\widetilde{u}(\xi, n) = \int_0^H \mathcal{F}u(\xi, z)\Phi(\xi)V(\xi, n, z)dz = (2\pi)^{-1/2}\mathcal{F}\left(\int_0^H u(., z) * w(., z)dz\right)$$

and

$$\|\Phi\widetilde{u}(., n)\|_{H^{\pm s}(\mathbb{R})} = (2\pi)^{-1/2}\left\|\int_0^H u(., z) * w(., z)dz\right\|_{L^2_{\pm s}(\mathbb{R})}.$$

Let us set $t = \pm s$. From $(1+x^2)^t \leq c(s)(1+(x-X)^2)^t(1+X^2)^s$, the Schwarz inequality, and the Fubini theorem, we deduce the following estimates:

$$\begin{aligned}
&\left\|\int_0^H u(., z) * w(., z)dz\right\|^2_{L^2_t(\mathbb{R})} \\
&= \int_{-\infty}^{+\infty} (1+x^2)^t \left|\int_0^H \int_{-\infty}^{+\infty} u(x-X, z)w(X, z)dXdz\right|^2 dx \\
&\leq c(s)\int_{-\infty}^{+\infty}\left|\int_{-\infty}^{+\infty}\int_0^H (1+(x-X)^2)^{t/2}u(x-X, z)(1+X^2)^{s/2}w(X, z)dzdX\right|^2 dx \\
&\leq c(s)\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty}\left(\int_0^H (1+(x-X)^2)^t|u(x-X, z)|^2dz\right)^{1/2}\right. \\
&\qquad\qquad \left.\left(\int_0^H (1+X^2)^s|w(X, z)|^2dz\right)^{1/2} dX\right]^2 dx \\
&\leq c(s)\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty}\int_0^H (1+X^2)^{-1}(1+(x-X)^2)^t|u(x-X, z)|^2dzdX\right] \\
&\qquad\qquad \left[\int_{-\infty}^{+\infty}\int_0^H (1+X^2)^{s+1}|w(X, z)|^2dzdX\right] dx \\
&\leq c(s)\|w\|^2_{L^2_{s+1}(\Omega)}\int_{\mathbb{R}\times\Omega}(1+X^2)^{-1}(1+(x-X)^2)^t|u(x-X, z)|^2dXdxdz.
\end{aligned}$$

Using (3.6) with $r = s+1$, we obtain the estimate

$$\forall u \in L^2(\Omega), \quad \|\Phi\widetilde{u}(., n)\|_{H^t(\mathbb{R})} \leq C\|u\|_{L^2_t(\Omega)} \tag{3.7}$$

with a constant $C = C_n(t, \Phi) \leq \widetilde{c}(s)\widetilde{c}_n(s+1, \Phi)$. The Φ dependence of the constant C appears in the constant $\widetilde{c}_n(s+1, \Phi)$ of (3.6). The proof is then complete. □

Choosing a larger s, we can improve the description of $\Phi\widetilde{u}(., n)$ with the help of Lemma B.2.

PROPOSITION 3.3. *Let $s > 3/2$ be a real number. Let $n \geq 1$ be an integer, f be a function in $NL^2_s(n)$, and Φ be a function in $C_0^\infty(\mathbb{R})$. Then there exists $\rho = \rho_{\Phi,f}$ in $H^{s-1}(\mathbb{R}) \cap L^1(\mathbb{R})$ such that*

$$\Phi(\xi)\widetilde{f}(\xi,n) = \xi\,\rho(\xi), \quad \xi \in \mathbb{R}. \tag{3.8}$$

Moreover, there exist constants $c(s), C = C_n(s,\Phi)$, $\delta \in [0,1]$ and $\delta < s - 3/2$, and $M = M_n(s,\Phi,\delta)$, independent of f in $NL^2_s(n)$ and of the real numbers ξ and ξ', such that

$$|\rho(\xi)| \leq c(s)\|\rho\|_{H^{s-1}(\mathbb{R})} \leq C\|f\|_{L^2_s(\Omega)}, \tag{3.9}$$

$$|\rho(\xi) - \rho(\xi')| \leq M|\xi - \xi'|^\delta \|f\|_{L^2_s(\Omega)}. \tag{3.10}$$

The constants $C = C_n(s,\Phi)$ and $M = M_n(s,\Phi,\delta)$ can be chosen with the same dependence on Φ as in Proposition 3.2.

Proof. Let us recall (2.12): $NL^2_s(n) = \{f \in L^2_s(\Omega) \ / \ \widetilde{f}(0,n) = 0\}$. Using Proposition 3.2, the function $\xi \longmapsto v(\xi) = \Phi(\xi)\widetilde{f}(\xi,n)$ is in $H^s(\mathbb{R})$ and is compactly supported, and $\|v\|_{H^s(\mathbb{R})} \leq c_n(s,\Phi)\|f\|_{L^2_s(\Omega)}$. Then we apply Lemma B.2 to get $\rho(\xi) = (v(\xi)/\xi)$ in $H^{s-1}(\mathbb{R}) \cap L^1(\mathbb{R})$ and $\|\rho\|_{H^{s-1}(\mathbb{R})} \leq \tilde{c}_n(s,\Phi)\|f\|_{L^2_s(\Omega)}$. Estimates (3.9) and (3.10) are then derived from classical ones for functions in $H^t(\mathbb{R})$. Specifically,

$$\begin{aligned}
|\rho(\xi)| &= (2\pi)^{-1/2}\left|\int_{-\infty}^{+\infty} \mathcal{F}\rho(x)e^{ix\xi}dx\right| \\
&\leq (2\pi)^{-1/2}\left(\int_{-\infty}^{+\infty} (1+x^2)^{-t}dx\right)^{1/2}\left(\int_{-\infty}^{+\infty} |\mathcal{F}\rho(x)|^2(1+x^2)^t dx\right)^{1/2} \quad \text{with } t > 1/2
\end{aligned}$$

so that $|\rho(\xi)| \leq c(t)\|\rho\|_{H^t(\mathbb{R})}$, and

$$\begin{aligned}
|\rho(\xi) - \rho(\xi')| &= (2\pi)^{-1/2}\left|\int_{-\infty}^{+\infty} \mathcal{F}\rho(x)(e^{ix\xi} - e^{ix\xi'})dx\right| \\
&\leq (2\pi)^{-1/2}\int_{-\infty}^{+\infty} |\mathcal{F}\rho(x)|\ 2^{1-\delta}|\xi - \xi'|^\delta|x|^\delta dx \quad \text{with } \delta \in [0,1]
\end{aligned}$$

so that $|\rho(\xi) - \rho(\xi')| \leq c(t,\delta)|\xi - \xi'|^\delta\|\rho\|_{H^t(\mathbb{R})}$ with $\delta < t - 1/2$. □

Precise results follow about the behavior of the trace functions near thresholds.

COROLLARY 3.4. *Let $s > 3/2$ be a real number, $n \geq 1$ be an integer, and $j = 1$ or 2. Let f be a function in $NL^2_s(n)$. Then there exist locally Hölder continuous functions $\widetilde{F}(.,n) = \widetilde{F}_f(.,n)$ defined on $I\!R$ and $F^j(.,n) = F^j_f(.,n)$ defined on $I_n = [\lambda(0,n), +\infty)$ such that*

$$\widetilde{f}(\xi,n) = \xi\widetilde{F}(\xi,n) \quad \textit{and} \quad f^j(\lambda,n) = (\lambda - \lambda(0,n))^{1/4}\, F^j(\lambda,n). \tag{3.11}$$

Specifically, for δ in $[0,1]$, $\delta < s - 3/2$, there exist a function $\widetilde{M}(\xi,\xi') = \widetilde{M}_n(s,\delta,\xi,\xi')$, continuous with respect to ξ and ξ', and a function $M(\lambda,\lambda') = M_n(s,\delta,\lambda,\lambda')$, continuous with respect to $\lambda, \lambda' \geq \lambda(0,n)$, such that

$$\forall f \in NL^2_s(n), \quad |\widetilde{F}(\xi,n) - \widetilde{F}(\xi',n)| \leq \widetilde{M}(\xi,\xi')|\xi - \xi'|^\delta\|f\|_{L^2_s(\Omega)}. \tag{3.12a}$$

$$\forall f \in NL^2_s(n), \quad |F^j(\lambda,n) - F^j(\lambda',n)| \leq M(\lambda,\lambda')|\lambda - \lambda'|^{\delta/2}\|f\|_{L^2_s(\Omega)}. \tag{3.12b}$$

Proof. Note that we already have the Hölder continuity (2.11) for $\widetilde{f}(.,n)$. From Proposition 3.3, we deduce $\widetilde{f}(\xi,n) = \xi\widetilde{F}(\xi,n)$ and (3.12a) with δ in $[0,1]$, $\delta < s-3/2$. Let $\lambda, \lambda' \geq \lambda(0,n)$ and $\xi, \xi' \geq 0$ be connected by $\xi = \xi(\lambda,n)$ and $\xi' = \xi(\lambda',n)$. From (2.6), (2.17c), and (2.17b) and setting $\varepsilon_j = (-1)^j$, we get

$$\begin{aligned} f^j(\lambda,n) &= (\partial_\lambda\xi(\lambda,n))^{1/2}\,\widetilde{f}(\varepsilon_j\xi,n) = (\lambda-\lambda(0,n))^{-1/2}H_n(\xi)^{1/2}\varepsilon_j\xi\,\widetilde{F}(\varepsilon_j\xi,n) \\ &= (\lambda-\lambda(0,n))^{-1/4}H_n(\xi)^{1/2}\,\varepsilon_j(\lambda-\lambda(0,n))^{1/2}G_n(\xi)\widetilde{F}(\varepsilon_j\xi,n). \end{aligned}$$

Thus $f^j(\lambda,n) = (\lambda-\lambda(0,n))^{1/4}F^j(\lambda,n)$ with

$$F^j(\lambda,n) = (-1)^j H_n(\xi)^{1/2}G_n(\xi)\,\widetilde{F}((-1)^j\xi,n). \tag{3.13}$$

The analyticity and strict positivity of H_n and G_n and the Hölder continuity (3.12a) of $\widetilde{F}(.,n)$ imply

$$|F^j(\lambda,n) - F^j(\lambda',n)| \leq M(\xi,\xi')|\xi-\xi'|^\delta \|f\|_{L^2_s(\Omega)}.$$

Again with (2.17b), we get (3.12b). □

Propositions 3.2 and 3.3 imply particular differentiability properties for the functions $\widetilde{f}(.,n) = \widetilde{\tau}_n(.)f$. Such properties are made explicit in the next proposition.

PROPOSITION 3.5 (derivatives of the trace operators and the spaces $NL^2_s(m,k)$). *Let $s > 1/2$ be a real number and $n \geq 1$ and $k \geq 0$ be integers with $k < s-1/2$. Then the function $\xi \longrightarrow \widetilde{\tau}_n(\xi)$ defined on $\mathbb{R}$ and valued in $L^2_s(\Omega)'$ has a derivative $(d^k\widetilde{\tau}_n/d\xi^k)$. There exists a function $C(\xi) = C_n(s,\xi)$, continuous with respect to ξ, such that*

$$\forall f \in L^2_s(\Omega), \quad \left|\frac{d^k\widetilde{\tau}_n}{d\xi^k}(\xi)f\right| \leq C(\xi)\|f\|_{L^2_s(\Omega)}. \tag{3.14}$$

This derivative is locally Hölder continuous with order δ in $[0,1]$, $\delta < s-1/2-k$. Specifically, there exists a function $M(\xi,\xi') = M_n(s,\delta,\xi,\xi')$, continuous with respect to ξ and ξ', such that

$$\forall f \in L^2_s(\Omega), \quad \left|\frac{d^k\widetilde{\tau}_n}{d\xi^k}(\xi)f - \frac{d^k\widetilde{\tau}_n}{d\xi^k}(\xi')f\right| \leq M(\xi,\xi')|\xi-\xi'|^\delta\|f\|_{L^2_s(\Omega)}. \tag{3.15}$$

The space

$$NL^2_s(n,k) = \left\{f \in L^2_s(\Omega)\ /\ \frac{d^l\widetilde{\tau}_n}{d\xi^l}(0)f = 0,\ 0 \leq l \leq k\right\} \tag{3.16}$$

is a closed subspace of $L^2_s(\Omega)$ with codimension $k+1$. In particular, we have

$$NL^2_s(n,0) = NL^2_s(n) = \left\{f \in L^2_s(\Omega)\ /\ \int_\Omega f(x,z)V(0,n,z)dxdz = 0\right\}. \tag{3.17}$$

Proof. The existence and continuity of the trace operator $\xi \longmapsto \widetilde{\tau}_n(\xi)$ and estimate (3.15) with $s > 1/2$ and $k = 0$ have been proved in Proposition 3.2 of [CD95a]. Let us now assume that $s > 3/2$. Let k be an integer such that $0 < k < s-3/2$. We consider a function f in $L^2_s(\Omega)$ and its Fourier transform $\mathcal{F}f$ with respect to the variable x. Since $k < s-1/2$, the function $\mathcal{F}f(.,z)$ is in the space $C^k(\mathbb{R})$ for almost every z in

$(0, H)$. For any integer l in $[0, k]$, we have

$$\begin{aligned}|\partial_\xi^l(\mathcal{F}f)(\xi, z)| &= (2\pi)^{-1/2}\left|\int_{-\infty}^{+\infty} x^l e^{-i\xi x} f(x, z)dx\right| \\ &\leq (2\pi)^{-1/2}\left(\int_{-\infty}^{+\infty}(1+x^2)^{-s+l}dx\right)^{1/2}\left(\int_{-\infty}^{+\infty}|f(x,z)|^2(1+x^2)^s dx\right)^{1/2} \\ &\leq C(s,k)\|f\|_{L^2_s(\Omega)}.\end{aligned}$$

The differentiability of the trace function $\xi \mapsto \widetilde{f}(\xi, n) = (\mathcal{F}f(\xi, .)\ /\ V(\xi, n, .))_{L^2((0,H))}$ follows from the differentiability of $\xi \mapsto V(\xi, n, .)$. Therefore, the trace map $\widetilde{\tau}_n$ has a derivative of order k such that

$$\frac{d^k\widetilde{\tau}_n}{d\xi^k}(\xi)f = \sum_{l=0}^{k}\binom{k}{l}\int_0^H \partial_\xi^l(\mathcal{F}f)(\xi, z)\partial_\xi^{k-l}V(\xi, n, z)dz. \tag{3.18}$$

As in Proposition 3.2 of [CD95a], estimates (3.14) and (3.15) are deduced from

$$\|\partial_\xi^l V(\xi, n, .) - \partial_\xi^l V(\xi', n, .)\|_{L^2((0,H))} \leq C_n(l, \xi, \xi')|\xi - \xi'|$$

and from

$$|e^{-i\xi x} - e^{-i\xi' x}| \leq 2^{1-\delta}|\xi - \xi'|^\delta |x|^\delta.$$

Now the forms $(d^l\widetilde{\tau}_n/d\xi^l)(0)$, $0 \leq l \leq k$, are continuous on $NL_s^2(m)$, and the properties of the spaces $NL_s^2(m, k)$ are clear. □

When f is a function in $NL_s^2(n)$, $s > 3/2$, we can determine Hölder continuous functions $\widetilde{F}_f(., n)$ and $F_f^j(., n)$, $j = 1$ or 2, through $\widetilde{f}(., n)$ and $f^j(., n)$ with the help of (3.11). When f is in $NL_s^2(n, 1)$, these functions have supplementary properties of nullity.

Proposition 3.6 (the space $NL_s^2(n, 1)$). *Let $s > 3/2$ be a real number and $n \geq 1$ be an integer. Then the space $NL_s^2(n, 1)$ defined by* (3.16) *is a closed hyperplane of $NL_s^2(n)$ such that*

$$\begin{aligned}NL_s^2(n, 1) &= \{f \in NL_s^2(n)\ /\ \widetilde{F}_f(0, n) = 0\} \\ &= \{f \in NL_s^2(n)\ /\ F_f^j(0, n) = 0,\ j = 1, 2\} \\ &= \left\{f \in NL_s^2(n)\ /\ \int_\Omega x f(x, z)V(0, n, z)dxdz = 0\right\}.\end{aligned} \tag{3.19}$$

Moreover, for δ in $[0, 1/2]$, $\delta < (s-3/2)/2$, there exists a function $C(\lambda) = C_n(s, \delta, \lambda)$, continuous with respect to $\lambda \geq \lambda(0, n)$, such that

$$\forall f \in NL_s^2(n, 1), \quad |f^j(\lambda, n)| \leq C(\lambda)(\lambda - \lambda(0, n))^{1/4+\delta}\|f\|_{L_s^2(\Omega)}. \tag{3.20}$$

Proof. When f is in $NL_s^2(n)$, $s > 3/2$, we have $\widetilde{f}(0, n) = 0$, and $\widetilde{F}_f(\xi, n) = \widetilde{f}(\xi, n)/\xi$ is locally Hölder continuous with respect to ξ. Therefore, $\widetilde{F}_f(0, n) = \lim_{\xi\to 0}\widetilde{f}(\xi, n)/\xi = (d\widetilde{\tau}_n/d\xi)(0)f$. Since $\xi \mapsto V(\xi, n, .)$ is even, $\partial_\xi V(0, n, .) = 0$. From this and from (3.18) with $k = 1$, we derive

$$\frac{d\widetilde{\tau}_n}{d\xi}(0)f = \int_0^H \partial_\xi(\mathcal{F}f)(0, z)V(0, n, z)dz = -i(2\pi)^{-1/2}\int_\Omega x f(x, z)V(0, n, z)dxdz.$$

The description (3.19) of $NL_s^2(n, 1)$ follows. Estimate (3.12b) yields (3.20). □

3.3. Investigations of the terms of the resolvent. Let μ be a real number in $\sigma(A) = [\lambda(0,1), +\infty)$. We fix $\lambda(0,0) = 0$. Let $m \geq 1$ be the integer such that μ is in the interval $(\lambda(0, m-1), \lambda(0,m)]$ and let $n \geq 1$ be an integer. The following quantities B_n and r_n are defined in [CD95a]:

If $n \geq 1$ and $\zeta \notin \sigma(A)$,

$$B_n(\zeta, f, \varphi) = \int_{-\infty}^{+\infty} \frac{\widetilde{f}(\xi,n)\,\overline{\widetilde{\varphi}(\xi,n)}}{\lambda(\xi,n)-\zeta} d\xi = \sum_{j=1}^{2} \int_{\lambda(0,n)}^{+\infty} \frac{f^j(\lambda,n)\,\overline{\varphi^j(\lambda,n)}}{\lambda-\zeta} d\lambda \tag{3.21a}$$

$$\text{with } f \text{ and } \varphi \text{ in } L^2(\Omega). \tag{3.21b}$$

If $n \geq m$ (which is equivalent to $\lambda(0,n) \geq \mu$),

$$B_n(\mu, f, \varphi) = \int_{-\infty}^{+\infty} \frac{\widetilde{f}(\xi,n)\,\overline{\widetilde{\varphi}(\xi,n)}}{\lambda(\xi,n)-\mu} d\xi = \sum_{j=1}^{2} \int_{\lambda(0,n)}^{+\infty} \frac{f^j(\lambda,n)\,\overline{\varphi^j(\lambda,n)}}{\lambda-\mu} d\lambda \tag{3.22a}$$

$$\text{with } f \text{ and } \varphi \text{ in } L^2(\Omega) \quad \text{if } n > m \text{ (and } \lambda(0,n) > \mu) \tag{3.22b}$$

$$\text{with } f \text{ and } \varphi \text{ in } NL^2_s(m), \quad s > 1, \quad \text{if } n = m \text{ (and } \mu = \lambda(0,m)). \tag{3.22c}$$

If $n < m$ (which is equivalent to $\lambda(0,n) < \mu$),

(3.23a)

$$B_n(\mu, f, \varphi) = \text{p.v.} \int_{-\infty}^{+\infty} \frac{\widetilde{f}(\xi,n)\,\overline{\widetilde{\varphi}(\xi,n)}}{\lambda(\xi,n)-\mu} d\xi = \sum_{j=1}^{2} \text{p.v.} \int_{\lambda(0,n)}^{+\infty} \frac{f^j(\lambda,n)\,\overline{\varphi^j(\lambda,n)}}{\lambda-\mu} d\lambda,$$

$$r_n(\mu, f, \varphi) = \sum_{j=1}^{2} f^j(\mu,n)\,\overline{\varphi^j(\mu,n)} \tag{3.23b}$$

$$\text{with } f \text{ and } \varphi \text{ in } L^2_s(\Omega), \quad s > 1/2. \tag{3.23c}$$

The existence of the following limits $B_n^{\pm}$ is used in [CD95a]:

If $n \geq m$ and if both f and φ satisfy (3.22b) in the case where $n > m$ or (3.22c) in the case where $n = m$,

$$B_n^+(\mu, f, \varphi) = B_n^-(\mu, f, \varphi) = \lim_{\zeta \to \mu,\ \pm \mathrm{Im}\zeta > 0} B_n(\zeta, f, \varphi) = B_n(\mu, f, \varphi). \tag{3.24}$$

If $n < m$ and if both f and φ satisfy (3.23c),

$$B_n^{\pm}(\mu, f, \varphi) = \lim_{\zeta \to \mu,\ \pm \mathrm{Im}\zeta > 0} B_n(\zeta, f, \varphi) = B_n(\mu, f, \varphi) \ \pm\ i\,\pi\, r_n(\mu, f, \varphi). \tag{3.25}$$

For $\zeta \in \overline{\mathbb{C}^{\pm}} = \{\zeta \in \mathbb{C} \ / \ \pm \mathrm{Im}\zeta \geq 0\}$, $\zeta \notin \sigma(A)$, with f and $\varphi \in L^2(\Omega)$, we set

$$B_n^{\pm}(\zeta, f, \varphi) = B_n(\zeta, f, \varphi). \tag{3.26}$$

The limiting absorption principle and the properties of the resolvent on $\overline{\mathbb{C}^{\pm}}$ are based on a detailed study of the functions $\zeta \mapsto B_n^{\pm}(\zeta, n, \varphi)$ near the spectrum $\sigma(A)$. This study is presented in Propositions 3.6 and 3.7 of [CD95a] and consists of proving particular properties for the Hilbert transform of a Hölder continuous function. The main result—namely, the Hölder continuity of the Hilbert transform—is also proved by Muskhelishvili [Mu] in a theorem on the behavior of a Cauchy integral near the boundary. We now give these properties in the form which we use for the proof of Theorem 2.3.

THEOREM 3.7. *Let a, b $(a < b)$, and $\delta \in [0,1)$ be three real numbers. Let h be a function that is Hölder continuous on $[a,b]$ with order δ. We suppose that $h(a) = h(b) = 0$ and that there exists a constant $A(h)$ such that*

$$\forall \lambda, \lambda' \in [a,b], \quad |h(\lambda) - h(\lambda')| \le A(h)|\lambda - \lambda'|^{\delta}. \tag{3.27}$$

For ζ in $\mathbb{C}^{\pm} = \{\zeta \in \mathbb{C} \ / \ \pm \mathrm{Im}\zeta > 0\}$, we set $\mathcal{H}h^{\pm}(\zeta) = \int_a^b (h(\lambda)/(\lambda - \zeta))d\lambda$. Then the following hold:

(i) *For a real number μ in $[a,b]$, the following limits exist:*

$$\text{if } h(\mu) \neq 0, \quad \mathcal{H}h^{\pm}(\mu) = \lim_{\zeta \to \mu, \ \pm\mathrm{Im}\zeta > 0} \mathcal{H}h^{\pm}(\zeta) = \text{p.v.} \int_a^b \frac{h(\lambda)}{\lambda - \mu} d\lambda \pm i\pi h(\mu); \tag{3.28}$$

$$\text{if } h(\mu) = 0, \quad \mathcal{H}h^{+}(\mu) = \mathcal{H}h^{-}(\mu) = \lim_{\zeta \to \mu, \ \pm\mathrm{Im}\zeta > 0} \mathcal{H}h^{\pm}(\zeta) = \int_a^b \frac{h(\lambda)}{\lambda - \mu} d\lambda. \tag{3.29}$$

(ii) *The extended functions $\mathcal{H}h^{\pm}$ on $V^{\pm} = \{\zeta \in \overline{\mathbb{C}^{\pm}} \ / \ a \le \mathrm{Re}\zeta \le b\}$, are Hölder continuous on $V^{\pm}$ with order δ. Specifically, there exist functions $C(\zeta, \zeta') = C(a, b, \delta, \zeta, \zeta')$ and $D(\zeta) = D(a, b, \delta, \zeta)$, continuous with respect to ζ and ζ' and independent of h, such that for ζ and ζ' in $V^{\pm}$, we have*

$$|\mathcal{H}h^{\pm}(\zeta) - \mathcal{H}h^{\pm}(\zeta')| \le A(h)C(\zeta, \zeta')|\zeta - \zeta'|^{\delta}, \tag{3.30}$$

$$|\mathcal{H}h^{\pm}(\zeta)| \le A(h)D(\zeta). \tag{3.31}$$

Proof. We refer to [Mu, Chap. 2, §22] (see also [G, Chap. I, §5]). Note that the use of the maximum-modulus theorem allows us to increase the Hölder order for $\mathcal{H}h^{\pm}$ up to $\delta < 1$ and not only to any δ', $\delta' < \delta$. An attentive reading of the proof shows the precise dependence on h in (3.30) and (3.31), as in our proofs of Propositions 3.6 or 3.7 of [CD95a]. □

The next two propositions collect estimates that can be found more or less explicitly in §3 of [CD95a].

PROPOSITION 3.8 (Hölder continuity and bounds for $B_n^{\pm}$ and r_n outside of $\Gamma(A)$). *Let $m \ge 2$ be an integer, I_m be the interval $(\lambda(0, m-1), \lambda(0, m))$, and $I_m^{\pm} = \{\zeta \in \overline{\mathbb{C}^{\pm}} \ / \ \mathrm{Re}\zeta \in I_m\}$. Let s and δ be real numbers such that*

$$s > 1/2 \quad \text{and} \quad \delta \in [0, \mathrm{Min}(1, s - 1/2)). \tag{3.32}$$

Then there exist functions $C(\zeta, \zeta') = C_m(s, \delta, \zeta, \zeta')$ and $D(\zeta) = D_m(s, \delta, \zeta)$, continuous with respect to ζ and ζ', such that for any functions f and φ in $L_s^2(\Omega)$,

(i) *for any complex numbers* ζ *and* ζ' *in* $I_m^{\pm}$, *we have*

$$\sum_{n\geq m} |B_n^{\pm}(\zeta,f,\varphi) - B_n^{\pm}(\zeta',f,\varphi)| \leq C(\zeta,\zeta')|\zeta-\zeta'| \, \|f\|_{L^2(\Omega)} \, \|\varphi\|_{L^2(\Omega)}, \tag{3.33}$$

$$\sum_{n\geq m} |B_n^{\pm}(\zeta,f,\varphi)| \leq D(\zeta)\|f\|_{L^2(\Omega)}\|\varphi\|_{L^2(\Omega)}, \tag{3.34}$$

$$\sum_{1\leq n<m} |B_n^{\pm}(\zeta,f,\varphi) - B_n^{\pm}(\zeta',f,\varphi)| \leq C(\zeta,\zeta')|\zeta-\zeta'|^{\delta}\|f\|_{L^2_s(\Omega)}\|\varphi\|_{L^2_s(\Omega)}, \tag{3.35}$$

$$\sum_{1\leq n<m} |B_n^{\pm}(\zeta,f,\varphi)| \leq D(\zeta)\|f\|_{L^2_s(\Omega)}\|\varphi\|_{L^2_s(\Omega)}; \tag{3.36}$$

(ii) *for any real numbers* λ *and* λ' *in* I_m, *we have*

$$\sum_{1\leq n<m} |r_n(\lambda,f,\varphi) - r_n(\lambda',f,\varphi)| \leq C(\lambda,\lambda')|\lambda-\lambda'|^{\delta}\|f\|_{L^2_s(\Omega)}\|\varphi\|_{L^2_s(\Omega)}, \tag{3.37}$$

$$\sum_{1\leq n<m} |r_n(\lambda,f,\varphi)| \leq D(\lambda)\|f\|_{L^2_s(\Omega)}\|\varphi\|_{L^2_s(\Omega)}. \tag{3.38}$$

Moreover, in (3.37), *the order* δ *can go up to* 1 *if* $s > 3/2$.

PROPOSITION 3.9 (Hölder continuity and bounds for $B_n^{\pm}$ and r_n near $\Gamma(A)$). *Let* $m \geq 1$ *be an integer and* $\lambda(0,m)$ *be the corresponding threshold of* A. *Let* J_m *be the interval* $(\lambda(0,m-1),\lambda(0,m+1))$ *and* $J_m^{\pm} = \{\zeta \in \overline{\mathbb{C}^{\pm}} \ / \ \mathrm{Re}\zeta \in J_m\}$.

(i) *Let* s *and* δ *be real numbers satisfying* (3.32). *Then there exist functions* $C(\zeta,\zeta') = C_m(s,\delta,\zeta,\zeta')$ *and* $D(\zeta) = D_m(s,\delta,\zeta)$, *continuous with respect to* ζ,ζ', *such that for any functions* f *and* φ *in* $L^2_s(\Omega)$ *and for any complex numbers* ζ *and* ζ' *in* $J_m^{\pm}$, *we have on the one hand*

$$\sum_{n>m} |B_n^{\pm}(\zeta,f,\varphi) - B_n^{\pm}(\zeta',f,\varphi)| \leq C(\zeta,\zeta')|\zeta-\zeta'| \, \|f\|_{L^2(\Omega)} \, \|\varphi\|_{L^2(\Omega)}, \tag{3.39}$$

$$\sum_{n>m} |B_n^{\pm}(\zeta,f,\varphi)| \leq D(\zeta)\|f\|_{L^2(\Omega)}\|\varphi\|_{L^2(\Omega)}, \tag{3.40}$$

and on the other hand, for $1 \leq n < m$, *estimates* (3.35) *and* (3.36) *and estimates* (3.37) *and* (3.38) *for any real numbers* λ *and* λ' *in* J_m.

(ii) *Let* s *and* δ *be real numbers such that*

$$s > 1 \quad \text{and} \quad \delta \in [0, \mathrm{Min}(1/4,(s-1)/2)). \tag{3.41}$$

Then there exist functions $C(\zeta,\zeta') = C_m(s,\delta,\zeta,\zeta')$ *and* $D(\zeta) = D_m(s,\delta,\zeta)$, *continuous with respect to* ζ *and* ζ', *such that for any complex numbers* ζ *and* ζ' *in* $J_m^{\pm}$ *and for any functions* f *and* φ *in* $NL^2_s(m)$, *one has*

$$|B_m^{\pm}(\zeta,f,\varphi) - B_m^{\pm}(\zeta',f,\varphi)| \leq C(\zeta,\zeta')|\zeta-\zeta'|^{\delta}\|f\|_{L^2_s(\Omega)}\|\varphi\|_{L^2_s(\Omega)}, \tag{3.42}$$

$$|B_m^{\pm}(\zeta,f,\varphi)| \leq D(\zeta)\|f\|_{L^2_s(\Omega)}\|\varphi\|_{L^2_s(\Omega)}. \tag{3.43}$$

Remark 3.1. If $s > 3/2$, we prove in Theorem 5.2 in §5 that the Hölder order in point (ii) can be improved up to δ in $[0,1/2]$, $\delta < (s-3/2)/2$.

We now use Agmon's result from §3.1.

PROPOSITION 3.10 (a new estimate for B_n). *Let $m \geq 2$ and $n \geq 1$ be integers such that $n < m$. Let μ be a real number in $\sigma(A)$ such that $\lambda(0, m-1) < \mu \leq \lambda(0,m)$. Let $s > 1/2$ be a real number. Then for any function f such that*

$$f \in L^2_s(\Omega), \qquad \tilde{f}(\pm\xi(\mu,n),n) = 0 \tag{3.44}$$

and for any function $\varphi \in L^2_s(\Omega)$, the principal-value integral (3.23a) *is an ordinary integral written as*

$$B_n(\mu,f,\varphi) = \int_{-\infty}^{+\infty} \frac{\tilde{f}(\xi,n)\,\overline{\tilde{\varphi}(\xi,n)}}{\lambda(\xi,n)-\mu}d\xi. \tag{3.45}$$

With a constant $C = C(\mu,s)$ independent of f and φ, it satisfies the estimate

$$|B_n(\mu,f,\varphi)| \leq C\|f\|_{L^2_s(\Omega)}\|\varphi\|_{L^2_{\tilde{s}}(\Omega)} \quad \textit{with } \tilde{s} = \mathrm{Max}(0, 1-s). \tag{3.46}$$

Moreover, if $s \geq 1$, for any function f satisfying (3.44) *and for any function φ in $L^2(\Omega)$, the integral* (3.45) *is still defined and the estimate* (3.46) *holds with $\tilde{s} = 0$.*

Proof. When $n < m$, the equation $\lambda(\xi,n) - \mu = 0$ has two roots $\xi = \pm\xi(\mu,n)$ which are of order 1. We fix a real number $\delta = \delta(\mu,n) > 0$ such that $\pm\xi(\mu,n)$ is the only root for $\lambda(\xi,n)-\mu = 0$ in the interval $J_n^\pm = (\pm\xi(\mu,n)-2\delta,\ \pm\xi(\mu,n)+2\delta)$. Then we choose a function $\Phi = \Phi_{n,\mu}$ in $C_0^\infty(\mathbb{R})$ equal to 1 on $(\pm\xi(\mu,n)-\delta,\ \pm\xi(\mu,n)+\delta)$ with support in $J_n^\pm$ and valued in $[0,1]$. In order to study the integral in (3.23a), we rewrite it in the form $B_n(\mu,f,\varphi) = b_n(\mu,f,\varphi) + s_n(\mu,f,\varphi)$ with

$$b_n(\mu,f,\varphi) = \int_{-\infty}^{+\infty} \frac{(1-\Phi(\xi)^2)\tilde{f}(\xi,n)\,\overline{\tilde{\varphi}(\xi,n)}}{\lambda(\xi,n)-\mu}d\xi, \tag{3.47}$$

$$s_n(\mu,f,\varphi) = \int_{-\infty}^{+\infty} \frac{\Phi(\xi)\tilde{f}(\xi,n)\,\overline{\Phi(\xi)\tilde{\varphi}(\xi,n)}}{\lambda(\xi,n)-\mu}d\xi. \tag{3.48}$$

The integral in the right-hand side of (3.47) is well defined when f and φ are in $L^2(\Omega)$. From (2.9), we get a bound for b_n with a constant $C = C(\mu)$,

$$|b_n(\mu,f,\varphi)| \leq C\int_{-\infty}^{+\infty} |\tilde{f}(\xi,n)\overline{\tilde{\varphi}(\xi,n)}|d\xi \leq C\|f\|_{L^2(\Omega)}\|\varphi\|_{L^2(\Omega)}. \tag{3.49}$$

Let us now consider the term $s_n(\mu,f,\varphi)$. Using Proposition 3.2, we have the function $\Phi\tilde{f}(.,n)$ in $H^s(\mathbb{R})$ and the estimate $\|\Phi\tilde{f}(.,n)\|_{H^s(\mathbb{R})} \leq C(s,n,\Phi)\ \|f\|_{L^2_s(\mathbb{R})}$. Moreover, since $\tilde{f}(\pm\xi(\mu,n),n) = 0$, with the help of Lemma B.2, we get the function $V = (\Phi\tilde{f}(.,n))/(\lambda(.,n)-\mu)$ in $H^{s-1}(\mathbb{R}) \cap L^1(\mathbb{R})$ and the estimate

$$\|V\|_{H^{s-1}(\mathbb{R})} \leq C(s,n)\|\Phi\tilde{f}(.,n)\|_{H^s(\mathbb{R})} \leq C(s,n,\Phi)\|f\|_{L^2_s(\Omega)}. \tag{3.50}$$

When φ is in $L^2_s(\Omega)$, $s > 1/2$, the function $\tilde{\varphi}(.,n)$ is continuous. From this and from our conditions on s, f, and φ, it follows that the function $V\overline{\Phi\tilde{\varphi}(.,n)}$ is in $L^1(\mathbb{R})$. Then the integral in the right-hand side of (3.48) is well defined, and (3.45) is valid.

Moreover, as φ belongs to $L^2_{\tilde{s}}(\Omega) \subseteq L^2_{1-s}(\Omega)$, again from Proposition 3.2, it follows that $\Phi\widetilde{\varphi}(.,n)$ in $H^{1-s}(\mathbb{R})$ and

$$\|\Phi\widetilde{\varphi}(.,n)\|_{H^{1-s}(\mathbb{R})} \leq C(s,n,\Phi)\ \|\varphi\|_{L^2_{1-s}(\Omega)}. \tag{3.51}$$

When φ is in $\mathcal{S}(\mathbb{R})$, the integral in (3.48) can be written as the duality brackets

$$s_n(\mu,f,\varphi) = \langle V, \overline{\Phi\widetilde{\varphi}(.,n)}\rangle_{\mathcal{S}'(\mathbb{R}),\,\mathcal{S}(\mathbb{R})} = \langle \mathcal{F}V, \mathcal{F}^{-1}(\overline{\Phi\widetilde{\varphi}(.,n)})\rangle_{L^2_{s-1}(\Omega),\ L^2_{1-s}(\Omega)}.$$

By a density argument, for any φ in $L^2_{\tilde{s}}(\Omega)$ if $1/2 < s < 1$ and $t = 1-s$ or for any φ in $L^2(\Omega)$ if $s \geq 1$ and $t = 0$, we get

$$|s_n(\mu,f,\varphi)| \ \leq\ C\|f\|_{L^2_s(\Omega)}\|\varphi\|_{L^2_t(\Omega)} \tag{3.52}$$

with $C = C(\mu,s)$. Estimate (3.46) follows from (3.49) and (3.52). □

PROPOSITION 3.11 (a new estimate for B_m at the threshold $\mu = \lambda(0,m)$). *Let $m \geq 1$ be an integer, $\mu = \lambda(0,m)$ be the corresponding threshold of A, and $s > 3/2$ be a real number. Then for any function f in $NL^2_s(m,1)$, the integral*

$$I_m(\mu,f,\varphi) = \int_{-\infty}^{+\infty} \frac{\widetilde{f}(\xi,m)\ \overline{\widetilde{\varphi}(\xi,m)}}{\lambda(\xi,m)-\mu}d\xi \tag{3.53}$$

is defined for any function φ in $L^2_{\tilde{s}}(\Omega)$ or, if $s \geq 2$, for any function φ in $L^2(\Omega)$. It satisfies the estimate

$$|I_m(\mu,f,\varphi)| \leq C\|f\|_{L^2_s(\Omega)}\|\varphi\|_{L^2_{\tilde{s}}(\Omega)} \quad \textit{with } \tilde{s} = \operatorname{Max}(0, 2-s) \tag{3.54}$$

with a constant $C = C(\mu,s)$ independent of f and φ. Moreover, if φ is a function in $NL^2_s(m)$, we have $I_m(\mu,f,\varphi) = B_m(\mu,f,\varphi)$.

Proof. We proceed in the same way as in the proof of Proposition 3.10. We have (2.4), that is, $\lambda(\xi,m) - \mu = \xi^2 a_m(\xi)$ with $a_m(\xi)$ an analytical function on $\mathbb{R}$ which does not vanish. With the help of a function $\Phi = \Phi_\mu$ in $C_0^\infty(\mathbb{R})$ equal to 1 in a neighborhood of $\xi = 0$ and valued in $[0,1]$, we again represent the integral in the form of the sum

$$B_m(\mu,f,\varphi) = b_m(\mu,f,\varphi) + s_m(\mu,f,\varphi) \tag{3.55}$$

with b_m and s_m defined by (3.47) and (3.48). Estimate (3.49) is still valid for b_m. Since $s > 3/2$ and f is in $NL^2_s(m,1)$, according to the definition of $NL^2_s(m,1)$ in (3.16), the function $\widetilde{f}(.,m)$ and its first derivative vanish at $\xi = 0$. With the help of Proposition 3.2 and Corollary 3.1, we get $\Phi\widetilde{f}(.,m)$ in $H^s(\mathbb{R})$, $V = (\Phi\widetilde{f}(.,m))/(\xi^2 a_m(\xi))$ in $H^{s-2}(\mathbb{R}) \cap L^1(\mathbb{R})$, and the estimate

$$\|V\|_{H^{s-2}(\mathbb{R})} \leq C(s,m)\|\Phi\widetilde{f}(.,m)\|_{H^s(\mathbb{R})} \leq C(s,m,\Phi)\|f\|_{L^2_s(\Omega)}.$$

The end of the proof is similar to the end of the proof of Proposition 3.10. □

4. Proofs of the division theorems. These theorems concern the operators $R^{\pm}_A(\mu)$ defined by LAP1 or LAP2. We refer to Proposition 3.8 of [CD95a] for their explicit calculations. We recall that they are obtained by studying the limits of

$$(R^{\pm}_A(\zeta)f \mid \varphi)_{L^2(\Omega)} \ =\ ((A-\zeta I)^{-1} \mid \varphi)_{L^2(\Omega)} \ =\ \sum_{n\geq 1} B^{\pm}_n(\zeta,f,\varphi) \tag{4.1}$$

when $\zeta \in \mathbb{C}^{\pm} = \{\zeta \in \mathbb{C} \;/\; \pm \mathrm{Im}\zeta > 0\}$, $\zeta \notin \sigma(A)$, and ζ tends to $\mu \in \sigma(A)$. Here the quantities $B_n^{\pm}$ are given by (3.26) and (3.21).

Let μ be a real number in $\sigma(A) = [\lambda(0,1), +\infty)$. We set $\lambda(0,0) = 0$. Let $m \geq 1$ be the integer such that μ is in the interval $(\lambda(0,m-1), \lambda(0,m)]$. Let s be a real number and $E_s(\mu)$ be the space of functions defined on Ω such that $s > 1/2$ and $E_s(\mu) = L^2_s(\Omega)$ if $\mu \neq \lambda(0,m)$, $s > 1$, and $E_s(\mu) = NL^2_s(m)$ if $\mu = \lambda(0,m)$. With such a μ, m, and s and with f and φ in $E_s(\mu)$, we have

$$\begin{aligned} \langle R_A^{\pm}(\mu)f, \overline{\varphi}\rangle_{E_s(\mu)',\, E_s(\mu)} &= \lim_{\zeta \to \mu,\ \pm\mathrm{Im}\zeta > 0} (R_A^{\pm}(\zeta)f \mid \varphi)_{L^2(\Omega)} \\ &= \sum_{n \geq 1} B_n(\mu, f, \varphi) \;\pm\; i\pi \sum_{n<m} r_n(\mu, f, \varphi), \end{aligned} \tag{4.2}$$

where the quantities B_n and r_n are given by (3.22a)–(3.23b). When the zero-trace conditions (2.21) are fulfilled, we have the simpler formula

$$\langle R_A^{+}(\mu)f, \overline{\varphi}\rangle_{E_s(\mu)',E_s(\mu)} = \langle R_A^{-}(\mu)f, \overline{\varphi}\rangle_{E_s(\mu)',\, E_s(\mu)} = \sum_{n \geq 1} B_n(\mu, f, \varphi). \tag{4.3}$$

Proof of Theorem 2.1. Let μ be in $I_m = (\lambda(0,m-1), \lambda(0,m))$. Let s and f satisfy the assumptions of Theorem 2.1. The function f is in $L^2_s(\Omega)$ and satisfies the zero-trace conditions (2.21). Formula (4.3) defines $u = R_A^{\pm}(\mu)f$ and holds in particular with φ in $C_0^{\infty}(\Omega)$. We apply Proposition 3.8 and Proposition 3.10 with $\tilde{s} = \mathrm{Max}(0, 1-s)$. We get

$$|\langle u, \overline{\varphi}\rangle_{\mathcal{D}'(\Omega), C_0^{\infty}(\Omega)}| = \left| \sum_{n \geq 1} B_n(\mu, f, \varphi) \right| \leq C \|f\|_{L^2_s(\Omega)} \|\varphi\|_{L^2_{\tilde{s}}(\Omega)} \tag{4.4}$$

with $C = C(\mu, s)$ independent of f in $L^2_s(\Omega)$, with φ in $C_0^{\infty}(\Omega)$, and continuous with respect to μ in I_m. Therefore, u is in $L^2_{-\tilde{s}}(\Omega)$ and satisfies (2.22). □

Proof of Theorem 2.2. Let $\mu = \lambda(0,m)$ and let s and f satisfy the assumptions of Theorem 2.2. The function f is in $NL^2_s(m,1)$, a closed hyperplane of $NL^2_s(m)$, and LAP2 is valid. With the zero-trace conditions (2.21), formula (4.3) defines $u = R_A^{\pm}(\mu)f$ and holds in particular with φ in $W(m) = NL^2_s(m) \cap C_0^{\infty}(\Omega)$. We apply Propositions 3.9, 3.10, and 3.11. The last of these requires f to be in $NL^2_s(m,1)$. We get

$$|\langle u, \overline{\varphi}\rangle_{NL^2_s(m)', NL^2_s(m)}| = \left| \sum_{n \geq 1} B_n(\mu, f, \varphi) \right| \leq C \|f\|_{L^2_s(\Omega)} \|\varphi\|_{L^2_{\tilde{s}}(\Omega)}, \tag{4.5}$$

where $\tilde{s} = \mathrm{Max}(0, 2-s)$ and $C = C(\mu, s)$ independent of f in $NL^2_s(m,1)$ and φ in $W(m)$. A density argument ends the proof: as $s > 3/2$, we have $\tilde{s}$ in $[0, 1/2)$, and the space $W(m)$ is dense in $L^2_{\tilde{s}}(\Omega)$. Therefore, u is in $L^2_{-\tilde{s}}(\Omega)$ and satisfies (2.24). □

Remark 4.1. The density argument about $W(m)$ can be omitted. Let us consider the form $U(\varphi) = \sum_{n \geq 1} B_n(\mu, f, \varphi)$ occuring in the right-hand side of (4.3). We apply point (i) of Proposition 3.9 if $n > m$, Proposition 3.10 if $n < m$, and Proposition 3.11 if $n = m$. With f in $NL^2_s(m,1)$ and φ in $C_0^{\infty}(\Omega)$, we get

$$|U(\varphi)| \leq C \|f\|_{L^2_s(\Omega)} \|\varphi\|_{L^2_{\tilde{s}}(\Omega)}.$$

The form U—and therefore u—is in $L^2_{-\tilde{s}}(\Omega)$.

5. Returning to the limiting absorption principle at thresholds. Let $m \geq 1$ be an integer and $\mu = \lambda(0, m)$ be the corresponding threshold of A. For $s > 1$ and f in $NL^2_s(m)$, the limit form $u^{\pm} = R^{\pm}_A(\mu)f$ in LAP2 is not defined on $L^2_{\tilde{s}}(\Omega)$. The trouble is caused by the following fact: when φ is in $L^2_{\tilde{s}}(\Omega) \setminus NL^2_s(m)$ and when ζ is not in $\sigma(A)$ and tends to μ, the modulus of $B_m(\zeta, f, \varphi)$ defined by (3.22) can tend to $+\infty$.

Let us consider the form $B_m(\mu, f, .)$ defined by (3.22) on $NL^2_s(m)$ when $s > 1$ and f is in $NL^2_s(m)$. On the basis of Proposition 3.11, this form can be extended to $L^2_{\tilde{s}}(\Omega)$, $\tilde{s} = \text{Max}(0, 2 - s)$, when $s > 3/2$ and f is in $NL^2_s(m, 1)$. On the subspace $L^2_s(\Omega)$ of $L^2_{\tilde{s}}(\Omega)$, this continuation is given by the integral in (3.53).

Let us now fix $s > 3/2$, f in $NL^2_s(m, 1)$, and φ in $L^2_s(\Omega)$. The question with respect to the prospect of a "better" limiting absorption principle at thresholds is whether the limit in (3.24) is still valid and if this limit is equal to $I_m(\mu, f, \varphi)$. A quick answer can be obtained. For ζ not in $\sigma(A)$, we consider

$$B_m(\zeta, f, \varphi) = \int_{-\infty}^{+\infty} \frac{\widetilde{f}(\xi, m)\, \overline{\widetilde{\varphi}(\xi, m)}}{\xi^2 a_m(\xi) - (\zeta - \mu)} d\xi. \tag{5.1}$$

As seen in Proposition 3.11, the function $\xi \mapsto (\widetilde{f}(\xi, m))/(\xi^2 a_m(\xi))$ is in $L^1_{\text{loc}}(\mathbb{R})$, and the function $\xi \mapsto \widetilde{\varphi}(\xi, m)$ is continuous on $\mathbb{R}$. Then with the condition $\text{Re}\zeta \leq \mu$, we may apply the Lebesgue dominated-convergence theorem and get

$$\lim_{\zeta \to \mu,\ \text{Re}\zeta\ \leq\ \mu,\ \zeta \notin \sigma(A)} B_m(\zeta, f, \varphi) = I_m(\mu, f, \varphi). \tag{5.2}$$

In fact, the restriction $\text{Re}\zeta \leq \mu$ in (5.2) can be raised. To do this, we have to use the finer properties of the trace functions, which are contained in Corollary 3.4 and Proposition 3.6. They allow us to apply Theorem 3.7.

5.1. Analysis of the term B_m near $\lambda(0, m)$.

It is more convenient to use the spectral variable λ in order to study the integral (3.22). For $j = 1$ or 2, for ζ not in $\sigma(A)$, and for f in $NL^2_s(m, 1)$ and φ in $L^2_s(\Omega)$, $s > 3/2$, we consider

$$B^j_m(\zeta, f, \varphi) = \int_{\lambda(0,m)}^{+\infty} \frac{f^j(\lambda, m)\, \overline{\varphi^j(\lambda, m)}}{\lambda - \zeta} d\lambda. \tag{5.3}$$

In view of (3.11), for $\lambda \in \overline{I_m} = [\lambda(0, m), +\infty)$, we set

$$h^j(\lambda) = f^j(\lambda, m)\, \overline{\varphi^j(\lambda, m)} = (\lambda - \lambda(0, m))^{1/4} F^j(\lambda, m) \overline{\varphi^j(\lambda, m)}. \tag{5.4}$$

It follows from Corollary 3.4 and Proposition 3.6 that the function $F^j(., m)$ is locally Hölder continuous on $\overline{I_m}$ with order δ in $[0, 1/2]$, $\delta < (s - 3/2)/2$, and vanishes at $\lambda = \lambda(0, m)$. We now specify the behavior of $\Phi^j(\lambda, m) = (\lambda - \lambda(0, m))^{1/4} \varphi^j(\lambda, m)$.

PROPOSITION 5.1. *Let $m \geq 1$ be an integer. Let $s > 1/2$ be a real number and φ be a function in $L^2_s(\Omega)$. Then the function $\Phi^j(., m)$ defined on $\overline{I_m} = [\lambda(0, m), +\infty)$ by*

$$\Phi^j(\lambda, m) = (\lambda - \lambda(0, m))^{1/4} \varphi^j(\lambda, m) \tag{5.5}$$

is locally Hölder continuous with order δ in $[0, 1/2]$, $\delta < (s - 1/2)/2$. Specifically, there exists a function $C(\lambda, \lambda') = C_m(s, \delta, \lambda, \lambda')$, continuous with respect to $\lambda, \lambda' \geq \lambda(0, m)$, such that

$$\forall \varphi \in L^2_s(\Omega), \quad |\Phi^j(\lambda, m) - \Phi^j(\lambda', m)| \leq C(\lambda, \lambda') |\lambda - \lambda'|^{\delta} \|\varphi\|_{L^2_s(\Omega)}. \tag{5.6}$$

In particular, if $s > 3/2$, $\Phi^j(.,m)$ *is locally Hölder continuous with order* $1/2$.

Proof. It is similar to the one concerning $F^j(.,m)$ in Corollary 3.4. Let $\lambda, \lambda' \geq \lambda(0,m)$ and let $\xi, \xi' \geq 0$ be connected by $\xi = \xi(\lambda, m)$ and $\xi' = \xi(\lambda', m)$. Estimate (2.15) already shows that $\Phi^j(.,m)$ is bounded. From (2.17d), we get

$$\Phi^j(\lambda, m) = H_m(\xi)^{1/2}\widetilde{\varphi}((-1)^j\xi, m).$$

The analyticity and strict positivity of H_m, the Hölder condition (2.11) of $\widetilde{\varphi}(.,m)$, with order δ' in $[0,1]$, $\delta' < s - 1/2$, imply

$$|\Phi^j(\lambda,m) - \Phi^j(\lambda',m)| \leq M(\xi,\xi')|\xi - \xi'|^{\delta'} \|\varphi\|_{L^2_s(\Omega)}.$$

With (2.17b), we get (5.6). □

We are now in a position to state the main result that will allow us to prove Theorem 2.3. This result has to be compared with point (ii) of Proposition 3.9.

First, we recall and fix some notation. According to (3.26), (3.24), (3.25), and (3.53), for a function f in $NL^2_s(m,1)$ and a function φ in $L^2_s(\Omega)$, we can define the function $B^{\pm}_m(.,f,\varphi)$ on $\overline{\mathbb{C}^{\pm}}$ by

$$B^{\pm}_m(\zeta, f, \varphi) = B_m(\zeta, f, \varphi) \quad \text{if } \pm\text{Im}\zeta > 0, \tag{5.7a}$$

$$B^{\pm}_m(\lambda, f, \varphi) = B_m(\lambda, f, \varphi) \quad \text{if } \lambda < \lambda(0,m), \tag{5.7b}$$

$$B^{\pm}_m(\lambda, f, \varphi) = B_m(\lambda, f, \varphi) \pm i\,\pi\, r_m(\lambda, f, \varphi) \quad \text{if } \lambda > \lambda(0,m), \tag{5.7c}$$

$$B^{\pm}_m(\lambda(0,m), f, \varphi) = I_m(\lambda(0,m), f, \varphi). \tag{5.7d}$$

THEOREM 5.2 (Hölder continuity and bounds for $B^{\pm}_m$ near $\lambda(0,m)$). *Let* $m \geq 1$ *be an integer and* $\lambda(0,m)$ *be the corresponding threshold of* A. *Let* J_m *be the interval* $(\lambda(0,m-1), \lambda(0,m+1))$ *and* $J^{\pm}_m = \{\zeta \in \overline{\mathbb{C}^{\pm}}\ /\ \text{Re}\zeta \in J_m\}$. *Let* s *and* δ *be real numbers such that*

$$s > 3/2 \quad \textit{and} \quad \delta \in [0,1/2], \quad \delta < (s - 3/2)/2. \tag{5.8}$$

Then there exist functions $C(\zeta,\zeta') = C_m(s,\delta,\zeta,\zeta')$ *and* $D(\zeta) = D_m(s,\delta,\zeta)$, *continuous with respect to* ζ *and* ζ', *such that for any complex numbers* ζ *and* ζ' *in* $J^{\pm}_m$ *and for any functions* f *in* $NL^2_s(m,1)$ *and* φ *in* $L^2_s(\Omega)$, *we have*

$$|B^{\pm}_m(\zeta, f, \varphi) - B^{\pm}_m(\zeta', f, \varphi)| \leq C(\zeta,\zeta')|\zeta - \zeta'|^{\delta}\|f\|_{L^2_s(\Omega)}\|\varphi\|_{L^2_s(\Omega)}, \tag{5.9}$$

$$|B^{\pm}_m(\zeta, f, \varphi)| \leq D(\zeta)\|f\|_{L^2_s(\Omega)}\|\varphi\|_{L^2_s(\Omega)}. \tag{5.10}$$

In particular, we have

$$\lim_{\zeta \to \lambda(0,m),\ \pm\text{Im}\zeta \geq 0} B^{\pm}_m(\zeta, f, \varphi) = I_m(\lambda(0,m), f, \varphi). \tag{5.11}$$

Proof. Let ζ be in $J^{\pm}_m$ and be not in $\sigma(A)$. We have $B^{\pm}_m(\zeta, f, \varphi) = B_m(\zeta, f, \varphi) = \sum_{j=1}^{2} B^j_m(\zeta, f, \varphi)$ with B^j_m defined in (5.3). We fix $\Lambda > 0$ and set $K = [\lambda(0,m), \lambda(0,m) + \Lambda]$. First, using (2.9), we easily get estimates similar to (5.9) with $\delta = 1$ and to (5.10) for the quantity $J(\zeta, f, \varphi) = \sum_{j=1}^{2} \int_{\lambda(0,m)+\Lambda}^{+\infty} (f^j(\lambda, m)\, \overline{\varphi^j(\lambda, m)})/(\lambda - \zeta) d\lambda$. Second, for $j = 1$ or 2, we study

$$I^j(\zeta, f, \varphi) = \int_K \frac{f^j(\lambda, m)\, \overline{\varphi^j(\lambda, m)}}{\lambda - \zeta} d\lambda = \int_K \frac{h^j(\lambda)}{\lambda - \zeta} d\lambda.$$

Our comments preceding Proposition 5.1 and the proposition itself imply that the function h^j is Hölder continuous on K with order δ in $[0, 1/2]$, $\delta < (s - 3/2)/2$, and there exists a constant $C_m = C_m(s, \delta, \Lambda)$ such that

(5.12)

$$\forall \lambda, \lambda' \in K, \quad \forall \varphi \in L_s^2(\Omega), \quad |h^j(\lambda) - h^j(\lambda')| \leq C|\lambda - \lambda'|^\delta \|f\|_{L_s^2(\Omega)} \|\varphi\|_{L_s^2(\Omega)}.$$

Moreover, the choice of f in $NL_s^2(m, 1)$ implies $F^j(\lambda(0, m), m) = 0$ so that $h^j(\lambda(0, m)) = 0$. Theorem 3.7 is then valid and gives the needed results to end the proof. □

5.2. Proof of Theorem 2.3. Let $\mu = \lambda(0, m)$ and let s and f satisfy the assumptions of Theorem 2.3. Since f is in $NL_s^2(m)$, $s > 3/2 > 1$, LAP2 is valid and formula (4.2) defines $u^\pm = R_A^\pm(\mu) f$ in the dual $NL_s^2(m)'$.

We proceed as in §3.2 of [CD95a].

Using Proposition 3.9 and Theorem 5.2, we get the limits in (2.18) in the norm topology on $B(NL_s^2(m, 1), L_{-s}^2(\Omega))$ and the Hölder conditions in (2.26) for $R_A^\pm$ under condition (5.8) for s and δ.

Formula (2.27) for $\langle R_A^\pm(\mu) f, \varphi \rangle_{L_{-s}^2(\Omega), L_s^2(\Omega)}$, is a consequence of formulas (3.53) and (5.11) for I_m, (3.22) for B_n if $n > m$, and (3.23) and (3.25) for B_n and r_n if $n < m$. When φ is in $NL_s^2(m)$, we again find $\langle u^\pm, \varphi \rangle_{NL_s^2(m)', NL_s^2(m)}$.

Let us now look at the gradient of $R_A^\pm(\zeta) f$. First, we recall a property already used in [CD95a] concerning ∇v when v in $D(A)$. Let t be a real number and v be a function in $D(A) \cap L_t^2(\Omega)$ such that Av is also in $L_t^2(\Omega)$. Then there exists a constant $C = C(t, A)$ such that

$$\|\nabla v\|_{L_t^2(\Omega)} \leq C\, (\|v\|_{L_t^2(\Omega)} + \|Av\|_{L_t^2(\Omega)}).$$

Therefore, when ζ is not in $\sigma(A)$, with $t = -s$ and $v = R_A^\pm(\zeta) f - R_A^\pm(\zeta') f$, we get an estimate similar to (2.26) for $\nabla R_A^\pm(\zeta)$ since $Av = \zeta R_A^\pm(\zeta) f - \zeta' R_A^\pm(\zeta') f$.

Finally, the fact that $U^\pm$ belongs to $D(A)_{\text{loc}}$ and the differential equation (2.28) follow from $\nabla .\, c^2 \nabla R_A(\zeta) = I + \zeta R_A(\zeta)$. □

We are now ready to study the perturbed operator presented in the introduction. We shall develop a perturbative method with ingredients adapted from Majda [Ma]. Preliminary results are presented in [BCD] and detailed and completed in [CD95b].

REFERENCES

[A] S. Agmon, *Spectral properties of Schrödinger operators and scattering theory*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 2 (1975), pp. 151–218.

[BCD] J.-L. Boelle, E. Croc, and Y. Dermenjian, *Spectral and numerical analysis of wave equation* 2D *in a stratified acoustical or elastic media with two welded stratifications*, in Proc. 2nd International SIAM–INRIA Conference on Mathematical and Numerical Aspects of Wave Propogation, Society for Industrial and Applied Mathematics, Philadelphia, 1993, pp. 82–91.

[CD95a] E. Croc and Y. Dermenjian, *Analyse spectrale d'une bande acoustique multistratifiée partie* I: *Principe d'absorption limite pour une stratification simple*, SIAM J. Math. Anal., 26 (1995), pp. 880–924.

[CD95b] ———, *Spectral analysis of a multistratified acoustic strip: Distribution of eigenvalues and perturbative method*, Rapport interne 6, Jeune Equipe EDP–AN, Université de Provence, Marseille, France, 1995; SIAM J. Math. Anal., submitted.

[DG] Y. Dermenjian and J. C. Guillot, *Théorie spectrale de la propagation des ondes acoustiques dans un milieu stratifié perturbé*, J. Differential Equations, 62 (1986), pp. 357–409.

[G] F. D. Gakhov, *Boundary Value Problems*, Pergamon Press, London, 1966.

[H] L. Hörmander, *The analysis of linear partial differential operators* II: *Differential operators with constant coefficients*, in Grundelehren der mathematischen Wissenschaften 257, Springer-Verlag, Berlin, New York, 1983.

[Ma] A. Majda, *Outgoing solutions for perturbations of $-\Delta$ with applications to spectral and scattering theory*, J. Differential Equations 16 (1974), pp. 515–547.

[Mu] N. I. Muskhelishvili, *Singular Integral Equations*, J. R. M. Radok, ed., Noordhoff International Publishing, Leiden, the Netherlands, 1977 (revised translation from the Russian).

[T81a] N. Tamura, *The principle of limiting absorption for uniformly propagative systems with perturbations of long range class*, Nagoya Math. J., 82 (1981), pp. 141–174.

[T81b] ———, *The principle of limiting absorption for propagative systems in crystal optics with perturbations of long range class*, Nagoya Math. J., 84 (1981), pp. 169–193.

[V] B. Vainberg, *Asymptotics Methods in Equations of Mathematical Physics*, Gordon and Breach, New York, 1989.

[We] R. Weder, *Spectral and Scattering Theory for Wave Propagation in Perturbed Stratified Media*, Appl. Math. Sci. 87, Springer-Verlag, New York, Berlin, 1991.

SIAM J. MATH. ANAL.
Vol. 27, No. 6, pp. 1653–1660, November 1996

SEMILINEAR ELLIPTIC EQUATIONS IN $\mathbb{R}^N$ WITH ALMOST PERIODIC OR UNBOUNDED FORCING TERM[*]

GILLES FOURNIER[†], ANDRZEJ SZULKIN[‡], AND MICHEL WILLEM[§]

Our colleague and friend, Gilles Fournier, died untimely on August 13, 1995.
A. Szulkin, M. Willem

Abstract. This work is devoted to the existence and uniqueness of almost periodic solutions in $\mathbb{R}^N$ of the equation $-\Delta u + \sum_{j=1}^N c_j\partial_j u + g(u) = h(x)$. We also prove the existence of solutions with the same growth as some unbounded forcing terms. Under a local monotonicity assumption, the method of upper and lower solutions is used.

Key words. semilinear elliptic equations, almost periodic functions, upper and lower solutions, unbounded forcing term

AMS subject classifications. 34C27, 35B15, 58G20.

1. Introduction.

Let us consider the problem

$$-\Delta u + \sum_{j=1}^{N} c_j\partial_j u + g(u) = h(x), \quad x \in \mathbb{R}^N. \tag{1.1}$$

Many papers are devoted to semilinear elliptic problems of this type; see, e.g., [10]. If h has some additional properties, one may look for solutions of (1.1) which also have such properties. In particular, if h is almost periodic, the existence of almost periodic oscillations for (1.1) is a natural problem (see [2], [5], [7], [12]). The method of [1]–[4] and [7] in treating (1.1) is to use minimization on a Besicovitch–Sobolev space of almost periodic functions. It is then a delicate (and in most cases unsolved) problem to show that the minimizing Besicovitch–Sobolev class contains a function which is a solution of (1.1) in the usual sense. Moreover, in [1]–[5] and [7], only ordinary differential equations are considered, and the variational approach there excludes the possibility of having first-order terms in (1.1).

In this paper, we use the method of upper and lower solutions. In §2, we prove the existence of a locally bounded solution of (1.1) and obtain a basic estimate. In §3, we show that (1.1) has a unique almost periodic solution if g is increasing and h almost periodic. We would like to emphasize here that while in [1]–[5] and [7] it is essential that one has an ordinary differential equation without first-order terms, our method makes it possible to also treat problems *with* first-order terms (like the pendulum with friction—see Example 3.3 below—with $x = t$ and $N = 1$) and *partial* differential equations with almost periodic forcing.

In §4, we consider equation (1.1) with forcing term h having polynomial or subexponential growth. Using appropriate upper and lower solutions, we prove the existence of solutions, with the same growth rate as h.

[*] Received by the editors August 11, 1994; accepted for publication (in revised form) August 1, 1995.

[†] Départment de Mathématiques, Université de Sherbrooke, Sherbrooke, PQ J1K 2R1, Canada.

[‡] Matematiska Institutionen, Stockholms Universitet, S-10691 Stockholm, Sweden. The research of this author was supported in part by the Swedish Natural Science Research Council.

[§] Département de Mathématiques, Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium.

2. Upper and lower solutions. We first consider the Dirichlet problem

$$\text{(2.1)}\qquad \begin{cases} -\Delta u + \sum_{j=1}^{N} c_j \partial_j u + f(x,u) = h(x), & x \in \Omega \\ u(x) = \gamma(x), \quad x \in \Gamma, \end{cases}$$

where Ω is an open-bounded domain in $\mathbb{R}^N$, $\Gamma := \partial\Omega$ is a smooth submanifold, the c_j's are real constants, $f \in \mathcal{C}(\bar{\Omega} \times \mathbb{R})$, $h \in L^\infty(\Omega)$, and $\gamma \in H^{1/2}(\Gamma)$.

DEFINITION 2.1. *The function $\alpha \in H^1(\Omega) \cap L^\infty(\Omega)$ is a lower solution of* (2.1) *if, for every $v \in \mathcal{D}(\Omega)$,*

$$\text{(2.2)}\qquad v \geq 0 \Rightarrow \int_\Omega \sum_{j=1}^{N} \partial_j \alpha (\partial_j v + c_j v)\, dx + \int_\Omega f(x,\alpha) v\, dx \leq \int_\Omega hv\, dx$$

and, for almost every $x \in \Gamma$,

$$\alpha(x) \leq \gamma(x).$$

The function $\beta \in H^1(\Omega) \cap L^\infty(\Omega)$ is an upper solution of (2.1) *if, for every $v \in \mathcal{D}(\Omega)$,*

$$\text{(2.3)}\qquad v \geq 0 \Rightarrow \int_\Omega \sum_{j=1}^{N} \partial_j \beta (\partial_j v + c_j v)\, dx + \int_\Omega f(x,\beta) v\, dx \geq \int_\Omega hv\, dx$$

and, for almost every $x \in \Gamma$,

$$\beta(x) \geq \gamma(x).$$

Remark 2.2. It is easy to verify that (2.2) and (2.3) hold for every $v \in H_0^1(\Omega)$ if α is a lower and β an upper solution.

The following result is contained in a theorem of Deuel and Hess [8]. We give a sketch of the proof for the sake of completeness.

THEOREM 2.3. *Assume that α is a lower and β an upper solution of* (2.1) *and that $\alpha \leq \beta$. Then problem* (2.1) *has a weak solution $u \in H^1(\Omega)$ such that $\alpha \leq u \leq \beta$. Moreover, $u \in \mathcal{C}^1(\Omega)$.*

Proof. 1. Consider the modified problem

$$\text{(2.4)}\qquad \begin{cases} -\Delta u + \sum_{j=1}^{N} c_j \partial_j u = -F(x,u) + h(x), & x \in \Omega, \\ u(x) = \gamma(x), \quad x \in \Gamma, \end{cases}$$

where F is defined on $\bar{\Omega} \times \mathbb{R}$ by

$$\begin{aligned} F(x,u) &:= f(x,\alpha(x)) \quad \text{if } u < \alpha(x), \\ &:= f(x,u) \quad \text{if } \alpha(x) \leq u \leq \beta(x), \\ &:= f(x,\beta(x)) \quad \text{if } \beta(x) < u. \end{aligned}$$

Since the linear operator in (2.4) is invertible, (2.4) is equivalent to the fixed-point problem

$$\text{(2.5)}\qquad u = Au, \quad u \in K = \{w \in H^1(\Omega) : w|\Gamma = \gamma\}.$$

Since A is completely continous and has a bounded range, problem (2.5) has a solution u by the Schauder fixed-point theorem. Moreover, $u \in W^{2,p}_{\text{loc}}(\Omega)$ for $1 < p < \infty$. In particular, $u \in \mathcal{C}^{1,\alpha}_{\text{loc}}(\Omega)$ for $0 < \alpha < 1$.

2. It remains only to prove that $\alpha \le u \le \beta$ on Ω. Since $v := (\alpha - u)^+ \in H_0^1(\Omega)$, we obtain, from (2.2) and (2.4),

$$\int_\Omega \sum_{j=1}^N \partial_j(\alpha - u)(\partial_j v + c_j v)\, dx \le \int_\Omega (F(x,u) - f(x,\alpha))v\, dx = 0$$

so that

$$\int_\Omega \sum_{j=1}^N (\partial_j(\alpha - u)^+)^2\, dx = 0.$$

Hence $(\alpha - u)^+$ is constant on Ω and, necessarily, $(\alpha - u)^+ = 0$ on Ω. Similarly, we have that $(u - \beta)^+ = 0$ on Ω. □

We now consider the problem

$$-\Delta u + \sum_{j=1}^N c_j \partial_j u + f(x,u) = h(x), \quad x \in \mathbb{R}^N, \tag{2.6}$$

where $f \in \mathcal{C}(\mathbb{R}^N \times \mathbb{R})$ and $h \in L^\infty_{\text{loc}}(\mathbb{R}^N)$.

DEFINITION 2.4. *The function $\alpha \in H^1_{\text{loc}}(\mathbb{R}^N) \cap L^\infty_{\text{loc}}(\mathbb{R}^N)$ is a lower solution of* (2.6) *if, for every $v \in \mathcal{D}(\mathbb{R}^N)$,* (2.2) *holds with $\Omega = \mathbb{R}^N$.*

The function $\beta \in H^1_{\text{loc}}(\mathbb{R}^N) \cap L^\infty_{\text{loc}}(\mathbb{R}^N)$ is an upper solution of (2.6) *if, for every $v \in \mathcal{D}(\mathbb{R}^N)$,* (2.3) *holds with $\Omega = \mathbb{R}^N$.*

THEOREM 2.5. *Assume that α is a lower and β an upper solution of* (2.6) *and $\alpha \le \beta$. Then equation* (2.6) *has a weak solution $u \in \mathcal{C}^1(\mathbb{R}^N)$ such that $\alpha \le u \le \beta$.*

Proof. 1. Let $n \ge 1$ be a fixed integer. Theorem 2.3, applied to $\Omega_n := B(0,n)$, implies the existence of a solution $u_n \in H^1(\Omega_n) \cap \mathcal{C}^1(\Omega_n)$ of

$$\begin{gathered} -\Delta u + \textstyle\sum_{j=1}^N c_j \partial_j u + f(x,u) = h(x), \quad |x| < n, \\ u(x) = \alpha(x), \quad |x| = n, \end{gathered}$$

such that $\alpha \le u_n \le \beta$ on Ω_n.

2. By construction, $(u_n)_{n\ge 2}$ and $(Lu_n)_{n \ge 2}$, where $L = -\Delta + \sum_{j=1}^N c_j \partial_j$, are bounded in $L^\infty(\Omega_2)$. Inequality (4.6) in [11] implies that $(u_n)_{n\ge 2}$ is bounded in $H^1(\Omega_1)$. Thus there exists a subsequence (u_n^1) such that, in $H^1(\Omega_1)$,

$$u_n^1 \rightharpoonup u^1, \quad n \to \infty.$$

By a repeated selection of subsequences, we obtain a sequence (u_n^k) such that in $H^1(\Omega_k)$,

$$u_n^k \rightharpoonup u^k, \quad n \to \infty.$$

It is clear that $u^{k-1} = u^k$ on Ω_{k-1}. Since by the Rellich theorem, in $L^2(\Omega_k)$,

$$u_n^k \to u^k, \quad n \to \infty,$$

it is clear that $\alpha \le u^k \le \beta$ and u^k is a weak solution of

$$-\Delta u + \sum_{j=1}^N c_j \partial_j u + f(x,u) = h(x)$$

on Ω_k. Let us define $u \in H^1_{\text{loc}}(\mathbb{R}^N)$ by $u = u^k$ on Ω_k. Then u is a weak solution of (2.6), $u \in \mathcal{C}^1(\mathbb{R}^N)$, and $\alpha \le u \le \beta$. □

Remark 2.6. The above result generalizes Theorem 2.10 in [10], where it is assumed that $c_1 = \cdots = c_N = 0$, f is a locally Hölder continuous function which is locally Lipschitz continuous in u, and $h = 0$.

We shall now prove a basic estimate.

THEOREM 2.7. *Let $h_1, h_2, \alpha, \beta \in L^\infty(\mathbb{R}^N)$. Assume that there exists $\delta > 0$ such that*

$$\alpha(x) \le s \le t \le \beta(x) \Rightarrow f(x,t) - f(x,s) \ge \delta(t-s). \tag{2.7}$$

If, for $j = 1, 2$, $u_j \in \mathcal{C}^1(\mathbb{R}^N)$ is a weak solution of

$$-\Delta u + \sum_{j=1}^N c_j \partial_j u + f(x,u) = h_j(x)$$

such that $\alpha \le u_j \le \beta$, then

$$|u_1 - u_2|_\infty \le \delta^{-1} |h_1 - h_2|_\infty.$$

Proof. Let us define $c := \delta^{-1}|h_1 - h_2|_\infty$, $v := u_2 - u_1 - c$, and $\Omega := \{x \in \mathbb{R}^N : v(x) > 0\}$. It follows from (2.7) that on Ω,

$$\begin{aligned} -\Delta v + \delta v + \sum_{j=1}^N c_j \partial_j v &= f(x,u_1) - f(x,u_2) + h_2 - h_1 + \delta v \\ &\le -\delta(u_2 - u_1) + |h_2 - h_1|_\infty + \delta v \\ &= |h_2 - h_1|_\infty - \delta c = 0. \end{aligned} \tag{2.8}$$

Let $\omega := v/\psi$, where $\psi(x) = \Pi_{j=1}^N \cosh \alpha x_j$, $\alpha > 0$, and

$$\alpha^2 N + \sum_{j=1}^N \alpha |c_j| < \delta.$$

A simple computation using (2.8) shows that on Ω,

$$-\Delta\omega + \sum_{j=1}^N \partial_j \omega (c_j - 2\alpha \tanh \alpha x_j) + \left(\delta - \alpha^2 N + \sum_{j=1}^N \alpha c_j \tanh \alpha x_j\right) \omega \le 0. \tag{2.9}$$

Assume that $\Omega \ne \phi$ and let $\Omega_R := \Omega \cap B(0,R)$. Since $v \in L^\infty(\mathbb{R}^N)$, $\omega(x) \to 0$ as $|x| \to \infty$. Therefore, for R large enough, there exists $\epsilon > 0$ such that

$$\max_{\partial\Omega_R} \omega < \epsilon < \max_{\Omega_R} \omega.$$

Multiplying (2.9) by $h := (\omega - \epsilon)^+ \in H^1_0(\Omega_R)$ and integrating by parts, we obtain

$$\begin{aligned} \int_{\Omega_R} |\nabla h|^2 \, dx + \alpha^2 \sum_{j=1}^N \int_{\Omega_R} h^2/\cosh^2 \alpha x_j \, dx \\ + \int_{\Omega_R} \left(\delta - \alpha^2 N + \sum_{j=1}^N \alpha c_j \tanh \alpha x_j\right)(h+\epsilon) h \, dx \le 0. \end{aligned}$$

Hence $h = 0$ and $\omega \le \epsilon$ on Ω_R—a contradiction. Thus $u_2 - u_1 \le c$. Interchanging u_1 and u_2 gives $u_1 - u_2 \le c$. □

Remark 2.8. The growth-damping factor ψ has been used to study linear elliptic equations on $\mathbb{R}^N$ (see, e.g., [9, p. 187]).

COROLLARY 2.9. *Suppose that $f(x,t) - f(x,s) \ge \delta(t-s)$ for some $\delta > 0$ and all $t \ge s$. Then equation* (2.6) *has at most one weak solution $u \in \mathcal{C}^1(\mathbb{R}^N)$ which is subexponential in the sense that $u(x)/\prod_{j=1}^N \cosh \alpha x_j$ is bounded for each $\alpha > 0$.*

Proof. In the proof of Theorem 2.7, we now have $c = 0$ and $v = u_2 - u_1$. Since v is subexponential, $\omega(x) = v(x)/\psi(x) \to 0$ as $|x| \to \infty$ and the argument of Theorem 2.7 gives $u_2 - u_1 \le 0$. Similarly, $u_1 - u_2 \le 0$. □

3. Uniformly almost periodic solutions. Let us recall that the translate of a function is defined by $\tau_c u(x) := u(x-c)$.

The following definition is due to Bochner (see, e.g., [6]).

DEFINITION 3.1. *A function $u \in \mathcal{BC}(\mathbb{R}^N)$ is uniformly almost periodic (UAP) if for every sequence $(c_n) \subset \mathbb{R}^N$, $(\tau_{c_n} u)$ contains a uniformly convergent subsequence on $\mathbb{R}^N$.*

We now consider the problem

$$\begin{cases} -\Delta u + \sum_{j=1}^N c_j \partial_j u + g(u) = h(x), & x \in \mathbb{R}^N, \\ u \in \mathcal{BC}(\mathbb{R}^N), \end{cases} \tag{3.1}$$

where $g \in \mathcal{C}(\mathbb{R})$ and $h \in L^\infty(\mathbb{R}^N)$.

THEOREM 3.2. *Assume that $\alpha \in H^1_{\text{loc}}(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$ is a lower and $\beta \in H^1_{\text{loc}}(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$ an upper solution of* (3.1) *and $\alpha \le \beta$. If there exists $\delta > 0$ such that*

$$\inf \text{ess}\, \alpha \le s \le t \le \sup \text{ess}\, \beta \Rightarrow g(t) - g(s) \ge \delta(t-s),$$

then problem (3.1) *has a unique weak solution $u \in \mathcal{C}^1(\mathbb{R}^N)$ such that $\alpha \le u \le \beta$. Moreover, if h is UAP, then u is also UAP.*

Proof. Theorem 2.5 implies the existence of a solution $u \in \mathcal{C}^1(\mathbb{R}^N)$ of (3.1) such that $\alpha \le u \le \beta$. Uniqueness follows from Theorem 2.7.

Now assume that h is UAP and consider a sequence $(c_n) \subset \mathbb{R}^N$. For simplicity, we write

$$h_n := \tau_{c_n} h, \qquad u_n := \tau_{c_n} u.$$

Going if necessary to a subsequence, we can assume that (h_n) converges uniformly on $\mathbb{R}^N$. It is clear that

$$-\Delta u_n + \sum_{j=1}^N c_j \partial_j u_n + g(u_n) = h_n.$$

Theorem 2.7 implies that

$$|u_j - u_k|_\infty \le \delta^{-1} |h_j - h_k|_\infty \to 0, \quad j, k \to \infty.$$

Hence (u_n) converges uniformly on $\mathbb{R}^N$ and u is UAP. □

Example 3.3. Consider the equation

$$-\Delta u + \sum_{j=1}^N c_j \partial_j u - \sin u = h(x), \tag{3.2}$$

where h is UAP. If $\|h\|_\infty < 1$, then equation (3.2) has a unique UAP solution such that $\pi/2 < u < 3\pi/2$. Indeed, for ϵ and δ positive and small enough, $\pi/2 + \epsilon$ is a lower solution, $3\pi/2 - \epsilon$ an upper solution, and

$$\sin s - \sin t \geq \delta(t - s)$$

for $\pi/2 + \epsilon \leq s \leq t \leq 3\pi/2 - \epsilon$. The particular case

$$-\ddot{u} = \sin u + h(x)$$

is treated in [7]. Assuming that $\|h\|_\infty < 1$, uniqueness is proved there for every h and existence for a dense set of h. See also [1]–[4] for the existence of generalized solutions.

Example 3.4. Consider the equation

$$-\Delta u + \sum_{j=1}^{N} c_j \partial_j u + u + u^3 = h(x), \tag{3.3}$$

where h is UAP. Using Theorem 3.2, it is easy to verify that equation (3.3) has a unique UAP solution. The particular case

$$-\ddot{u} + u + u^3 = h(x)$$

is treated in [5].

4. Unbounded forcing term. In this section, we consider problem (2.6) with a measurable function h having polynomial or subexponential growth.

DEFINITION 4.1. *A function h is said to be of polynomial growth if there is an integer $m \geq 0$ and a constant C such that*

$$|h(x)| \leq C\left(1 + \sum_{j=1}^{N} |x_j|^m\right) \quad \textit{for almost all } x \in \mathbb{R}^N, \tag{4.1}$$

and h is of subexponential growth if for each $\gamma > 0$, there is a C_γ such that

$$|h(x)| \leq C_\gamma \prod_{j=1}^{N} \cosh \gamma x_j \quad \textit{for almost all } x \in \mathbb{R}^N.$$

THEOREM 4.2. *Suppose that there are $\delta > 0$ and $R > 0$ such that $f(x,t)t \geq \delta t^2$ whenever $|t| \geq R$. If h is a measurable function which satisfies* (4.1) *for some integer $m \geq 0$, then* (2.6) *has a weak solution $u \in C^1(\mathbb{R}^N)$ which also satisfies* (4.1) (*with the same m*).

Proof. Let $\varphi(x) := D(1 + \gamma \sum_{j=1}^{N} |x_j|^m)$, where $\gamma > 0$, $D \geq R$, and $m \geq 2$. Then

$$-\Delta\varphi + \sum_{j=1}^{N} c_j \partial_j \varphi + f(x, \varphi) \geq -\gamma m(m-1) D \sum_{j=1}^{N} |x_j|^{m-2}$$
$$+ \gamma m D \sum_{j=1}^{N} c_j |x_j|^{m-2} x_j + \delta D\left(1 + \gamma \sum_{j=1}^{N} |x_j|^m\right).$$

Choosing γ small enough, we obtain

$$-\Delta\varphi + \sum_{j=1}^{N} c_j \partial_j \varphi + f(x, \varphi) \geq \epsilon D \left(1 + \sum_{j=1}^{N} |x_j|^m\right) \tag{4.2}$$

for some $\epsilon > 0$. Hence $\alpha := -\varphi$ is a lower and $\beta := \varphi$ an upper solution if D is large enough. Therefore, the conclusion follows from Theorem 2.5.

If $m = 1$, let $\varphi_0(x) := D(1 + \gamma \sum_{j=1}^{N} |x_j|)$, where $D \geq R$, and let φ be obtained from φ_0 by "rounding off the corner" in each $|x_j|$. Then it is easy to see that (4.2) is satisfied (with $m = 1$) whenever γ is small enough. Finally, if $m = 0$, there is a constant lower and a constant upper solution. □

THEOREM 4.3. *If f satisfies the hypotheses of Theorem* 4.2 *and h is a measurable function of subexponential growth, then* (2.6) *has a weak solution $u \in C^1(\mathbb{R}^N)$ which is also of subexponential growth.*

Proof. Let $\varphi(x) := D \prod_{j=1}^{N} \cosh \gamma x_j$, where $\gamma > 0$ and $D \geq R$. Choosing γ sufficiently small, we obtain

$$-\Delta\varphi + \sum_{j=1}^{N} c_j \partial_j \varphi + f(x, \varphi) \geq \epsilon\varphi$$

for some $\epsilon > 0$, and again we can take $\alpha := -\varphi$ and $\beta := \varphi$ with D large enough. □

THEOREM 4.4. *Suppose that $f(x,t) - f(x,s) \geq \delta(t - s)$ for some $\delta > 0$ and all $t \geq s$, $x \in \mathbb{R}^N$. Then the solution of* (2.6) *which was obtained in Theorems* 4.2 *and* 4.3 *is unique in the class of all functions $u \in C^1(\mathbb{R}^N)$ having subexponential growth.*

Proof. This is an immediate consequence of Corollary 2.9. □

Remark 4.5. (i) Under the hypotheses of Theorem 4.4, equation (2.6) may have many solutions as is demonstrated by the example $-u'' + u = 0$, but only one solution is subexponential.

(ii) In [13], it was shown that if $f(x,t)t \geq \delta|t|^{p+1}$ for large $|t|$ and $h \in L^1_{\text{loc}}(\mathbb{R}^N)$, then (2.6) has a solution $u \in L^p_{\text{loc}}(\mathbb{R}^N)$. Moreover, this solution is unique in $L^p_{\text{loc}}(\mathbb{R}^N)$ if f satisfies a suitable monotonicity assumption.

Acknowledgments. This work was done while the first and second authors were visiting Louvain-la-Neuve. They would like to thank the members of the Department of Mathematics at Universite Catholique de Louvain for their hospitality. The second author would also like to thank Jan Chabrowski, from whom he learned the method of introducing the growth damping factor.

REFERENCES

[1] J. M. BELLEY, G. FOURNIER, AND S. HAYES, *Existence of almost periodic weak type solutions for the conservative forced pendulum equation*, preprint.

[2] J. M. BELLEY, G. FOURNIER, AND H. SAADI DRISSI, *Almost periodic weak solutions to forced pendulum type equations without friction*, Aequationes Math., 44 (1992), pp. 100–108.

[3] ———, *Solutions presque periodiques d'equations differentielles du type pendule force*, Acad. Roy. Belg. Bull. Cl. Sci. (6), 3 (1992), pp. 173–186.

[4] ———, *Solutions presque periodiques de systeme differentiel du type pendule force couple*, preprint.

[5] M. S. BERGER AND Y. Y. CHEN, *Forced quasi periodic and almost periodic oscillations of nonlinear Duffing equations*, Nonlinear Anal., 19 (1992), pp. 249–257.

[6] A. S. BESICOVITCH, *Almost periodic functions*, Dover, New York, 1954.

[7] J. Blot, *Une methode hilbertienne pour les trajectoires presque periodiques*, C. R. Acad. Sci. Paris Ser. I Math., 313 (1991), pp. 487–490.

[8] J. Deuel and P. Hess, *A criterion for the existence of solutions of non-linear elliptic boundary value problems*, Proc. Roy. Soc. Edinburgh Sect. A, 74 (1974), pp. 49–54.

[9] M. Krzyżański, *Partial Differential Equations of Second Order*, vol. I, PWN, Warsaw, 1971.

[10] W. M. Ni, *On the elliptic equation $\Delta u + K(x)u^{(n+2)/(n-2)} = 0$, its generalizations and applications in geometry*, Indiana Univ. Math. J., 31 (1982), pp. 493–529.

[11] L. Nirenberg, *Remarks on strongly elliptic partial differential equations*, Comm. Pure Appl. Math., 8 (1955), pp. 648–674.

[12] K. Schmitt and J. Ward, *Almost periodic solutions of nonlinear second order differential equations*, Results Math., 21 (1992), pp. 190–199.

[13] H. Brezis, *Semilinear equations in $\mathbb{R}^N$ without condition at infinity*, Appl. Math. Optim., 12 (1984), pp. 271–282.

SIAM J. MATH. ANAL.
Vol. 27, No. 6, pp. 1661–1679, November 1996

008

WHAT IS THE SUBDIFFERENTIAL OF THE CLOSED CONVEX HULL OF A FUNCTION?*

J. BENOIST† AND J.-B. HIRIART-URRUTY‡

Abstract. Given a function $f : \mathbb{R}^n \to (-\infty, +\infty]$ and its closed convex hull $\overline{\text{co}}f$, we consider the question of expressing the subdifferential of $\overline{\text{co}}f$ in terms of the subdifferential of f. Under a fairly general assumption on the behavior of f at infinity, we obtain an explicit formula of the subdifferential of $\overline{\text{co}}f$ from that of f and its asymptotic function.

Key words. convex hull, subdifferential, asymptotic function

AMS subject classifications. 26B05, 26B25, 49H05

1. Introduction.

Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ be any function. The *closed convex hull* (or closed convex envelope) $\overline{\text{co}}f$ of f can be defined in various ways such as the following: $\overline{\text{co}}f$ is the greatest closed convex function maximized by f on $\mathbb{R}^n$; or (when f is minimized by some affine function) $\overline{\text{co}}f$ is the Legendre–Fenchel biconjugate of f. In view of the generality of the class of functions f that we are considering, the calculation of $(\overline{\text{co}}f)(x)$ is, as a general rule, extremely complicated. However, we should be able to derive qualitative properties on $\overline{\text{co}}f$ from the corresponding properties of f without having to compute $(\overline{\text{co}}f)(x)$ for all x. For example, is $\overline{\text{co}}f$ differentiable whenever f is differentiable? The answer is, in general, no: there are C^∞ functions $f : \mathbb{R}^2 \to \mathbb{R}$ whose closed convex hulls are not (even once) differentiable. However, the answer is yes under a fairly general assumption on the behavior of f at infinity, as will be shown in §4. One of the key ingredients for studying such differentiability properties is *a formula linking the subdifferential of f and that of $\overline{\text{co}}f$*. To derive such an explicit formula is precisely the aim of this paper. This paper is organized as follows. General assumptions on f and properties of $\overline{\text{co}}f$ as well as preliminary results from subdifferential theory are presented in §2. The case where f is 1-coercive (i.e., f satisfies $\lim_{\|x\| \to +\infty} f(x)/\|x\| = +\infty$) is recalled in §3 from [9, Chap. X, §1.5]. This is motivated by the importance of 1-coercive functions f in the context of calculating $\overline{\text{co}}f$ as well as by the necessity of preparing the more general situation treated in §4. When f is no longer 1-coercive, to calculate $\overline{\text{co}}f$, the behavior of f at infinity must somehow be taken into account. This is done via the so-called asymptotic function f_∞ of f whose essential properties are displayed at the beginning of §4. Under a general assumption on f_∞, in §4, we give an explicit formula of the subdifferential of $\overline{\text{co}}f$ in terms of the subdifferentials of f and f_∞.

2. Preliminaries.

2.1. From a function to its closed convex hull.

We assume throughout that $\mathbb{R}^n$ is equipped with the standard inner product denoted by $\langle .,. \rangle$, and $f : \mathbb{R}^n \to (-\infty, +\infty]$ satisfies (at least)

$$\text{dom} f := \{x \in \mathbb{R}^n \mid f(x) < +\infty\} \text{ is nonempty;} \tag{1}$$

$$\text{there is an affine function minimizing } f \text{ on } \mathbb{R}^n. \tag{2}$$

* Received by the editors April 6, 1994; accepted for publication (in revised form) June 9, 1995.

† Département de Mathématiques, Université de Limoges, 123 Rue A. Thomas, 87060 Limoges, France.

‡ U.F.R. Mathématiques, Informatique, Gestion, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse cedex, France.

Such an f is convex if and only if its epigraph $\text{epi}f := \{(x,r) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq r\}$ is convex. If not, a natural way to "convexify" f is to take the convex hull $\text{co}(\text{epi}f)$ of $\text{epi}f$. This gives a convex set, which is not necessarily an epigraph (consider $x \in \mathbb{R} \mapsto f(x) = \sqrt{|x|}$), but which can be made so by "closing its bottom"; we thus define the *convex hull* $\text{co}f$ of f by

$$(3) \qquad x \in \mathbb{R}^n \ \mapsto \ (\text{co}f)(x) = \inf\{r \in \mathbb{R} \ \mid \ (x,r) \in \text{co}(\text{epi}f)\}$$

($\inf \emptyset = +\infty$ by convention).

There are several ways of getting at $(\text{co}f)(x)$ (cf. [14], for example); statements (4) and (5) recall them. By the unit simplex of $\mathbb{R}^{n+1}$, we mean

$$\Delta_{n+1} := \left\{(\alpha_1, \ldots, \alpha_{n+1}) \ \in \ \mathbb{R}^{n+1} \mid \sum_{i=1}^{n+1} \alpha_i = 1, \quad \alpha_i \geq 0 \quad \text{for } i = 1, \ldots, n+1\right\}.$$

For all $x \in \mathbb{R}^n$, we have

$$(4) \qquad (\text{co}f)(x) = \sup\{g(x) \mid g : \mathbb{R}^n \to (-\infty, +\infty] \text{ convex}, \quad g \leq f\}$$

$$(5) \qquad (\text{co}f)(x) = \inf\left\{\sum_{i=1}^{n+1} \alpha_i f(x_i) \mid (\alpha_1, \ldots, \alpha_{n+1}) \in \Delta_{n+1}, \quad x_i \in \text{dom}\, f, \sum_{i=1}^{n+1} \alpha_i x_i = x\right\}.$$

Instead of $\text{co}(\text{epi}f)$, we can take the closed convex hull $\overline{\text{co}}($
epif) of $\text{epi}f$; we obtain a closed set, which is now always an epigraph. The so-called *closed convex hull* $\overline{\text{co}}f$ of f is thus defined by

$$(6) \qquad x \ \in \ \mathbb{R}^n \ \mapsto \ (\overline{\text{co}}f)(x) \ = \ \min\{r \in \mathbb{R} \mid (x,r) \in \overline{\text{co}}(\text{epi}f)\}.$$

Similarly to (4), for all $x \in \mathbb{R}^n$, we have

$$(7) \qquad (\overline{\text{co}}f)(x) = \sup\{g(x) \ \mid \ g : \mathbb{R}^n \to (-\infty, +\infty] \text{ closed convex}, \ g \leq f\}.$$

More interesting to note is the following:

$$(8) \qquad (\overline{\text{co}}f)(x) = \sup\{\langle s, x\rangle - b \mid \langle s, y\rangle - b \ \leq \ f(y) \quad \text{for all } y \in \mathbb{R}^n\},$$

or, equivalently, $\overline{\text{co}}f$ is the Legendre–Fenchel biconjugate of f:

$$(9) \qquad \overline{\text{co}}f \ = \ f^{\star\star}$$

(Recall that the Legendre–Fenchel conjugate $f^\star$ of f is defined by $s \in \mathbb{R}^n \mapsto f^\star(s) = \sup_x[\langle s, x\rangle - f(x)]$ and that $f^{\star\star}$ stands for $(f^\star)^\star$.)

In view of (5), it is clear that $\text{dom}(\text{co}f) = \text{co}(\text{dom}\, f)$. By construction, the closure of the convex function $\text{co}f$ is precisely $\overline{\text{co}}f$. Thus both functions coincide at least on the relative interior of $\text{co}(\text{dom}\, f)$. In many cases, we have $\text{co}f = \overline{\text{co}}f$ (cf. §3 below, for example), but there are important instances where $\text{co}f$ is not a closed function.

The convex-hull and closed-convex-hull operations are *global* in the sense that they require to know—a priori—the behavior of f on the whole of $\mathbb{R}^n$. That is the main source of difficulties in the calculation of $(\overline{\text{co}}f)(x)$. Actually, the calculus rules on the closed-convex-hull operation rely mainly on (9), twice using the transformation rules for the Legendre–Fenchel conjugacy.

2.2. The subdifferential theory. At a point x where f is finite, the *subdifferential* $\partial f(x)$ of f at x is the set of $s \in \mathbb{R}^n$ satisfying

$$f(y) \geq f(x) + \langle s, y - x\rangle \quad \text{for all } y \in \mathbb{R}^n. \tag{10}$$

From this definition and the construction (8) of $\overline{\text{co}}f$, it is clear that $\partial(\overline{\text{co}}f)(x) = \partial f(x)$ whenever $(\overline{\text{co}}f)(x) = f(x)$.

The following are classical results (cf. [14, Part V] or [9, Chap. X, §1.4]) and will be used in what follows. First,

$$\partial f(x) \text{ nonempty} \Longrightarrow (\overline{\text{co}}f)(x) = f(x); \tag{11}$$

$$\begin{matrix}(\overline{\text{co}}f)(x) = f(x) \text{ and} \\ f \text{ finite in a neighborhood of } x\end{matrix} \Longrightarrow \partial f(x) \text{ is nonempty.} \tag{12}$$

In particular, if f is Gâteaux differentiable at x, having a nonempty $\partial f(x)$ is equivalent to having $(\overline{\text{co}}f)(x) = f(x)$, and in such a case,

$$\partial f(x) = \partial(\overline{\text{co}}f)(x) = \{\nabla f(x)\}. \tag{13}$$

Consider the example $x \in \mathbb{R} \mapsto f(x) = e^{-x^2}$ and realize that one may have $(\overline{\text{co}}f)(x) < f(x)$ (i.e., $\partial f(x) = \emptyset$) for all x.

If f is convex, $\partial f(x)$ is nonempty for (at least) all the x's in the relative interior of $\operatorname{dom} f$. When f is not convex, $\partial f(x) \neq \emptyset$ occurs at very peculiar points x of $\operatorname{dom} f$ (as seen from (11) and (12)).

An equivalent way of expressing (10) is

$$s \in \partial f(x) \text{ if and only if } f^\star(s) + f(x) - \langle s, x\rangle = 0 \text{ (or } \leq 0). \tag{14}$$

This is a very useful characterization of the elements of $\partial f(x)$, especially since $(\overline{\text{co}}f)^\star = f^\star$. Concerning the minimization of f and $\overline{\text{co}}f$, we recall that

$$\begin{aligned} \inf_{y \in \mathbb{R}^n} f(y) &= \inf_{y \in \mathbb{R}^n} (\overline{\text{co}}f)(y) \quad \text{and} \\ \operatorname{Argmin}(\overline{\text{co}}f) &:= \{x \in \mathbb{R}^n \mid (\overline{\text{co}}f)(x) = \inf_{y \in \mathbb{R}^n} (\overline{\text{co}}f)(y)\} \\ &= \{x \in \mathbb{R}^n \mid 0 \in \partial(\overline{\text{co}}f)(x)\}. \end{aligned} \tag{15}$$

Finally, for differentiability purposes, we keep in mind the following:

$$\begin{matrix}\text{co}f \text{ and } \overline{\text{co}}f \text{ coincide on the interior of } \text{co}(\operatorname{dom} f), \text{and } \text{co}f \text{ is} \\ \text{differentiable at } x \in \operatorname{int}(\text{co}(\operatorname{dom} f)) \text{ if and only if } \partial(\text{co}f)(x) \\ \text{contains a single element (which is therefore } \nabla(\text{co}f)(x)).\end{matrix} \tag{16}$$

3. The subdifferential of $\overline{\text{co}}f$ for 1-coercive f.

In this section, we assume that $f : \mathbb{R}^n \to (-\infty, +\infty]$ satisfies (1) and

$$f \text{ is closed on } \mathbb{R}^n \text{(i.e., epi} f \text{ is a closed set of } \mathbb{R}^{n+1}); \tag{17}$$

$$f \text{ is 1-coercive on } \mathbb{R}^n, \quad \text{that is,} \quad \lim_{\|x\| \to +\infty} \frac{f(x)}{\|x\|} = +\infty. \tag{18}$$

As far as the closed-convex-hull operation is concerned, assumption (17) is not severe: indeed, both f and its closure $\overline{f}$ yield the same closed convex hull ($\overline{\text{co}}f = \overline{\text{co}}\overline{f}$).

Assumption (18) (which, in particular, implies (2) for closed functions) is more restrictive and will be removed in the next section. There are, however, many situations where it holds.

Example 3.1. Let S be a nonempty compact subset of $\mathbb{R}^n$ and $f : S \to \mathbb{R}$ a continuous function. Then the function f, extended to the whole $\mathbb{R}^n$ setting $f(x) = +\infty$ if $x \notin S$, satisfies (17) and (18). This is actually the most frequent case in applications where $\overline{\text{co}}f$ must be explicitly calculated (see [10, Chap. IV, §4.3] and [15]).

Example 3.2. Let f satisfy only (1) and (17). Then the "restricted" function

$$f_k : x \in \mathbb{R}^n \mapsto f(x) \quad \text{if } \|x\| \leq k, \quad +\infty \text{ if not,}$$

verifies (for k large enough) assumptions (1), (17), and (18). Indeed, $\overline{\text{co}}f_k$ ($=\text{co}f_k$ as will seen below) is an approximation of co$f$ since $(\overline{\text{co}}f_k)_k$ is a decreasing sequence of closed convex functions converging to co$f$ when $k \to +\infty$ (cf. (5)). The first important property of 1-coercive functions is the following.

LEMMA 3.3. *Let f satisfy* (1), (17), *and* (18). *Then the following hold:*

(i) co(epi$f$) *is a closed set.*

(ii) *For any $x \in$ dom(co$f$) $=$ co(dom$f$), there are $x_i \in$ dom f and $(\alpha_1, \ldots, \alpha_{n+1}) \in \Delta_{n+1}$ such that*

$$x = \sum_{i=1}^{n+1} \alpha_i x_i \quad \text{and} \quad (\text{co}f)(x) = \sum_{i=1}^{n+1} \alpha_i f(x_i). \tag{19}$$

Part (i) of the lemma is due to Valadier [17, p. 69]; a more readable and detailed proof can be found in [9, Chap. X, §1.5]. Part (ii) is simply a consequence of (i): it suffices to express $(x, (\text{co}f)(x))$—which lies on the boundary of co(epi$f$)—as a convex combination of $n+1$ elements (x_i, y_i) in epi$f$ and to realize that the y_i's have to be $f(x_i)$. To directly prove that the infimum in the definition of $(\text{co}f)(x)$ in (5) is achieved may be arduous, as exemplified in [8, Thm. 2.1].

We are thus in a situation where $\overline{\text{co}}f = \text{co}f$. Given $x \in$ dom(co$f$) $=$ co(dom f), we dub "called by x" the subfamily $\{x_i\}_{i \in I}$ of the family $\{x_1, \ldots, x_{n+1}\}$ as described in Lemma 3.3(ii) corresponding to strictly positive α_i's ($i \in I$ whenever $\alpha_i > 0$). Where do the x_i's called by a given x lie? As a general rule, this is very difficult to answer. In view of their definition, however, we can say that

$$\begin{gathered} \text{the } x_i\text{'s called by } x \text{ belong to dom} f \\ \text{and} \\ \text{to the smallest face of co(dom} f\text{)containing } x. \end{gathered} \tag{20}$$

For example, if x is an extreme point of co(dom f), then $x \in$ dom f and the smallest face containing x is $\{x\}$; hence the only possibility for x is to call itself. Consequently, $(\text{co}f)(x) = f(x)$.

Note also that, due to the 1-coerciveness assumption on f, the x_i's called by x remain in a compact set when the x's lie in a compact set.

Remark 3.4. Even if the original function f has a very "regular" behavior on (a convex) dom $f(:= C)$, the resulting co$f$ may not be continuous on C. To see this, let us recall the counterexample by Kruskal [12]. In $\mathbb{R}^3$, let C be the convex hull of the circle $\{(a,b,c) \mid c = 0 \text{ and } a^2 + (b-1)^2 = 1\} =: \Gamma$ and the line segment $\{(0,0,c) \mid -1 \leq c \leq 1\} =: L$ (see Fig. 1). Now let $f : (a,b,c) \in \mathbb{R}^3 \mapsto f(a,b,c) = -c^2$ if $(a,b,c) \in C$ and $+\infty$ if not. Clearly, f satisfies all the assumptions invoked in this section. What is $(\mathrm{co} f)(x)$, $x = (a, b, c)$? If $x \in \Gamma \setminus \{(0,0,0)\}$, $(\mathrm{co} f)(x) = f(x) = 0$; if $x = (0,0,1)$ or $(0,0,-1)$, $(\mathrm{co} f)(x) = f(x) = -1$. Let $x = (0,0,0)$. The smallest face of C containing x is the line segment L, and the x_i's called by such an x are $x_1 = (0,0,1)$ and $x_2 = (0,0,-1)$; whence $(\mathrm{co} f)(x) = (1/2)[f(x_1) + f(x_2)] = -1$. Thus $\mathrm{co} f$ is not continuous on Γ (contained in the boundary of C). We know that $\mathrm{co} f$, being a convex function, is continuous on the interior of C. As for the differentiability properties, the answer will be provided by a subdifferential relationship linking ∂f to $\partial(\mathrm{co} f)$.

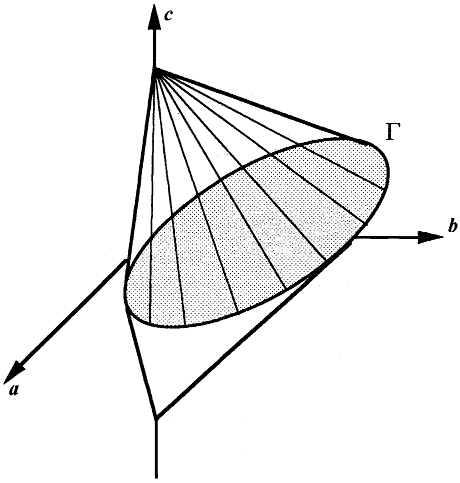


FIG. 1.

LEMMA 3.5. *For a given $x \in \mathrm{co}(\mathrm{dom} f)$, consider a family $\{x_i\}_{i\in I}$ called by x. Then the following hold:*

(i) *$f(x_i) = (\mathrm{co} f)(x_i)$ for all $i \in I$;*

(ii) *$\mathrm{co} f$ is affine on the compact convex polyhedron $\mathrm{co}\{x_i \mid i \in I\}$.*

Proof. See [8, p. 697] or [9, Chap. 10, §1.5]. □

The fact that $\mathrm{co} f$ is affine on $\mathrm{co}\{x_i \mid i \in I\}$ is easy to imagine and visualize (for $n = 1$ or 2). The coincidence property (i) shows that the points x_i that are called are very particular: not all the points x_i, even those satisfying (20), are going to be called.

We now are ready to obtain the subdifferential formula of $\mathrm{co} f$ from that of f.

THEOREM 3.6. *Let f satisfy (1), (17), and (18). For a given $x \in \mathrm{co}(\mathrm{dom} f)$, consider a collection of $\{x_i\}_{i\in I}$ called by x. Then*

$$\partial(\mathrm{co} f)(x) = \bigcap_{i\in I} \partial f(x_i). \tag{21}$$

For any $s \in \partial(\mathrm{co} f)(x)$,

$$\langle s, x\rangle - (\mathrm{co} f)(x) = \langle s, x_i\rangle - f(x_i) \quad \text{for all } i \in I. \tag{22}$$

Proof. See [9, Chap. X, §1.5]; a proof of a different kind and covering a more general situation will be given later (the proof of Theorem 4.6 in §4). □

Relation (22) expresses an "equilibrium property" achieved between $(\mathrm{co} f)(x)$ and the $f(x_i)$'s at the x_i's that the point x calls; it reminds us of the affinity property of $\mathrm{co} f$ on the convex compact polyhedron generated by the x_i's (cf. Lemma 3.5(ii)). Actually, the following explains (22): if $\partial(\mathrm{co} f)(x)$ is nonempty and $s \in \partial(\mathrm{co} f)(x)$, then $(x, (\mathrm{co}f)(x))$ belongs to the subset of epi(co$f$) that maximizes the linear form $\langle (s, -1), . \rangle$.

Comment 3.7. 1. It is surprising that f enters the subdifferential formula (21) via its subdifferential in the sense of convex analysis (i.e., the stringent condition (10)) since we know that for a nonconvex function f, this subdifferential is empty more often than not. However, the x_i's considered are "called points," and at such points f and co$f$ coincide, so there is a good chance that the $\partial f(x_i)$'s are nonempty (cf. (12)).

2. If x lies in the relative interior of co(dom f), the subdifferential of co$f$ at x is nonempty, which implies (by (21)) that the subdifferential of f at the x_i's called by x is necessarily nonempty. This again restricts the set of possible candidates $x_i \in \mathrm{dom}\, f$ to be called.

3. In view of subdifferential theory as recalled in §2.2 (especially (16)), the differentiability of co$f$ at $x \in \mathrm{int}(\mathrm{co}(\mathrm{dom}\, f))$ is secured whenever $\partial f(x_i)$ contains a single element for just one of the x_i's called by x. In fact, all of the conditions on the geometry of dom f near x_i and the behavior of f around x_i, which have been carefully studied by Griewank and Rabier in [8, §3] to ensure the differentiability of co$f$ on int(co(dom f)), are sufficient for $\partial(\mathrm{co}f)$ to be single valued; formula (21) clearly shows what to expect about $\partial(\mathrm{co}f)(x)$ from information about $\partial f(x_i)$ at points x_i called by x. To illustrate, consider f (satisfying (1), (17), and (18)) such that

$$\text{(23)}\quad \begin{array}{l}\partial f(x_i) \text{ is empty for all } x_i \text{ on the boundary of } \mathrm{dom}\, f, \text{and } \partial f(x_i)\\ \text{is a singleton at any } x_i \in \mathrm{int}(\mathrm{dom}\, f) \text{ where } \partial f(x_i) \text{ is nonempty}\end{array}$$

(for example, if dom f is open and f is Gâteaux differentiable on dom f). Then it immediately follows from (21) that co$f$ is differentiable (and, therefore, continuously differentiable) on the interior of its domain.

4. For f satisfying (1), (17), and (18), the set Argmin$f$ of its minimum points is a nonempty compact set. It then follows from (15) and (21) that

$$\text{(24)}\qquad \mathrm{Argmin}(\mathrm{co}f) = \mathrm{co}(\mathrm{Argmin}f).$$

5. Even if we impose greater regularity on f (say C^k, $k \geq 2$), co$f$ as a rule is not C^2. However, if f is locally $C^{1,\alpha}$, $0 < \alpha \leq 1$, then so is co$f$ (under some additional technical assumptions on f that are explained in detail in [8, §4]). See also [13, Prop. 3.1] for an ellipticity condition on $\overline{\mathrm{co}}f$ when f is C^2 on $\mathbb{R}^n$ and [11, §5] for a study of what regularity (between C^1 and C^2) co$f$ must have.

Remark 3.8. In connection with the end of Comment 3.7(3), let us note that even if C is a compact convex set and f a C^∞ function on int$C$, the resulting $\overline{\mathrm{co}}f$ need not be differentiable on int$C$. The following is a counterexample in that respect. Let C be the convex hull of the five points $(1,0), (1,2), (0,3), (-1,2)$, and $(-1,0)$ in $\mathbb{R}^2$ and let $f : (a,b) \in \mathbb{R}^2 \mapsto f(a,b) = 1 - a^2$ if $(a,b) \in C$ and $+\infty$ if not. Then $\overline{\mathrm{co}}f$ is not differentiable on the line segment L joining $(1,2)$ and $(-1,2)$. The reason is that for any x lying in the relative interior of L, the points called by x are $x_1 = (1,2)$ and $x_2 = (-1,2)$ so that $\partial f(x_1)$ $(= \partial(\overline{\mathrm{co}}f)(x_1))$ and $\partial f(x_2)$ $(= \partial(\overline{\mathrm{co}}f)(x_2))$ have more than one element in common. The situation would be different if x_1 and x_2 were "smooth" boundary points of C. See Fig. 2.

4. The subdifferential of $\overline{\mathrm{co}}f$ for epi-pointed f. If we assume only (1) and (17) for f, can we express the subdifferential of $\overline{\mathrm{co}}f$ in terms of that of f only (as in (21))? The answer is, in general, no, as shown by the next example.

Example 4.1. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$(a,b) \in \mathbb{R}^2 \mapsto f(a,b) = \sqrt{a^2 + e^{-b^2}}.$$

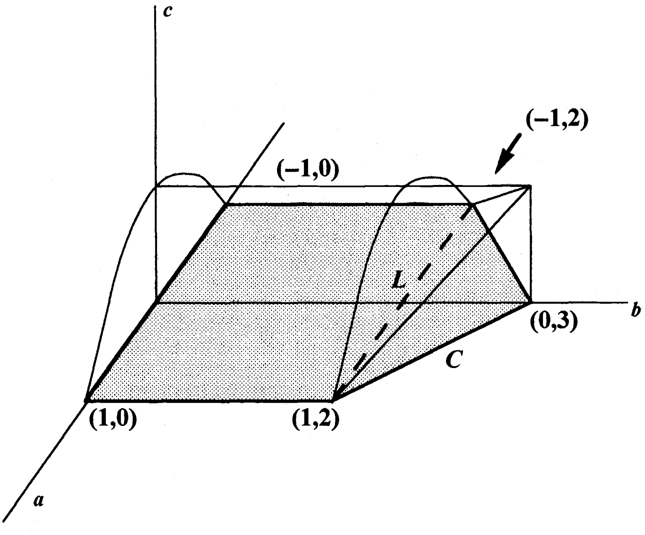


FIG. 2.

f is C^∞ on $\mathbb{R}^2$, and it is easy to verify that

$$(\text{co}f =) \quad \overline{\text{co}}f : (a,b) \longmapsto (\overline{\text{co}}f)(a,b) = |a|.$$

Should a formula like (21) hold true, the differentiability of f would induce that of $\overline{\text{co}}f$, which is not the case.

To obtain a formula similar to (21)—at least for a large class of functions that we call "epi-pointed" (the definition comes later)—we must somehow take into account the behavior of f at infinity. This is done via the asymptotic function of f.

4.1. The asymptotic cone of a set and the asymptotic function of a function. Given a nonempty closed set S (the closedness is not so important since the proposed concepts are blind to the closure operation on sets), the *asymptotic cone* S_∞ of S is defined as follows:

$$S_\infty := \{d \in \mathbb{R}^n \mid \exists (x_k)_k \text{ in } S,\ \exists (t_k)_k \text{ in } \mathbb{R}_+ \text{ with } \lim_{k\to+\infty} t_k = 0 \text{ such that } d = \lim_{k\to+\infty} t_k x_k\}. \tag{25}$$

S_∞ is a *closed cone* (with apex 0), and, moreover, if S is convex, S_∞ coincides with the asymptotic cone (or recession cone or characteristic cone) of S used in the context of convex analysis (cf. [14, §8] or [16, p. 107]). If K is already a nonempty closed cone, then clearly $K_\infty = K$; in particular, $(S_\infty)_\infty = S_\infty$. Before going further, we comment on the introduction and use of asymptotic cones for nonconvex sets. To our knowledge, it is Debreu [3, p. 26] who first proposed the current definition (under a different but equivalent form) of the asymptotic cone and gave (without proofs) some of its uses—for example, sufficient conditions for the sum of closed sets to be closed. Later, the concept was proposed again by Dedieu [4, 5, 6], who studied it in full detail (and even in a more general setting—that of a real topological vector space). The definition has been rediscovered several times since then, and it has proved useful in various areas such as nonconvex optimization, existence theory in variational problems, mathematical economics, etc.

DEFINITION 4.2. *A closed cone K of* $\mathbb{R}^n$ *is said to be* pointed *if*

$$\text{(26)}\qquad \left.\begin{array}{l} m \in \mathbb{N}^\star, \\ c_i \in K \quad \textit{for all } i = 1, \dots, m, \\ \displaystyle\sum_{i=1}^{m} c_i = 0 \end{array}\right\} \Longrightarrow (c_i = 0 \quad \textit{for all } i = 1, \dots, m).$$

Note that as in Carathéodory's theorem (and its proof), we might well limit ourselves to the integers $m \leq n+1$ in (26). Indeed, suppose that there is a family $\{c_i\}_{i \in I} \subset K \setminus \{0\}$ with card$I > n+1$ and $\sum_{i \in I} c_i = 0$ and that property (26) holds true with $1 \leq m \leq n+1$. From the equality $0 = \sum_{i \in I} \frac{1}{\text{card} I} c_i$, with Carathéodory's theorem, we deduce that there exist $J \subset I$ and $\{\beta_j\}_{j \in J}$ such that

$$1 \leq \text{card} J \leq n+1, \qquad \beta_j > 0 \quad \text{for all } j \in J,$$

$$\sum_{j \in J} \beta_j = 1 \quad \text{and} \quad \sum_{j \in J} \beta_j c_j = 0.$$

Now, however, the $\{c'_j := \beta_j c_j\}_{j \in J}$'s form a family of at most $n+1$ elements of K, and property (26) would induce that $c'_j = 0$, and hence $c_j = 0$ for all $j \in J$. This contradicts the initial assumption on c_j, namely, $c_j \in K \setminus \{0\}$.

It is also easy to check that a nonempty closed cone K is pointed if and only if

$$\text{(27)}\qquad (\text{co}K) \cap (-\,\text{co}\,K) = \{0\};$$

this expresses the pointedness property of the convex cone co$K$.

The following result plays a key role in what follows.

PROPOSITION 4.3. *Let S be a nonempty closed set and assume that S_∞ is pointed; then*

(i) $\overline{\text{co}}S = \text{co}S + \text{co}(S_\infty)$ $(= \text{co}(S + S_\infty))$;

(ii) $\text{co}(S_\infty) = (\overline{\text{co}}S)_\infty$ *and thus* $\text{co}(S_\infty)$ *is closed.*

To our knowledge, this statement first appeared in McFadden [7]; see the appendix for an alternate proof.

It is now natural to consider closed functions f and define their asymptotic functions via the asymptotic cones of their epigraphs. Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ satisfy (1), (2), and (17) (although all these requirements are not necessary for the definition to hold); then epi$f$ is a nonempty closed set in $\mathbb{R}^{n+1}$ and we can define (geometrically) the *asymptotic function* f_∞ of f by

$$\text{(28)}\qquad \text{epi}(f_\infty) = (\text{epi}f)_\infty.$$

Clearly, f_∞ is a *positively homogeneous closed function* satisfying $f_\infty(0) = 0$. Indeed, if f is minimized by the affine function $\langle s, . \rangle - r$, the asymptotic function f_∞ of f is minimized by the linear function $\langle s, . \rangle$. In fact, we have an *analytical* definition of f_∞ (suggested in [5, p. 943]) as follows.

PROPOSITION 4.4. *For all $d \in \mathbb{R}^n$,*

$$\text{(29)}\qquad f_\infty(d) = \liminf_{t \to 0^+, d' \to d} t f\left(\frac{d'}{t}\right).$$

Proof. See the appendix. □

At this stage, one may wonder what class of positively homogeneous closed functions one obtains when taking asymptotic functions. Actually, one obtains all of them since $(f_\infty)_\infty = f_\infty$. The following result, whose proof relies on specific approximation and regularization techniques from [1], goes even further: if $g : \mathbb{R}^n \longrightarrow (-\infty, +\infty]$ is a positively homogeneous closed function with a nonempty domain, there exists a C^1 function $f : \mathbb{R}^n \to \mathbb{R}$ such that $f_\infty = g$.

4.2. Epi-pointed closed functions. We say that a function $f : \mathbb{R}^n \to (-\infty, +\infty]$ satisfying (1), (2), and (17) is *epi-pointed* (or asymptotically epi-pointed) when $(\operatorname{epi} f)_\infty$ is pointed. The following gives various characterizations of epi-pointed functions.

PROPOSITION 4.5. *Assume that f satisfies* (1), (2), *and* (17). *Then the epi-pointedness of f is equivalent to one of the following properties:*

(i) *f is minimized by $n+1$ affine functions $\langle s_i, . \rangle - r_i$ with affinely independent slopes s_i*;

(ii) $\operatorname{dom} f^\star$ *has a nonempty interior*;

(iii) *there exist $s \in \mathbb{R}^n$, $\sigma > 0$, and $r \in \mathbb{R}$ such that*

$$f(x) \geq \langle s, x\rangle + \sigma\|x\| - r \quad \textit{for all } x \in \mathbb{R}^n; \tag{30}$$

(iv) *there exists $s \in \mathbb{R}^n$ such that*

$$\liminf_{\|x\| \to +\infty} \frac{f(x) - \langle s, x\rangle}{\|x\|} > 0. \tag{31}$$

Proof. See the appendix. □

The following are easy consequences of the characterizations above (the functions f and g below are assumed to satisfy (1), (2), and (17)):

- If f is epi-pointed, so is $f + \langle s, . \rangle + r$, where $s \in \mathbb{R}^n$ and $r \in \mathbb{R}$.
- If $f \geq g$ and g is epi-pointed ($g = \sigma\|.\| + r$ with $\sigma > 0$, $r \in \mathbb{R}$, for example), then f is epi-pointed.
- If f or g is epi-pointed and if $\operatorname{dom} f \cap \operatorname{dom} g$ is nonempty, then $\max(f, g)$ is epi-pointed (use characterization (iii)).

In spite of the exact formula $h_\infty = \min(f_\infty, g_\infty)$ for $h := \min(f, g)$, the epi-pointedness of f and g does not ensure that of h; however, note that if f is 1-coercive and g is epi-pointed (resp. 1-coercive), then h is epi-pointed (resp. 1-coercive).

- Consider f and g with g epi-pointed and $\operatorname{dom} f \cap \operatorname{dom} g$ nonempty. Then $f + g$ is epi-pointed (use characterization (iii)).
- If S is a nonempty closed set and I_S its indicator function ($I_S(x) = 0$ if $x \in S$, $+\infty$ if not), then $(I_S)_\infty = I_{S_\infty}$, whence S is pointed if and only if I_S is epi-pointed.

4.3. The main result. The theorem below gives an explicit formula for the subdifferential of $\overline{\text{co}} f$ in terms of those of f and f_∞ for the class of epi-pointed functions.

THEOREM 4.6. *Assume that the closed function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is epi-pointed. Then the following hold:*

(i) *For all* $x \in \operatorname{dom}(\overline{\text{co}} f)$*, there are points* $x_1, \ldots, x_p$ *in* $\operatorname{dom} f$*, real numbers* $\alpha_1, \ldots, \alpha_p$ $(p \in \mathbb{N}^\star)$*, and possibly points* $y_1, \ldots, y_q$ *in* $\operatorname{dom} f_\infty \setminus \{0\}$ $(q \in \mathbb{N})$ *such that*

$$(32)\qquad \begin{cases} \alpha_i > 0 \quad \textit{for all } i = 1, \ldots, p, \quad \sum_{i=1}^{p} \alpha_i = 1; \\ x = \sum_{i=1}^{p} \alpha_i x_i + \sum_{j=1}^{q} y_j; \\ (\overline{\text{co}} f)(x) = \sum_{i=1}^{p} \alpha_i f(x_i) + \sum_{j=1}^{q} f_\infty(y_j). \end{cases}$$

Moreover, one may choose a decomposition of the above type with

$$(33)\qquad q \leq n \quad \textit{and} \quad p + q \leq n + 1.$$

(ii) *For any decomposition of the type described in* (i)*, we have*

$$(34)\qquad \partial(\overline{\text{co}} f)(x) = \left[\bigcap_{i=1}^{p} \partial f(x_i)\right] \bigcap \left[\bigcap_{j=1}^{q} \partial f_\infty(y_j)\right].$$

Proof. See the appendix. □

In the above statements, $q = 0$ means that there is no y_j in the decomposition of x. In accordance with our earlier appellation, we term "called by x" a family $\{x_i, y_j\}$ of points as described in (i). We exclude the null y_j's because they add nothing in (32) (since $f_\infty(0) = 0$) or (34) (since, as seen in the course of the proof, $\partial f(x_i) \subset \partial f_\infty(0)$ for any $x_i \in \operatorname{dom} f$ and $\partial f_\infty(y_j) \subset \partial f_\infty(0)$ for any $y_j \in \operatorname{dom} f_\infty$). A given $x \in \operatorname{dom}(\overline{\text{co}} f)$ may call only x_i points or both x_i and y_j points for (32) to hold.

The 1-coercive case, recalled in §3, falls into this more general one. Indeed, if f is closed and 1-coercive, then f is epi-pointed (use Proposition 4.5(iv) with $s = 0$ or Proposition 4.5(ii) since $\operatorname{dom} f^\star = \mathbb{R}^n$ in that case) and $f_\infty = I_{\{0\}}$, i.e., $f_\infty : d \mapsto f_\infty(d) = 0$ if $d = 0$, $+\infty$ if not. Thus $\operatorname{epi}(\overline{\text{co}} f) = \operatorname{co}(\operatorname{epi} f) + \{0\} \times \mathbb{R}_+ = \operatorname{co}(\operatorname{epi} f)$, whence we recover $\overline{\text{co}} f = \text{co} f$. The points $x \in \operatorname{dom}(\text{co} f)$ are necessarily of the x_i type. Before going further, let us illustrate Theorem 4.6 with a simple example.

Example 4.7. Let $f : \mathbb{R} \to \mathbb{R}$ be defined as follows:

$$f(x) = \begin{cases} \sqrt{|x+1|} & \text{if } x \leq -1, \\ 1 - |x| & \text{if } -1 \leq x \leq 1 \\ (x-1)(1+e^{-x}) & \text{if } x \geq 1. \end{cases} \quad \text{(see Fig. 3)},$$

Then $f_\infty(d) = d^+$, and $(\overline{\text{co}} f)(x) = 0$ if $x \leq 1$ and $x - 1$ if $x \geq 1$. The assumptions of Theorem 4.6 are obviously satisfied by f. The point $x = 2$ can call only $x_1 = 1$ and $y_1 = 1$; as a result (see (34)),

$$(\{1\} =) \quad \partial(\overline{\text{co}} f)(x) = \partial f(x_1) \cap \partial f_\infty(y_1) = [0, 1] \cap \{1\}.$$

This also shows that the bounds in (33) are sharp. On the other hand, $x = 0$ can call only $x_1 = 1$ and $x_2 = -1$; then

$$(\{0\} =) \quad \partial(\overline{\text{co}} f)(x) = \partial f(x_1) \cap \partial f(x_2) = [0, 1] \cap \{0\}.$$

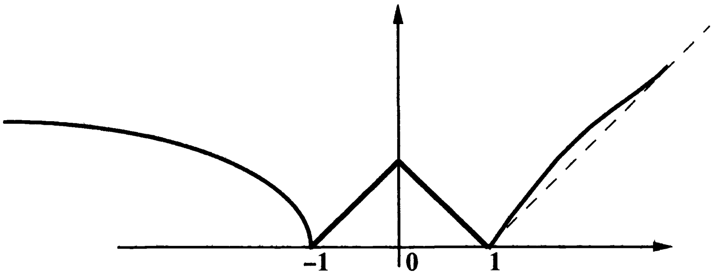


FIG. 3.

Comment 4.8 (under the assumptions of Theorem 4.6). 1. Since $(\overline{\text{co}}(\text{epi}f))_\infty$ equals $\text{co}((\text{epi}f)_\infty)$ (cf. Proposition 4.3(ii)), its "functional" counterpart is:

$$(\overline{\text{co}}f)_\infty = \text{co}(f_\infty) \quad (= \overline{\text{co}}(f_\infty)). \tag{35}$$

It goes without saying that f_∞ itself satisfies the assumptions of Theorem 4.6: it is closed and epi-pointed.

2. Similarly to the 1-coercive case (cf. Lemma 3.5), we note the following properties of $\overline{\text{co}}f$: if $x_1, \dots, x_p$ and $y_1, \dots, y_q$ are points called by a given $x \in \text{dom}(\overline{\text{co}}f)$, then

$$\begin{gathered} f(x_i) = (\overline{\text{co}}f)(x_i) \quad \text{for all } i = 1, \dots, p \\ \text{and} \\ f_\infty(y_j) = (\text{co}f_\infty)(y_j) \quad \text{for all } j = 1, \dots, q; \end{gathered} \tag{36}$$

$$\begin{gathered} \overline{\text{co}}f \text{ is affine on the closed convex polyhedron} \\ P = \text{co}\{x_1, \dots, x_p\} + \mathbb{R}_+ y_1 + \cdots + \mathbb{R}_+ y_q. \end{gathered} \tag{37}$$

To prove (36), we first note that

$$(\overline{\text{co}}f)(x) \le (\overline{\text{co}}f)\left(\sum_{i=1}^{p} \alpha_i x_i\right) + (\overline{\text{co}}f)_\infty \left(\sum_{j=1}^{q} y_j\right). \tag{38}$$

(This is a property of asymptotic functions for convex functions; see [14, p. 66].) Second, $\overline{\text{co}}f$ is convex and $(\overline{\text{co}}f)_\infty$ subadditive, so

$$(\overline{\text{co}}f)(x) \le \sum_{i=1}^{p} \alpha_i \, (\overline{\text{co}}f)(x_i) + \sum_{j=1}^{q} (\overline{\text{co}}f)_\infty(y_j). \tag{39}$$

Now $\overline{\text{co}}f$ (resp. $(\overline{\text{co}}f)_\infty$ ($= \text{co}f_\infty$ from (35))) minimizes f (resp. f_∞); hence the right-hand side of (39) is maximized by $\sum_{i=1}^p \alpha_i f(x_i) + \sum_{j=1}^q f_\infty(y_j) = (\overline{\text{co}}f)(x)$. Combining this with (39) yields (36).

To prove (37), we let $P = \text{co}\{x_1, \dots, x_p\} + \mathbb{R}_+ y_1 + \cdots + \mathbb{R}_+ y_q$; x lies in the relative interior *riP* of P. (Recall that $riP = \{\sum_{i=1}^p \alpha_i' x_i + \sum_{j=1}^q \beta_j' y_j : \alpha_i' > 0 \text{ for all } i, \sum_{i=1}^p \alpha_i' = 1, \beta_j' > 0 \text{ for all } j\}$.) Consider $g(x) = (\overline{\text{co}}f)(x)$ if $x \in P$ and $+\infty$ otherwise. It is a closed convex function whose domain is P; indeed,

$$\begin{aligned} \left(\overline{\text{co}}f)(\sum_{i=1}^{p} \alpha_i' x_i + \sum_{j=1}^{q} \beta_j' y_j\right) &\le (\overline{\text{co}}f)\left(\sum_{i=1}^{p} \alpha_i' x_i\right) + (\overline{\text{co}}f)_\infty \left(\sum_{j=1}^{q} \beta_j' y_j\right) \\ &\le \sum_{i=1}^{p} \alpha_i' (\overline{\text{co}}f)(x_i) + \sum_{j=1}^{q} \beta_j' (\overline{\text{co}}f)_\infty(y_j). \end{aligned}$$

Since $x \in ri\,\mathrm{dom}\, g$, we can choose s in the nonempty $\partial g(x)$ and define $\ell(x') := (\overline{\mathrm{co}}f)(x) + \langle s, x' - x\rangle$.

Step 1. The affine function ℓ minorizes $(\overline{\mathrm{co}}f)$ on P and coincides with $(\overline{\mathrm{co}}f)$ at x.

Step 2. Since $\ell \leq \overline{\mathrm{co}}f$ on P, we easily deduce that

$$\begin{cases} \ell(x_i) \leq \overline{\mathrm{co}}f(x_i) & \text{for all } i \in \{1,\dots,p\}; \\ \ell_\infty(y_j) = \langle s, y_j\rangle \leq (\overline{\mathrm{co}}f)_\infty(y_j) & \text{for all } j \in \{1,\dots,q\}. \end{cases}$$

Step 3. From the following chain of equalities and inequalities,

$$\begin{aligned} \overline{\mathrm{co}}f(x) &= \ell(x) \\ &= \sum_{i=1}^{p} \alpha_i \ell(x_i) + \sum_{j=1}^{q} \langle s, y_j\rangle && \text{[since } \ell \text{ is affine]} \\ &\leq \sum_{i=1}^{p} \alpha_i (\overline{\mathrm{co}}f)(x_i) + \sum_{j=1}^{q} (\overline{\mathrm{co}}f)_\infty(y_j) && \text{[see Step 2]} \\ &= \sum_{i=1}^{p} \alpha_i f(x_i) + \sum_{j=1}^{q} f_\infty(y_j) && \text{[from (36)]} \\ &= \overline{\mathrm{co}}f(x) && \text{[from (32)],} \end{aligned}$$

we deduce that

$$\begin{cases} \ell(x_i) = \overline{\mathrm{co}}f(x_i) & \text{for all } i \in \{1,\dots,p\}; \\ \langle s, y_j\rangle = (\overline{\mathrm{co}}f)_\infty(y_j) & \text{for all } j \in \{1,\dots,q\}. \end{cases}$$

Step 4. Let $x' = \sum_{i=1}^{p} \alpha'_i x_i + \sum_{j=1}^{q} \beta'_j y_j \in P$. We have

$$\begin{aligned} \overline{\mathrm{co}}f(x') &\leq (\overline{\mathrm{co}}f)\left(\sum_{i=1}^{p} \alpha'_i x_i\right) + (\overline{\mathrm{co}}f)_\infty \left(\sum_{j=1}^{q} \beta'_j y_j\right) \\ &\leq \sum_{i=1}^{p} \alpha'_i (\overline{\mathrm{co}}f)(x_i) + \sum_{j=1}^{q} \beta'_j (\overline{\mathrm{co}}f)_\infty(y_j) \\ &= \sum_{i=1}^{p} \alpha'_i \ell(x_i) + \sum_{j=1}^{q} \beta'_j \langle s, y_j\rangle && \text{[see Step 3]} \\ &= \ell(x') && \text{[since } \ell \text{ is affine]} \\ &\leq \overline{\mathrm{co}}f(x'). \end{aligned}$$

Then all of these inequalities are equalities and, in particular, $\ell = \overline{\mathrm{co}}f$ on P.

3. Concerning the comparison of $\mathrm{Argmin}(\overline{\mathrm{co}}f)$ with similar sets associated with f and f_∞, we have the following:

$$\begin{aligned} \mathrm{Argmin}(\overline{\mathrm{co}}f) &= \mathrm{co}(\mathrm{Argmin} f) + \mathrm{co}(\mathrm{Argmin} f_\infty) \\ &\quad (= \mathrm{co}(\mathrm{Argmin} f + \mathrm{Argmin} f_\infty)). \end{aligned} \tag{40}$$

For this, it suffices to combine the characterization (15) of $\mathrm{Argmin}(\overline{\mathrm{co}}f)$ with formula (34). When f is 1-coercive, $\mathrm{Argmin} f_\infty = \{0\}$ and we obtain (24).

4. The "equilibrium property" (22) of Theorem 3.6 has also its counterpart: if $x_1, \dots, x_p$ and $y_1, \dots, y_q$ are points called by a given $x \in \mathrm{dom}(\overline{\mathrm{co}}f)$, then for any $s \in \partial(\overline{\mathrm{co}}f)(x)$,

$$f_\infty(y_j) = \langle s, y_j\rangle \quad \text{for all } j = 1, \dots, q \tag{41}$$

and

$$\langle s, x\rangle - (\overline{\text{co}} f)(x) = \langle s, x_i\rangle - f(x_i) \quad \text{for all } i = 1, \dots, p. \tag{42}$$

To see that, observe that $s \in \partial f_\infty(y_j)$ $(= \partial(\text{co} f_\infty)(y_j))$ implies

$$s \in \partial f_\infty(0)\ (= \partial(\text{co} f_\infty)(0)) \quad \text{and} \quad f_\infty(y_j) = \langle s, y_j\rangle \quad \text{for all } j = 1, \dots, q. \tag{43}$$

(This is due to the fact that $\text{co} f_\infty$ $(= (\overline{\text{co}} f)_\infty)$ is a closed positively homogeneous function.) Next, $s \in \partial(\overline{\text{co}} f)(x)$ (resp. $s \in \partial f(x_i)$) is characterized by the equality $(\overline{\text{co}} f)(x) - \langle s, x\rangle = -(\overline{\text{co}} f)^\star(s)$ (resp. $f(x_i) - \langle s, x_i\rangle = -f^\star(s)$), whence we have (42) since $(\overline{\text{co}} f)^\star = f^\star$.

COROLLARY 4.9. *In addition to the assumptions of Theorem* 4.6 *on f suppose the following:*

The subdifferential of f is empty on the boundary of $\text{dom}\, f$;
f is Gâteaux differentiable on the interior of $\text{dom}\, f$.

(*A Fréchet-differentiable epi-pointed $f : \mathbb{R}^n \to \mathbb{R}$ satisfies all these requirements*). *Then $\overline{\text{co}} f$ is* (*continuously*) *differentiable on the interior of its domain.*

Indeed, the x_i's called by $x \in \text{int}\,\text{dom}(\overline{\text{co}} f)$—and there necessarily are such x_i's (see Theorem 4.6(i))—lie in the interior of $\text{dom}\, f$ and, at these points, $\partial f(x_i) = \{\nabla f(x_i)\}$. It then follows from (34) that

$$\nabla(\overline{\text{co}} f)(x) = \nabla f(x_i) \quad \text{for all } i = 1, \dots, p. \tag{44}$$

The function f exhibited in Example 4.1 is not epi-pointed. It is minimized by (at most) two affine functions with affinely independent slopes, while three would be necessary to make it epi-pointed (cf. Proposition 4.5(i)). Should the result of Theorem 4.6 hold, the differentiability of f would induce that of $\overline{\text{co}} f$, which is not the case. Thus the epi-pointedness assumption on f cannot be completely removed in Theorem 4.6.

Remark 4.10. This is also the place where the extended subdifferential calculus for nonconvex functions [2] can be useful. Suppose, for example, that we want to convexify the restriction of some closed function $g : \mathbb{R}^n \to (-\infty, +\infty]$ on a closed subset S, i.e., convexify $f = g + I_S : x \mapsto f(x) = g(x)$ if $x \in S$ and $+\infty$ if not. Various conditions on g or S guarantee that f is closed and epi-pointed (the assumptions of Theorem 4.6). At a point x_i called by x, under fairly general assumptions on g and S [2, pp. 95–109],

$$\partial^{cl} f(x_i) \subset \partial^{cl} g(x_i) + N_S^{cl}(x_i)$$

($\partial^{cl} g$ (resp. N_S^{cl}) denote the generalized subdifferential of g (resp. normal cone to S) in Clarke's sense), whence an *outer estimate* of $\partial f(x) \subset \partial f(x_i)$ (since $\partial f(x_i)$ is always included in $\partial^{cl} f(x_i)$) in terms of the behavior of g around x_i (via $\partial^{cl} g(x_i)$) and the geometry of S around x_i (via $N_S^{cl}(x_i)$).

Remark 4.11. As is clear from the proof of Theorem 4.6 (see the appendix), all the conclusions of Theorem 4.6, especially formula (34), hold provided that

$$\overline{\text{co}}(\text{epi} f) = \text{co}(\text{epi} f) + \text{co}(\text{epi} f_\infty) \tag{45}$$

for the function f considered. To ensure this, the epi-pointedness of f was specifically a general assumption. However, a function like $x \in \mathbb{R}^n \mapsto \sqrt{\|x\|}$ satisfies (45) without being epi-pointed.

5. Appendix.

5.1. Proof of Proposition 4.3.

Proof of (i). Let $x \in S$, $d \in S_\infty$; then for some $(x_k)_k$ in S and $(t_k)_k$ in $\mathbb{R}_+$ with $\lim_{k\to+\infty} t_k = 0$,

$$x + d = \lim_{k\to+\infty} (x + t_k x_k) = \lim_{k\to+\infty} ((1 - t_k)x + t_k x_k),$$

whence $x + d \in \overline{\text{co}}S$. Therefore, $S + S_\infty \subset \overline{\text{co}}S$ and

$$\text{co}S + \text{co}S_\infty = \text{co}(S + S_\infty) \subset \overline{\text{co}}S. \tag{46}$$

For the converse inclusion, consider $v \in \overline{\text{co}}S$; then $v = \lim_{k\to+\infty} v_k$ with

$$v_k = \alpha_k^1 x_k^1 + \cdots + \alpha_k^{n+1} x_k^{n+1} \tag{47}$$

for some $\alpha_k = (\alpha_k^1, \ldots, \alpha_k^{n+1}) \in \Delta_{n+1}$ and $x_k^1, \ldots, x_k^{n+1}$ in S.

Subsequencing if necessary, we may assume that, for all i, $\alpha_k^i \to \alpha^i$ when $k \to +\infty$. Since $(\alpha^1, \ldots, \alpha^{n+1}) \in \Delta_{n+1}$, we set $I := \{i \mid \alpha^i > 0\}$ and $J := \{i \mid \alpha^i = 0\}$. Suppose that one of the sequences $(\alpha_k^i x_k^i)_k$ is unbounded. Without loss of generality, we may consider that

(a) $\|\alpha_k^1 x_k^1\| = \max_{i=1,\ldots,n+1} \|\alpha_k^i x_k^i\|$ for all k;

(b) $\|\alpha_k^1 x_k^1\| \to +\infty$ when $k \to +\infty$.

According to (a), each sequence $(\alpha_k^i x_k^i / \|\alpha_k^1 x_k^1\|)_k$ is bounded and may thus be assumed to converge to some d^i which—due to (b)—lies in S_∞. Again from (b), dividing both sides of (47) by $\|\alpha_k^1 x_k^1\|$ and letting $k \to +\infty$ yields $0 = \sum_{i=1}^{n+1} d^i$. However, since $\|d^1\| = 1$, this contradicts the pointedness property of S_∞.

Thus all of the sequences $(\alpha_k^i x_k^i)_k$ are bounded and, subsequencing in an appropriate way, we can pass to the limit in (47) and obtain

$$v = \sum_{i\in I} \alpha^i x_i + \sum_{j\in J} d^j \in \text{co}\, S + \text{co}S_\infty,$$

whence the converse inclusion in (46) is proved.

Proof of (ii). Since the asymptotic cone of $\overline{\text{co}}S$ is the same as that of co$S$, what we have to prove is actually $\text{co}(S_\infty) = (\text{co}S)_\infty$.

From $S \subset \text{co}S$, we infer that $S_\infty \subset (\text{co}S)_\infty$, and since the latter is convex, we have $\text{co}(S_\infty) \subset (\text{co}S)$.

Conversely, let $v \in (\text{co}S)_\infty$. According to (25) and the definition of co$S$, there exist sequences $(t_k)_k$ in $\mathbb{R}_+$, $(\beta_k)_k$ in Δ_{n+1} with $\beta_k = (\beta_k^1, \ldots, \beta_k^{n+1})$, and $(x_k^1)_k, \ldots, (x_k^{n+1})_k$ in S such that

$$\lim_{k\to+\infty} t_k = 0, \qquad v = \lim_{k\to+\infty} t_k(\beta_k^1 x_k^1 + \cdots + \beta_k^{n+1} x_k^{n+1}).$$

Setting $\alpha_k^i = t_k \beta_k^i$ and $v_k = \alpha_k^1 x_k^1 + \cdots + \alpha_k^{n+1} x_k^{n+1}$, we continue the same reasoning as in the proof of (i) above ($I = \emptyset$ here) to get $v = \sum_{j=1}^{n+1} d^j$ with $d^j \in S_\infty$ for all $j \in J$. Whence $v \in \text{co}(S_\infty)$.

As the asymptotic cone of $\overline{\text{co}}S$, $\text{co}(S_\infty)$ is closed. This can also be viewed as a direct consequence of (i): apply the result of (i) to S_∞ and the relation $\overline{\text{co}}(S_\infty) = \text{co}(S_\infty)$ results. □

5.2. Proof of Proposition 4.4. Define

$$g : d \in \mathbb{R}^n \mapsto g(d) = \liminf_{t \to 0^+, d' \to d} tf\left(\frac{d'}{t}\right).$$

To prove $f_\infty = g$, it suffices to verify that epi $g = (\text{epi} f)_\infty$.

First, let $(v, \mu) \in (\text{epi} f)_\infty$; there exist sequences $((x_k, \mu_k))_k$ in epi$f$ and $(t_k)_k$ in $\mathbb{R}_+$ converging to 0 such that $(v, \mu) = \lim_{k \to +\infty} t_k(x_k, \mu_k)$. Therefore, $v = \lim_{k \to +\infty} t_k x_k$ and $\mu = \lim_{k \to +\infty} t_k \mu_k$ with $f(x_k) \leq \mu_k$ for all k so that

$$g(v) \leq \liminf_{k \to +\infty} t_k f\left(\frac{t_k x_k}{t_k}\right) \leq \liminf_{k \to +\infty} t_k \mu_k = \mu.$$

We thus have proved that $(v, \mu) \in$ epi g.

Conversely, let $(v, \mu) \in$ epi g. Following the definition of $g(v)$, there exist $(t_k)_k$ in $\mathbb{R}_+^\star$ converging to 0 and $(v_k)_k$ converging to v such that $(t_k f(v_k/t_k))_k$ converges and $\lim_{k \to +\infty} t_k f(v_k/t_k) \leq \mu$. Given $\epsilon > 0$, we have $t_k f(v_k/t_k) \leq \mu + \epsilon$ for k large enough. Then $(v, \mu + \epsilon) = \lim_{k \to +\infty} t_k(v_k/t_k, (\mu + \epsilon)/t_k)$ with $(v_k/t_k, (\mu + \epsilon)/t_k) \in$ epi$f$ for k large enough. This means that $(v, \mu + \epsilon) \in (\text{epi} f)_\infty$, and since this holds true for all $\epsilon > 0$, we have $(v, \mu) \in (\text{epi} f)_\infty$. □

5.3. Proof of Proposition 4.5.

(*f is epi-pointed*) $\Rightarrow$ (i). We make use of the following classical result on closed convex cones (deduced from [14, Cor. 14.6.1] or [16, §2.10], for example): if K is a nonempty closed convex cone,

$$(48) \qquad (K \text{ pointed}) \Longleftrightarrow (\text{polar cone } K^\circ \text{of } K \text{ has a nonempty interior});$$

$$(49) \qquad (v \in \text{int} K^\circ) \Longleftrightarrow (\langle v, d \rangle < 0 \quad \text{for all nonnull } d \text{ of } K).$$

Set $K = \text{co}(\text{epi} f_\infty)$; K is a pointed closed convex cone (cf. Proposition 4.3(ii)). Since $\text{int} K^\circ \neq \emptyset$ and since we know that $K^\circ \subset \mathbb{R}^n \times \mathbb{R}_-$, we may choose $n+1$ affinely independent s_i's such that $(s_i, -1) \in \text{int} K^\circ$ (a similar device was used in [17, p. 59]). We want to show that s_i is the slope of some affine function minimizing f. For a given i, suppose that for all k there exists $u_k \in$ epi$f$ such that

$$(50) \qquad \langle (s_i, -1), u_k \rangle \geq k.$$

Subsequencing if necessary, we may assume that $(u_k/\|u_k\|)_k$ converges to an element d which by construction lies in $(\text{epi} f)_\infty \setminus \{(0,0)\}$. Dividing (50) by $\|u_k\|$ and letting $k \to +\infty$ yields $\langle (s_i, -1), d \rangle \geq 0$, which contradicts $(s_i, -1) \in \text{int} K^\circ$ (cf. (49)). Thus for all i, there exists $r_i \in \mathbb{R}$ such that

$$\sup_{x \in \mathbb{R}^n} \langle (s_i, -1), (x, f(x)) \rangle = \sup_{u \in \text{epi} f} \langle (s_i, -1), u \rangle \leq r_i.$$

This is written as

$$\langle s_i, x \rangle - r_i \leq f(x) \quad \text{for all } x \in \mathbb{R}^n.$$

(i) $\Leftrightarrow$ (ii). From the definition of $f^\star$, f is minimized by the affine function $\langle s, . \rangle - r$ if and only if the slope s lies in dom $f^\star$. (i) expresses that there are $n+1$ affinely independent points in dom $f^\star$; this means that int(dom $f^\star$) is nonempty.

(ii) $\Rightarrow$ (iii). Let $s \in \operatorname{int}(\operatorname{dom} f^\star)$; we take $\sigma > 0$ such that the compact ball $\bar{B}(s,\sigma)$ is contained in $\operatorname{int}(\operatorname{dom} f^\star)$. Since $f^\star$ is continuous on $\operatorname{int}(\operatorname{dom} f^\star)$ [14, §10], it is bounded above on $\bar{B}(s,\sigma)$; we choose an $r \in \mathbb{R}$ for which

$$f^\star(s') \leq r \quad \text{for all } s' \in \bar{B}(s,\sigma)$$

or, equivalently,

$$-r + \langle s', x\rangle \leq f(x) \quad \text{for all } x \in \mathbb{R}^n \text{ and } s' \in \bar{B}(s,\sigma).$$

By considering $s' = s + \sigma(x/\|x\|)$ if $x \neq 0$ (and any $s' \in \bar{B}(s,\sigma)$ if $x = 0$), we conclude that

$$-r + \langle s, x\rangle + \sigma\|x\| \leq f(x) \quad \text{for all } x \in \mathbb{R}^n.$$

(iii) $\Leftrightarrow$ (iv). This is clear.

(iv) $\Rightarrow$ (*f is epi-pointed*). There is $\sigma > 0$ such that for $\|x\|$ large enough,

$$f(x) \geq \sigma\|x\| + \langle s, x\rangle.$$

The function $x \mapsto \sigma\|x\| + \langle s, x\rangle$ is epi-pointed; hence so is f. □

5.4. Proof of Theorem 4.6.

Proof of (i). Since $(\operatorname{epi} f)_\infty$ is pointed by assumption, it follows from Proposition 4.3 (i) that

$$(\operatorname{epi}(\overline{\operatorname{co}} f) =)\ \overline{\operatorname{co}}(\operatorname{epi} f) = \operatorname{co}(\operatorname{epi} f) + co(\operatorname{epi} f_\infty). \tag{51}$$

For $x \in \operatorname{dom}(\overline{\operatorname{co}} f)$, the point $(x, (\overline{\operatorname{co}} f)(x))$ lies in the epigraph of $(\overline{\operatorname{co}} f)$; according to the decomposition (51), there exist real numbers $\alpha_1, \ldots, \alpha_p$, points $(x_1, r_1), \ldots, (x_p, r_p)$ in $\operatorname{epi} f$, and points $(y_1, \rho_1), \ldots, (y_{q'}, \rho_{q'})$ in $\operatorname{epi} f_\infty$ such that

$$\begin{aligned} &\alpha_i > 0 \quad \text{for all } i = 1, \ldots, p, \quad \sum_{i=1}^p \alpha_i = 1; \\ &(x, (\overline{\operatorname{co}} f)(x)) = \sum_{i=1}^p \alpha_i (x_i, r_i) + \sum_{j=1}^{q'} (y_j, \rho_j) \end{aligned} \tag{52}$$

(remember that $\operatorname{epi} f_\infty$ is a cone). Actually, due to the definition of $(\overline{\operatorname{co}} f)(x)$ as the infimum of those r such that $(x, r) \in \operatorname{epi}(\overline{\operatorname{co}} f)$, each r_i (resp. each ρ_j) has to be $f(x_i)$ (resp. $f_\infty(y_j)$). Now since $f_\infty(0) = 0$, all the null y_j's can be removed in (52); there are thus real numbers $\alpha_1, \ldots, \alpha_p$ ($p \in \mathbb{N}^\star$), points $x_1, \ldots, x_p$ in $\operatorname{dom} f$, and possibly points $y_1, \ldots, y_q$ in $\operatorname{dom} f_\infty \setminus \{0\}$ ($q \in \mathbb{N}$) such that

$$\begin{aligned} &\alpha_i > 0 \quad \text{for all } i = 1, \ldots, p, \quad \sum_{i=1}^p \alpha_i = 1; \\ &(x, (\overline{\operatorname{co}} f)(x)) = \sum_{i=1}^p \alpha_i (x_i, f(x_i)) + \sum_{j=1}^{q} (y_j, f_\infty(y_j)). \end{aligned} \tag{53}$$

Hence the first part of Theorem 4.6(i) is proved.

To continue, we need the following variant of Carathéodory's theorem.

LEMMA 5.1. *Let $x_1, \dots, x_r$ be r points of $\mathbb{R}^n$, and let x be on the boundary of the closed convex cone $\mathbb{R}_+x_1 + \cdots + \mathbb{R}_+x_r$. Then there exists $\Gamma(x) \subset \{1, \dots, r\}$ with* $\mathrm{card}\Gamma(x) \leq n-1$ *such that* $x \in \sum_{\gamma \in \Gamma(x)} \mathbb{R}_+x_\gamma$.

Proof. Because x lies on the boundary of $K := \mathbb{R}_+x_1 + \cdots + \mathbb{R}_+x_r$, there exists a hyperplane $H_{s,r}$ supporting K at x: for some $s \neq 0$ and $r \in \mathbb{R}$,

$$\langle s, x \rangle - r = 0, \tag{54}$$

$$\langle s, d \rangle - r \leq 0 \quad \text{for all } d \in K. \tag{55}$$

Relabelling the indices if necessary, we may assume that x is a positive combination of the first δ elements $x_1, \dots, x_\delta$:

$$x = \alpha_1 x_1 + \cdots + \alpha_\delta x_\delta, \quad \alpha_i > 0 \quad \text{for all } i = 1, \dots, \delta \text{ and } \delta \leq r.$$

Setting $d = x_i$ successively in (55), using (54), we obtain

$$0 = \langle s, x \rangle - r = \sum_{i=1}^{\delta} \alpha_i [\langle s, x_i \rangle - r] \leq 0,$$

so each $\langle s, x_i \rangle - r$ is actually 0: each x_i is in $H_{s,r}$. Then Carathéodory's theorem for cones tells us that

$$x \in (\mathbb{R}_+x_1 + \cdots + \mathbb{R}_+x_\delta) \cap H_{s,r}$$

can be described as positive combination of only $n-1$ elements x_i. □

We begin with a decomposition of $(x, (\overline{\mathrm{co}}f)(x))$ as in (53),

$$(x, (\overline{\mathrm{co}}f)(x)) = \sum_{i=1}^{p} \alpha_i (x_i, f(x_i)) + \sum_{j=1}^{q} (y_j, f_\infty(y_j)),$$

and we set (in $\mathbb{R}^{n+2}$) $X = (x, (\overline{\mathrm{co}}f)(x), 1)$, $X_i = (x_i, f(x_i), 1)$ for $i = 1, \dots, p$, and, possibly, $X_{p+j} = (y_j, f_\infty(y_j), 0)$ for $j = 1, \dots, q$. It is clear that X lies on the boundary of the closed convex cone $\sum_{i=1}^{p+q} \mathbb{R}_+X_i$ (of $\mathbb{R}^{n+2}$). Thus according to Lemma 5.1, there exists an index set $\Gamma(X) = I \cup J$ ($I \subset \{1, \dots, p\}$ and $J \subset \{p+1, \dots, p+q\}$) with $\mathrm{card}\Gamma(X) \leq n+1$ such that

$$X = \sum_{i \in I} \mathbb{R}_+X_i + \sum_{j \in J} \mathbb{R}_+X_j.$$

This yields

$$(x, (\overline{\mathrm{co}}f)(x)) = \sum_{i \in I} \alpha_i (x_i, f(x_i)) + \sum_{j \in J} \beta_j (y_{j-p}, f_\infty(y_{j-p}))$$

$$\text{with } \sum_{i \in I} \alpha_i = 1, \qquad \alpha_i \geq 0 \quad \text{for all } i \in I, \quad \beta_j \geq 0 \quad \text{for all } j \in J.$$

However, since f_∞ is positively homogeneous, we can rewrite

$$\beta_j (y_{j-p}, f_\infty(y_{j-p})) = (y'_j, f_\infty(y'_j)),$$

where $y'_j = \beta_j y_{j-p}$ for all $j \in J$.

As for the bound on q, it suffices to note that $\sum_{j \in J}(y'_j, f_\infty(y'_j))$ lies on the boundary of $\mathrm{co}(\mathrm{epi} f_\infty)$ and again apply Lemma 4.12.

Proof of (ii). Recall that for a nonempty $S \subset \mathbb{R}^n$ and $u \in S$, the normal cone $N_S(u)$ to S at u is defined as the set of d such that $\langle d, v-u\rangle \leq 0$ for all $v \in S$. Its relationship with the subdifferential of a function is as follows:

$$s \in \partial g(x) \quad \text{if and only if } (s, -1) \in N_{\mathrm{epi}\, g}((x, g(x))). \tag{56}$$

For normal cones, the following calculus rules are useful.

LEMMA 5.2.

(i) *If* $u_1 \in S_1$ *and* $u_2 \in S_2$, *then*

$$N_{S_1+S_2}(u_1+u_2) = N_{S_1}(u_1) \cap N_{S_2}(u_2).$$

(ii) *For* $u \in \mathrm{co}\, S$ ($u = \sum_{i=1}^r \alpha_i u_i$ *with* $\sum_{i=1}^r \alpha_i = 1$, $\alpha_i > 0$, *and* $u_i \in S$ *for all* $i = 1, \dots, r$),

$$N_{\mathrm{co}\, S}(u) = \bigcap_{i=1}^r N_S(u_i).$$

(iii) *For a closed subset* S *and any* $u \in S$, *we have*

$$N_S(u) \subset N_{S_\infty}(0).$$

Proof. (i) This is immediate from the definition of normal cones.

(ii) First, note that $N_{\mathrm{co}\, S}(u) = N_S(u)$. If $d \in \bigcap_{i=1}^r N_S(u_i)$, we infer from the r inequalities

$$\langle d, v - u_i\rangle \leq 0 \quad \text{for all } v \in S \ \ (i = 1, \dots, r)$$

that $\langle d, v-u\rangle \leq 0$ for all $v \in S$, whence $d \in N_S(u)$.

Conversely, let $d \in N_{\mathrm{co}\, S}(u)$. Fix $i_0 \in \{1, \dots, r\}$ and for an arbitrary $w \in S$, set $v = \alpha_{i_0} w + \sum_{i \neq i_0} \alpha_i u_i$. Since $v \in \mathrm{co}\, S$,

$$\alpha_{i_0}\langle d, w - u_{i_0}\rangle = \langle d, v - u\rangle \leq 0;$$

that is, $d \in N_S(u_{i_0})$.

(iii) Let $d \in N_S(u)$ and consider $d' \in S_\infty$. For sequences $(x_k)_k$ in S and $(t_k)_k$ in $\mathbb{R}_+$ that give rise to d' (cf. the definition in (25)), we have

$$\langle d, t_k x_k - t_k u\rangle = t_k \langle d, x_k - u\rangle \leq 0 \quad \text{for all } k.$$

Then passing to the limit $k \to +\infty$ yields $\langle d, d'\rangle \leq 0$. □

In a decomposition like (53), we set $y_{q+1} = 0$ if $q = 0$ (i.e., there is no nonnull y_j involved). We have

$$\begin{aligned}
&N_{\overline{\mathrm{co}}(\mathrm{epi} f)}((x, (\overline{\mathrm{co}} f)(x))) \\
&= N_{\mathrm{co}(\mathrm{epi} f)+\mathrm{co}(\mathrm{epi} f_\infty)}\left(\sum_i \alpha_i(x_i, f(x_i)) + \sum_j (y_j, f_\infty(y_j))\right) \\
&= \left[N_{\mathrm{co}(\mathrm{epi} f)}\left(\sum_i \alpha_i(x_i, f(x_i))\right)\right] \cap \left[N_{\mathrm{co}(\mathrm{epi} f_\infty)}\left(\sum_j (y_j, f_\infty(y_j))\right)\right] \\
&\quad \text{(according to Lemma 5.2 (i))} \\
&= [\cap_i N_{\mathrm{epi} f}((x_i, f(x_i)))] \cap [\cap_j N_{\mathrm{epi} f_\infty}((y_j, f_\infty(y_j)))] \\
&\quad \text{(according to Lemma 5.2(ii)).}
\end{aligned}$$

According to Lemma 5.2 (iii), we can now remove all the null y_j's. With (56), we then have

$$\begin{aligned} s \in \partial(\overline{\text{co}}f)(x) &\Leftrightarrow (s,-1) \in N_{\text{epi}(\overline{\text{co}}f)}((x,(\overline{\text{co}}f)(x))) \\ &\Leftrightarrow \begin{cases} (s,-1) \in N_{\text{epi}f}((x_i,f(x_i))) & \text{for all } i \\ \text{and} \\ (s,-1) \in N_{\text{epi}f_\infty}((y_j,f_\infty(y_j))) & \text{for all } j \end{cases} \\ &\Leftrightarrow s \in \partial f(x_i) \quad \text{for all } i \quad \text{and} \quad s \in \partial f_\infty(y_j) \quad \text{for all } j. \qquad \square \end{aligned}$$

Acknowledgment. The authors would like to thank the referee for his careful study of the script and for pointing out that our original proof of (37) was somewhat questionable.

REFERENCES

[1] J. Benoist, *Approximation and regularization of arbitrary sets in finite dimension*, preprint, Department of Mathematics, University of Limoges, Limoges, France, 1992.

[2] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[3] G. Debreu, *Theory of Value*, John Wiley, New York, 1959.

[4] J.-P. Dedieu, *Cône asymptote d'un ensemble non convexe*, C. R. Acad. Sci. Paris Sér. I. Math., 285 (1977), pp. 501–503.

[5] ———, *Critère de fermeture pour l'image d'un fermé non convexe par une multiapplication*, C. R. Acad. Sci. Paris Sér. I. Math., 287 (1978), pp. 941–943.

[6] ———*Cônes asymptotes d'ensembles non convexes*, Bull. Soc. Math. France, 60 (1979), pp. 31–44.

[7] D. McFadden, *Convex Analysis*, in Production Economics: A Dual Approach to Theory and Applications, vol. I, North–Holland, Amsterdam, 1978, Appendix A3.

[8] A. Griewank and P. J. Rabier, *On the smoothness of convex envelopes*, Trans. Amer. Math. Soc., 322 (1990), pp. 691–709.

[9] J.-B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms*, Grundlehren der mathematischen Wissenschaften 305 and 306, Springer-Verlag, Berlin, New York, Heidelberg, 1993.

[10] R. Horst and H. Tuy, *Global Optimization: Deterministic Approaches*, Springer-Verlag, Berlin, New York, Heidelberg, 1990.

[11] C. O. Kiselman, *Regularity classes for operations in convexity theory*, preprint, Department of Mathematics, Uppsala University, Uppsala, Sweden, 1992.

[12] J. B. Kruskal, *Two convex counterexamples: A discontinuous envelope function and a nondifferentiable nearest-point mapping*, Proc. Amer. Math. Soc., 23 (1969), pp. 697–703.

[13] J.-P. Raymond, *Existence of minimizers for vector problems without quasiconvexity condition*, Nonlinear Anal., 18 (1992), pp. 815–828.

[14] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[15] H. D. Sherali and A. Alameddine, *An explicit characterization of the convex envelope of a bivariate bilinear function over special polytopes*, Ann. Oper. Res., 25 (1990), pp. 197–210.

[16] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions*, Springer-Verlag, Berlin, New York, Heidelberg, 1970.

[17] M. Valadier, *Intégration de convexes fermés, notamment d'épigraphes: Inf-convolution continue*, Rev. Inform. Rech. Opér., 1970, pp. 57–73.

SIAM J. MATH. ANAL.
Vol. 27, No. 6, pp. 1680–1703, November 1996

009

INVERSION DE CERTAINS OPÉRATEURS ELLIPTIQUES À COEFFICIENTS VARIABLES[*]

PHILIPPE TCHAMITCHIAN[†]

Abstract. We consider elliptic operators in divergence form with variable coefficients defined through an accretive sesquilinear form on the whole space. The coefficients of leading order are supposed to be lipschitzian. We show how wavelets bases allow us to explicitly compute the inverse of such operators. The first main ingredient is a detailed study, of independant interest, of the convergence of the usual Galerkin approximations. The second main ingredient is the notion of paraproduct, suitably adapted to our context.

Key words. elliptic operators, wavelets, Galerkin approximation, paraproducts

AMS subject classifications. 35A35, 35C10, 35J15, 65M60, 65M70

1. Introduction et exemples.

Cet article présente une analyse de certains opérateurs aux dérivées partielles à coefficients variables, à l'aide des analyses multirésolution et des bases d'ondelettes qui leur sont attachées.

Les opérateurs que nous traitons sont de la forme

$$L = -\operatorname{div} A\nabla + b,$$

vérifiant les hypothèses suivantes:

- ils sont définis sur $\mathbb{R}^n$ (le traitement de problèmes avec conditions au bord étant l'un des principaux problèmes ouverts à l'emploi des ondelettes);
- ils sont associés à une forme sesquilinéaire continue sur $H^1(\mathbb{R}^n)$ et strictement accrétive;
- enfin A est lipschitzienne, ce qui implique que le domaine de L soit l'espace $H^2(\mathbb{R}^n)$.

Notre but est de calculer le plus explicitement possible l'opérateur L^{-1}, au sens de la meilleure topologie possible, c'est-à-dire en tant qu'opérateur continu de L^2 dans H^2, et d'une façon qui soit utilisable en analyse numérique.

Nous avons proposé une première solution à ce problème dans [8]. Mais, reposant sur l'emploi de bases d'ondelettes trop particulières, elle n'était pas très utile et nous ne l'avons pas publiée. Néanmoins, certains ingrédients en sont repris ici.

Une deuxième solution a ensuite été proposée par Bénassi, Jaffard, et Roux, qui consiste à construire une famille de "vaguelettes" orthogonales pour la forme associée à l'opérateur. Mais, bien que constructive, la preuve présente de nombreuses étapes qui, à notre avis, rendent difficile la transposition de ce résultat en situation numérique.

D'un point de vue plus directement algorithmique, on peut penser à partir des idées développées par Beylkin, Coifman, et Rokhlin (BCR) pour la résolution de problèmes intégraux ou différentiels. Cependant, une hypothèse essentielle dans leurs travaux nous fait défaut.

En effet, pour que les algorithmes BCR donnent de bons résultats, ils doivent être appliqués à un opérateur dont le noyau est régulier en dehors de la diagonale, et d'autant plus décroissant à l'infini qu'il est dérivé. L'exemple typique est $K(x,y) = |x-y|^{-n+2}$. Avec un opérateur du type que nous considérons, une telle hypothèse

[*] Received by the editors April 16, 1994; accepted for publication July 31, 1995.

[†] Laboratoire APT, Faculté des Sciences et Techniques de Saint Jérôme, Université d'Aix-Marseille III, Avenue Escadrille Normandie Niemen, 13397 Marseille cedex 13, France.

n'est satisfaite que si les coefficients sont suffisamment réguliers, bien plus que ce que nous voulons supposer.

La question des hypothèses faites sur la régularité des coefficients n'est pas gratuite, même pour le numéricien. Car, outre son intérêt mathématique intrinsèque, elle est directement liée à la stabilité des algorithmes. Par exemple, dans notre calcul de L^{-1}, les quantités importantes sont le conditionnement de A, celui de la forme associée à L, et enfin la norme lipschitz de A. Cela implique que tout algorithme dérivé de nos résultats dépendra de ces constantes, et sera d'autant mieux conditionné qu'elles seront d'un ordre de grandeur raisonnable, et ce indépendamment de la taille des dérivées d'ordre supérieur des coefficients.

Les deux approches, de Benassi, Jaffard, Roux et de Beylkin, Coifman, et Rokhlin, utilisent la forme associée à L et les opérateurs de Galerkin qui approchent L^{-1}. Notre solution repose en partie également sur ces opérateurs, que nous sommes ainsi amenés à étudier. Nous prouvons que ces opérateurs permettent d'approcher L^{-1} pour la topologie forte de H^2, et non pas seulement celle de H^1: c'est là un premier résultat, qui montre que, même si la précision est du même ordre, la méthode de Galerkin converge plus finement dans des espaces associés à une analyse multi–résolution qu'en situation générale.

Ce résultat repose sur des inégalités a priori précisées,[1] que nous établissons en imitant, à l'intérieur de chaque espace d'approximation donné par l'analyse multi–résolution, le calcul fonctionnel de Calderón. Nous définissons en particulier une notion adaptée de transformées de Riesz, et étudions les commutateurs entre ces transformées et les opérateurs de multiplication ponctuelle par les fonctions lipschitziennes.

Il faut cependant souligner, comme nous l'a fait remarquer D. Gottlieb, que nous remercions, que le même résultat est vrai pour les méthodes spectrales, dans le cadre périodique. En revanche, alors que la multiplication par des coefficients variables est peu agréable en Fourier, les bases d'ondelettes permettent de presque diagonaliser les opérateurs de multiplication ponctuelle par des fonctions un peu régulières. C'est là le deuxième ingrédient de notre construction, le paraproduit, issu du calcul paradifférentiel de J. M. Bony, sous la forme simple qu'on peut lui donner avec une base d'ondelettes, qui a été explicitée par Y. Meyer [7].

Combinant les approximations fournies par la méthode de Galerkin et par le calcul paradifférentiel, nous inversons L, en suivant un schéma abstrait qui est expliqué dans la première partie de l'article. Nous y énonçons à la suite notre résultat principal.

Dans la deuxième partie, nous démontrons les inégalités précisées dont nous avons besoin dans la troisième. Celle-ci est consacrée à la preuve de nos résultats sur la qualité de l'approximation par la méthode de Galerkin dans une analyse multi-résolution et sur le calcul de L^{-1}.

L'ensemble de l'article utilise, en plusieurs points cruciaux, les propriétés essentielles des ondelettes. C'est pourquoi nous rappelons, à la fin de la première partie, le lemme fondamental qui régit (jusqu'à présent) l'emploi de ces bases en théorie des opérateurs.

2. Le schéma d'inversion.

2.1. Généralités. Nous le présentons d'abord de manière abstraite, avant de considérer des opérateurs plus concrets.

[1] Ce sont ces inégalités qui nous permettent de généraliser le théorème écrit dans [8] à toutes les analyses multi-résolution raisonnables de $L^2(\mathbb{R}^n)$, et en font ainsi un résultat à notre avis intéressant.

Soit L un opérateur non borné sur un espace de Hilbert $\mathcal{H}$, de domaine dense $D(L)$, réalisant un isomorphisme de $D(L)$ sur $\mathcal{H}$.

Soit également une suite $(V_j)_{j\in\mathbb{Z}}$ d'espaces d'approximation, fermés dans $\mathcal{H}$, emboîtés ($V_j \subset V_{j+1}$) et définissant une approximation de l'identité: $\bigcup_{j\in\mathbb{Z}} V_j$ est dense dans $\mathcal{H}$, et dans $D(L)$ muni de la topologie du graphe. On note

$$\pi_j : \mathcal{H} \to V_j$$

l'opérateur de restriction par projection orthogonale (pour le produit scalaire sur $\mathcal{H}$), et l'adjoint π_j^*, de V_j dans $\mathcal{H}$, est l'opérateur d'extension naturel. On note également $\pi_j^\perp$ la projection orthogonale de $\mathcal{H}$ sur $V_j^\perp$.

On suppose que sont satisfaites les hypothèses suivantes.

Hypothèse (H1). Pour tout $j \in \mathbb{Z}$, l'opérateur

$$\pi_j L \pi_j^* \ : \ V_j \to V_j$$

est inversible.

On note alors Γ_j l'opérateur inverse prolongé à $\mathcal{H}$:

$$\Gamma_j = \pi_j^* (\pi_j L \pi_j^*)^{-1} \pi_j.$$

Les opérateurs Γ_j sont appelés opérateurs de Galerkin.

Hypothèse (H2). Les opérateurs $L\Gamma_j$ sont uniformément bornés.

Remarquer que, par définition, on a

$$\pi_j L \Gamma_j = \pi_j. \tag{1}$$

L'hypothèse porte donc en réalité sur les opérateurs $\pi_j^\perp L \Gamma_j$: ils mesurent le résidu associé à la résolution approchée d'une équation $Lu = f$, où $f \in V_j$, par $u_j = \Gamma_j f$.

Enfin, on suppose qu'on sait approximativement inverser L sur $V_j^\perp$, au sens suivant.

Hypothèse (H3). Pour tout j, il existe un opérateur P_j, continu de $\mathcal{H}$ dans $D(L)$, tel que, si R_j est défini par la relation

$$L P_j = \pi_j^\perp - R_j, \tag{2}$$

alors on a

$$\lim_{j\to+\infty} \|R_j\| = 0.$$

Par exemple, si L admet une parametrix P, c'est-à-dire un opérateur continu de $\mathcal{H}$ dans $D(L)$ tel que

$$L P = I - K, \tag{3}$$

où K est compact, alors $P_j = P\pi_j^\perp$ convient. On aura en effet $R_j = K\pi_j^\perp$, et la compacité de K entraine $\lim_{j\to+\infty} \|R_j\| = 0$. Naturellement, si la relation (3) est satisfaite, il suffit d'inverser $I - K$ pour inverser L. La portée de cette remarque est cependant limitée, premièrement parce qu'il est souvent à peu près aussi ardu de construire une parametrix que de calculer L^{-1}, et deuxièmement parce que l'inversion d'une perturbation compacte de l'identité n'est pas, en pratique, facile a priori.

Le schéma que nous proposons, plus flexible, est fondé sur l'emploi conjugué des opérateurs de Galerkin Γ_j et des parametrix généralisées P_j.

THÉORÈME 2.1. *Avec les notations précédentes, et sous les hypothèses* (H1), (H2), *et* (H3), *si on définit les opérateurs* U_j *par*

$$L(\Gamma_j + P_j) = I - U_j,$$

alors on a

$$\lim_{j\to+\infty} \|U_j^2\| = 0.$$

Par conséquent, la série de Neumann $\sum_{p=0}^{+\infty} U_j^p$ *est normalement convergente pour j assez grand, et on peut écrire*

$$L^{-1} = (\Gamma_j + P_j) \sum_{p=0}^{+\infty} U_j^P.$$

L'argument repose sur une relation d'orthogonalité. On commence par calculer $L(\Gamma_j + P_j)$. Utilisant (1) et (2), il vient

$$\begin{aligned} L(\Gamma_j + P_j) &= \pi_j L\Gamma_j + \pi_j^\perp L\Gamma_j + \pi_j^\perp - R_j \\ &= I + \pi_j^\perp L\Gamma_j - R_j, \end{aligned}$$

c'est-à-dire

$$U_j = R_j - \pi_j^\perp L\Gamma_j.$$

L'opérateur U_j n'est pas petit, il est seulement contrôlé uniformément en norme, grâce à (H2). Mais il faut se rappeler que Γ_j est nul sur $V_j^\perp$, autrement dit que $\Gamma_j = \Gamma_j \pi_j^* \pi_j$. Par conséquent, $\pi_j^\perp L\Gamma_j$ est de carré nul. On a donc

$$U_j^2 = R_j^2 - R_j \pi_j^\perp L\Gamma_j - \pi_j^\perp L\Gamma_j R_j.$$

On conclut alors grâce à (H2) et (H3).

2.2. Enoncé des principaux résultats. Ce schéma sera utilisé dans le cas des opérateurs

$$L = -\operatorname{div} A\nabla + b,$$

définis sur $\mathbb{R}^n$, sous les hypothèses suivantes:

- $A = A(x) = (a_{\alpha\beta}(x))_{1\le\alpha,\beta\le n}$ et $b = b(x)$ sont bornées.
- La forme sesquilinéaire

$$B(f,g) = \int A\nabla f\cdot\nabla\bar{g} + \int b f\bar{g},$$

continue sur $H^1(\mathbb{R}^n)$, est strictement accrétive. Cela signifie qu'il existe $\delta > 0$ tel que

$$\forall f \in H^1, \quad \operatorname{Re} B(f,f) \ge \delta\|f\|_{H^1}^2,$$

ou encore, de façon équivalente,

$$\begin{aligned} \forall x,\xi \in \mathbb{R}^n, \quad &\operatorname{Re} A(x)\xi\cdot\bar{\xi} \ge \delta|\xi|^2. \\ &\operatorname{Re} b(x) \ge \delta. \end{aligned}$$

- La matrice A est lipschitzienne.

L'espace de Hilbert $\mathcal{H}$ est naturellement $L^2(\mathbb{R}^n)$. On note simplement $\|\cdot\|$ sa norme, ainsi que la norme d'opérateur induite.

Les espaces d'approximation V_j seront fournis par une analyse multirésolution tensorielle (AMR)[2] de régularité $r \geq 3$.

Cela signifie que:

- les espaces V_j sont engendrés par les fonctions $\varphi_{jk}(x) = 2^{jn/2}\,\varphi(2^j x - k)$, $k \in \mathbb{Z}^n$, orthonormales entre elles, déduites par translation et dilatation d'une fonction φ;
- cette fonction φ est construite par produit tensoriel à partir d'une fonction d'une variable, notée ϕ:

$$\varphi(x) = \phi(x_1) \cdots \phi(x_n);$$

- la fonction ϕ est au moins de classe C^{r-1} et à dérivée d'ordre $r-1$ lipschitzienne, avec $r \geq 3$, chaque dérivée de ϕ étant à décroissance rapide;
- pour tout $\varepsilon \in \{0,1\}^n$, $\varepsilon \neq 0$, il existe une fonction ψ_ε, de sorte que l'ensemble des $2^{jn/2}\,\psi_\varepsilon(2^j\,x - k)$, $k \in \mathbb{Z}^n$, forme une base orthonormée de W_j, le supplémentaire orthogonal de V_j dans V_{j+1};
- chaque ψ_ε, comme φ, est au moins de classe C^{r-1}, chaque dérivée d'ordre $r-1$ est au moins lipschitzienne, et chaque dérivée de ψ_ε est à décroissance rapide.
- On aura besoin de la description plus précise suivante. Il existe une ondelette d'une variable, Ψ, ayant les mêmes propriétés que ϕ, et telle qu'il existe un entier $N \geq r$ pour lequel on a

$$\forall k \leq N, \quad \int_{\mathbb{R}} t^k\,\Psi(t)\,dt = 0. \tag{4}$$

En convenant de noter $\psi_0 = \phi$ et $\psi_1 = \Psi$, et si $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$, alors on a

$$\psi_\varepsilon(x) = \psi_{\varepsilon_1}(x_1) \cdots \psi_{\varepsilon_n}(x_n).$$

On écrit

$$2^{jn/2}\,\psi_\varepsilon(2^j x - k) = \psi_\lambda(x),$$

où $\lambda = (k + \frac{1}{2}\varepsilon)\,2^{-j}$, et on désigne par D l'ensemble des λ. La correspondance

$$\lambda \mapsto (j,\, k,\, \varepsilon)$$

est biunivoque, et on notera parfois $j(\lambda)$, $k(\lambda)$, $\varepsilon(\lambda)$ les indices associés à λ.

La collection des ψ_λ, $\lambda \in \Lambda$, est une base orthonormée de $L^2(\mathbb{R}^n)$, et une base inconditionnelle des espaces de Sobolev $H^s(\mathbb{R}^n)$, $|s| < 3$.

Nous renvoyons le lecteur à [3] et [5] pour les connaissances de base sur les ondelettes. Nous nous contenterons d'indiquer que l'intuition avec laquelle on peut comprendre ce qui suit est que les fonctions φ_{jk} et ψ_λ se comportent comme si elles étaient supportées dans le cube dyadique $k2^{-j} + 2^{-j}\,[0,1]^n$, et, dans l'espace de Fourier, dans

[2] Une remarque de P. Auscher, que nous remercions, a permis de généraliser notre preuve au cas des AMR de régularité $r \geq 3$ quelconques. Cependant, ces analyses sont, à notre connaissance, peu utilisées. Nous nous contenterons donc d'indiquer rapidement les modifications à apporter, sans les détailler. Ces modifications sont, de toute façon, de nature technique et s'insèrent dans la stratégie générale de la preuve.

la boule $|\omega| \leq 2^j$ pour ce qui concerne φ_{jk}, et dans la couronne $2^{j-1} \leq |\omega| \leq 2^{j+1}$ pour ce qui concerne ψ_λ.

Nous démontrerons le théorème suivant.

THÉORÈME 2.2. *L'opérateur L et l'AMR (V_j) satisfont aux hypothèses* (H1), (H2), *et* (H3), *et avec les notations du Théorème* 2.1, *on a plus précisément*

$$\|U_j^2\| \leq C\,2^{-j}, \tag{5}$$

où C est une constante indépendante de j.

Remarquons que, L étant associé à une forme strictement accrétive, l'hypothèse (H1) est classiquement satisfaite. La difficulté est de prouver (H2) et (H3). Si la preuve de (H3) emploie des ingrédients plus ou moins déjà connus, celle de (H2) va nécessiter quelques développements assez nouveaux, croyons-nous, et qui conduisent à un résultat intéressant par lui-même, que nous énonçons maintenant.

THÉORÈME 2.3. *Reprenant les notations du paragraphe* 2.1, *on a*

$$\forall f \in L^2, \quad \lim_{j\to+\infty} \|L^{-1} f - \Gamma_j\, f\|_{H^2} = 0.$$

Autrement dit, les approximations u_j de la solution du problème $Lu = f$, données par $u_j = \Gamma_j f$ convergent non seulement en norme H^1 vers u, mais aussi en norme H^2.

Les AMR conduisent donc à des schémas de Galerkin plus précis, au sens que la convergence a lieu pour une topologie plus fine que la topologie habituelle, même si l'ordre de précision est inchangé.

Ainsi que nous l'avons annoncé, ces théorèmes sont une conséquence de l'existence des bases d'ondelettes. La base des φ_{jk}, $k \in \mathbb{Z}^n$, dans V_j, nous sera de faible utilité, et nous utiliserons presque constamment les ondelettes ψ_λ.

Rappelons donc de quelle façon on emploie ces fonctions en théorie des opérateurs.

2.3. Le lemme fondamental. Jusqu'à présent, les bases d'ondelettes ont servi à prouver la continuité de certains opérateurs, le plus souvent sur L^2. La démarche suivie consiste, à partir d'une analyse adéquate de l'opérateur étudié, à se ramener à une classe d'opérateurs définis à travers les images de chaque ondelette de base, qui soient des fonctions imitant le comportement des ondelettes (et qu'on appelle vaguelettes pour cette raison).

Or, la continuité des opérateurs de cette classe est entièrement caractérisée.

LEMME FONDAMENTAL. *Soit U un opérateur linéaire défini par $U(\psi_\lambda) = f_\lambda$, où les fonctions f_λ satisfont aux hypothèses suivantes:*

$$\forall x \in \mathbb{R}^n, \quad |f_\lambda(x)| \leq C\,2^{jn/2}\,(1 + 2^j|x-\lambda|)^{-n-s}, \tag{6}$$

$$\forall x \in \mathbb{R}^n, \quad |\nabla f_\lambda(x)| \leq C\,2^{jn/2}\,2^j\,(1 + 2^j|x-\lambda|)^{-n-s}, \tag{7}$$

où s est un réel > 0 indépendant de λ, ainsi que la constante C.

Alors, U est continu sur $L^2(\mathbb{R}^n)$ si et seulement si les scalaires $\int f_\lambda$ satisfont à la condition de Carleson:

$$\exists C > 0, \quad \forall \lambda \in D, \quad \sum_{Q_\mu \subset Q_\lambda} \left|\int f_\mu\right|^2 \leq C^2\,|Q_\lambda|,$$

où Q_λ est le cube dyadique contenant λ dont le côté est de longueur $2^{-j(\lambda)}$.

Si cette condition est satisfaite, la norme d'opérateur de U est contrôlée par la meilleure constante C possible dans les trois inégalités précédentes.

Le lecteur intéressé pourra trouver une preuve de ce résultat dans [6], sous une forme d'ailleurs un peu plus générale (c'est-à-dire avec des hypothèses (6) et (7) plus faibles). L'énoncé que nous en donnons est celui dont nous avons besoin. Signalons enfin qu'un argument court, mais sophistiqué, qui démontre le lemme fondamental est de remarquer que U est un opérateur d'intégrale singulière, dont le noyau est de Calderón–Zygmund, puis d'appliquer le critère de David et Journé, autrement dit le théorème T(1). On vérifie alors que tout se ramène à savoir si la fonction

$$\sum_\lambda \overline{\left(\int f_\lambda\right)}\, \psi_\lambda = U^*(1)$$

appartient ou non à BMO (bounded mean oscillation). Or, cette appartenance est équivalente à la condition de Carleson.

On convient, dans toute la suite, d'appeler *estimations standard de paramètre s* les estimations (6) et (7). On dira parfois qu'une famille (f_λ) vérifie les estimations standard, ce qui voudra dire qu'il existe $s > 0$ pour lequel les estimations de paramètre s sont vraies.

L'exemple le plus caractéristique d'emploi du lemme fondamental, et qui est le prototype de l'usage que nous en ferons, est donné par la preuve suivante du célèbre théorème du premier commutateur de Calderón, dont nous rappelons la teneur.

On se donne $a(x)$, $x \in \mathbb{R}$, une fonction lipschitzienne d'une variable; on note H la transformée de Hilbert. Il s'agit de montrer que le commutateur $[a, H]$ est régularisant d'ordre un, c'est-à-dire que l'opérateur $[a, H]\, \frac{d}{dx}$ est borné sur L^2.

Calculons l'image d'une ondelette ψ_λ par cet opérateur. Elle s'écrit

$$a(x)\, 2^j\, (H\, \psi')_\lambda(x) - H\{2^j a\, (\psi')_\lambda\}\, (x),$$

avec $(H\psi')_\lambda(x) = 2^{j/2}\, (H\, \psi')\, (2^j\, x - k)$, et de même pour $(\psi')_\lambda$.

Mais on peut soustraire à $a(x)$ n'importe quelle constante sans changer l'image de ψ_λ : on remplace donc $a(x)$ par $a(x) - a(\lambda)$.

Soient alors U_1 et U_2 les deux opérateurs définis par

$$U_1\, \psi_\lambda(x) = 2^j\, [a(x) - a(\lambda)]\, (H\, \psi')_\lambda(x),$$
$$U_2\, \psi_\lambda(x) = 2^j\, [a(x) - a(\lambda)]\, (\psi')_\lambda(x),$$

de sorte que

$$[a, H]\, \frac{d}{dx} = U_1 - H\, U_2.$$

Il suffit de prouver que U_1 et U_2 sont continus sur L^2 pour prouver le théorème de Calderón. Pour cela, on applique le lemme fondamental.

On vérifie aisément que U_1 et U_2 en satisfont les hypothèses. On calcule donc:

$$\begin{aligned} \int U_1\, (\psi_\lambda) &= - \int a'\, (H\psi)_\lambda \\ &= - \int (Ha')\, \psi_\lambda \\ \text{et } \int U_2\, (\psi_\lambda) &= - \int a'\, \psi_\lambda\, . \end{aligned}$$

Les deux conditions de Carleson sont remplies, parce que $a' \in L^\infty$ par hypothèse, ce qui entraîne que $H\,a' \in \mathrm{BMO}$, et permet de conclure.

Nous en venons maintenant à la preuve des Théorèmes 1.2 et 1.3, et commençons par la démonstration de quelques inégalités a priori.

3. Une inégalité précisée.

3.1. Enoncé du théorème. Soit $B = B(x) = (b_{\alpha\beta}(x))$ une fonction à valeurs matricielles, supposée bornée et lipschitzienne. On notera $\|B\|$ le supremum des normes $\|B(x)\|_{2,2}$, induites par la norme euclidienne sur $\mathbb{R}^n$, et $\|B\|_{\mathrm{lip}}$ la quantité $\sum_{\alpha,\beta}\|\nabla\, b_{\alpha\beta}\|_\infty$.

L'opérateur $b_{\alpha\beta}\,\partial_\alpha\,\partial_\beta$ (où la sommation sur tous les indices α et β est sous-entendue) est bien défini de H^2 dans L^2, et il existe une constante C telle que

$$\forall f \in H^2, \quad \|b_{\alpha\beta}\,\partial_a\,\partial_\beta\, f\| \leq C\|\Delta f\|$$

(où $\|\cdot\|$ désigne la norme L^2).

Dans le cas où B est en fait constante, la meilleure constante C est exactement $\|B\|$. Dans le cas général, cette observation reste vraie, à un terme d'ordre 1 près.

LEMME 3.1. *Il existe une constante absolue $C = C(n)$ telle que*

$$\forall f \in H^2, \quad \|b_{\alpha\beta}\,\partial_\alpha\,\partial_\beta\, f\| \leq \|B\|\;\|\Delta\, f\| + C\,\|B\|_{\mathrm{lip}}\;\|\nabla\, f\|.$$

La preuve de ce résultat est très simple, mais repose sur le résultat profond de Calderón que nous avons déjà mentionné: si b est lipschitzienne et si R_α est une transformée de Riesz, alors le commutateur $[b,\, R_\alpha]$ est régularisant d'ordre 1 (exactement, c'est la généralisation à plusieurs dimensions du théorème de Calderón que nous utilisons).

Soit $f \in H^2$. On a

$$\begin{aligned} b_{\alpha\beta}\,\partial_\alpha\,\partial_\beta\, f &= b_{\alpha\beta}\,R_\alpha\,R_\beta\,(\Delta f) \\ &= R_\alpha\,b_{\alpha\beta}\,R_\beta\,(\Delta f) \\ &\quad + [b_{\alpha\beta},\, R_\alpha]\,R_\beta\,(\Delta f). \end{aligned}$$

Le deuxième terme est celui auquel on applique le résultat de Calderón: il se majore par $C\|B\|_{lip}\,\|\nabla f\|$. Pour estimer le premier, il suffit de se donner $g \in L^2$; on a alors

$$\tag{8} \begin{aligned} |\langle R_\alpha\, b_{\alpha\beta}\, R_\beta\,\Delta f,\, g\rangle| &= |\langle b_{\alpha\beta}\, R_\beta\,\Delta f, R_\alpha\, g\rangle| \\ &\leq \|B\| \left\| \left(\sum |R_\beta\,\Delta f|^2\right)^{1/2} \right\| \left\| \left(\sum |R_\alpha g|^2\right)^{1/2} \right\| \\ &\leq \|B\|\|\Delta f\|\|g\|. \end{aligned}$$

(Remarquons que le choix de la matrice B, définissant l'opérateur $b_{\alpha\beta}\,\partial_\alpha\,\partial_\beta$, n'est pas unique. Si B est réelle, c'est en imposant à B d'être symétrique qu'on minimise la norme $\|B\|$.)

Le but de cette deuxième partie est d'établir une version plus forte du Lemme 3.1, où l'espace de départ est un espace d'approximation V_j, au lieu de H^2.

THÉORÈME 3.2. *On désigne par π_j la projection orthogonale de L^2 sur l'espace V_j. Alors, il existe une constante C, ne dépendant que de l'AMR, telle que*

$$\tag{9} \begin{gathered} \forall j \in \mathbb{Z}, \quad \forall f \in V_j, \\ \|\pi_j b_{\alpha\beta}\,\partial_\alpha\,\partial_\beta\, f\| \leq \|B\|\;\|\pi_j\,\Delta f\| + C\|B\|_{\mathrm{lip}}\,\|\nabla f\|. \end{gathered}$$

Noter que le passage à la limite $j \to +\infty$ permet de retrouver le Lemme 3.1.

L'uniformité par rapport à j dans l'inégalité précédente résulte simplement de son invariance par scaling. Ce théorème est donc en réalité un résultat sur l'espace V_0. Néanmoins, la structure d'AMR sera utilisée dans la preuve, de sorte que nous ne savons pas pour quels espaces V, en toute généralité, ce résultat reste vrai.

Remarquons cependant que, suivant D. Gottlieb, dans le cas des approximations de Fourier, on a $\pi_j \Delta f = \Delta f$ si $f \in V_j$. Le Lemme 3.1 implique alors aussitôt l'inégalité (9), puisque π_j est une projection orthogonale.

Pour prouver le Théorème 3.2, nous nous plaçons dans V_0. La démonstration suit alors à peu près le même schéma que celle du Lemme 3.1. Elle repose sur l'emploi de transformées de Riesz adaptées à l'espace V_0, qui permettent de passer de $\pi_0 \, \Delta f$ à $\pi_0 \, \partial_\alpha \, \partial_\beta \, f$, pour toute $f \in V_0$. En particulier, nous éviterons entièrement le recours aux commutateurs de la forme $[\pi_0, \, \partial_\alpha]$.

La section suivante est consacré à l'étude de ces transformées, avant de donner la preuve du Théorème 3.2.

3.2. Transformées de Riesz adaptées à l'espace V_0.

3.2.1. Opérateurs à coefficients constants et transformées de Riesz adaptées. On oublie ici, et jusqu'à la fin de cette deuxième partie, l'indice 0, et on note V, π au lieu de V_0, π_0. On rappelle que π est défini comme étant un opérateur de L^2 dans V, et son adjoint π^* de V dans L^2.

V étant invariant par translation selon $\mathbb{Z}^n$, les calculs se font, dans cette section, avec la transformée de Fourier. Rappelons que ([3], [5]) tout élément f de V est donné par

$$\hat{f}(\omega) = m(\omega) \, \hat{\varphi}(\omega),$$

où m est $2\pi \, \mathbb{Z}^n$-périodique et de carré intégrable (il suffit d'écrire $f(x) = \sum c_k \, \varphi(x - k)$ pour le voir). De plus, l'orthonormalité des fonctions $\varphi(\cdot - k)$ est équivalente à la relation

$$\sum_{k \in \mathbb{Z}^n} |\hat{\varphi}(\omega + 2\pi k)|^2 = 1, \tag{10}$$

qui est vraie en tout $\omega \in \mathbb{R}^n$. Finalement, si f est quelconque dans L^2, sa projection orthogonale sur V est donnée par

$$(\pi f)^\wedge(\omega) = m(\omega) \, \hat{\varphi}(\omega), \tag{11}$$

avec

$$m(\omega) = \sum_{k \in \mathbb{Z}^n} \hat{f}(\omega + 2\pi k) \, \overline{\hat{\varphi}(\omega + 2\pi k)}.$$

Soit maintenant $P(D)$ un opérateur aux dérivées partielles à coefficients constants, d'ordre 2 pour fixer les idées, et de symbole $\sigma(\omega)$:

$$(P(D) \, f)^\wedge(\omega) = \sigma(\omega) \, \hat{f}(\omega).$$

Il résulte de (11) que, si $f \in V$, avec $\hat{f}(\omega) = m(\omega) \, \hat{\varphi}(\omega)$, alors on a

$$(\pi \, P(D) \, f)^\wedge(\omega) = \sigma^V(\omega) \, m(\omega) \, \hat{\varphi}(\omega), \tag{12}$$

où

$$\sigma^V(\omega) = \sum_{k \in \mathbb{Z}^n} \sigma(\omega + 2\pi k)\, |\hat{\varphi}(\omega + 2\pi k)|^2$$

(l'hypothèse de régularité sur φ assure l'existence de σ^V).

En d'autres termes, l'opérateur $\pi\, P(D)\, \pi^*$ est l'opérateur de convolution associé au symbole σ^V.

En particulier, on a, si $f \in V$,

$$(\pi\, \Delta f)^\wedge(\omega) = -\left(\sum_{k \in \mathbb{Z}^n} |\omega + 2\pi k|^2\, |\hat{\varphi}(\omega + 2\pi k)|^2 \right) \hat{f}(\omega).$$

On définit alors l'opérateur Λ^V, de V dans lui-même, par

$$\forall f \in V, \quad (\Lambda^V f)^\wedge(\omega) = \left(\sum_k |\omega + 2\pi k|^2\, |\hat{\varphi}(\omega + 2\pi k)|^2 \right)^{1/2} \hat{f}(\omega),$$

de sorte que

$$(\Lambda^V)^2 = -\pi \Delta \pi^*.$$

On définit ensuite les transformées de Riesz adaptées à V, notées R^V_α et $R^V_{\alpha,\beta}$, par les formules

$$\begin{aligned} R^V_\alpha &= -i(\pi\, \partial_\alpha\, \pi^*)\, (\Lambda^V)^{-1}, \\ R^V_{\alpha,\beta} &= (\pi\, \partial_\alpha\, \partial_\beta\, \pi^*)\, (\pi\, \Delta \pi^*)^{-1}. \end{aligned}$$

Si on pose

$$k^V_\alpha(\omega) = -i\, \frac{\displaystyle\sum_k (\omega_\alpha + 2\pi k_\alpha)\, |\hat{\varphi}(\omega + 2\pi k)|^2}{\left(\displaystyle\sum_k |\omega + 2\pi k|^2\, |\hat{\varphi}(\omega + 2\pi k)|^2 \right)^{1/2}}$$

et

$$k^V_{\alpha,\beta}(\omega) = \frac{\displaystyle\sum_k (\omega_\alpha + 2\pi k_\alpha)\, (\omega_\beta + 2\pi k_\beta)\, |\hat{\varphi}(\omega + 2\pi k)|^2}{\displaystyle\sum_k |\omega + 2\pi k|^2\, |\hat{\varphi}(\omega + 2\pi k)|^2}$$

(ces fonctions sont $2\pi\mathbb{Z}^n$-périodiques, définies en tout $\omega \notin 2\pi\mathbb{Z}^n$ et bornées par 1), alors R^V_α et $R^V_{\alpha,\beta}$ sont les restrictions à V des opérateurs de convolution de symboles k^V_α et $k^V_{\alpha,\beta}$:

$$\begin{aligned} (R^V_\alpha\, f)^\wedge\, (\omega) &= k^V_\alpha(\omega)\, \hat{f}(\omega), \\ (R^V_{\alpha,\beta}\, f)^\wedge\, (\omega) &= k^V_{\alpha,\beta}(\omega)\, \hat{f}(\omega). \end{aligned}$$

Par construction, si $f \in V$, on a

$$\pi\, \partial_\alpha\, \partial_\beta\, f = R^V_{\alpha,\beta}\, (\pi \Delta f). \tag{13}$$

Dans $L^2(\mathbb{R}^n)$, on a l'identité algébrique

$$R_{\alpha,\beta} = R_\alpha \, R_\beta \ ,$$

qui, toute élémentaire qu'elle soit, est essentielle dans la preuve du Lemme 3.1. Ici, on n'a pas a priori la même relation entre $R^V_{\alpha,\beta}$, R^V_α, et R^V_β. Cependant, la structure de produit tensoriel de V nous donne le lemme suivant.

LEMME 3.3. *Si $\alpha \neq \beta$, $R^V_{\alpha,\beta} = R^V_\alpha \, R^V_\beta$.*

Cela provient des identités

$$\hat{\varphi}(\omega) = \hat{\phi}(\omega_1) \cdots \hat{\phi}(\omega_n)$$

et

$$\sum_{k_\alpha \in \mathbb{Z}} |\hat{\phi}(\omega_\alpha + 2\pi\, k_\alpha)|^2 = 1, \quad \alpha = 1, \dots, n,$$

d'où

$$\sum_k (\omega_\alpha + 2\pi k_\alpha)\,(\omega_\beta + 2\pi k_\beta)\,|\hat{\varphi}(\omega + 2\pi\, k)|^2$$
$$= \left(\sum_k (\omega_\alpha + 2\pi k_\alpha) |\hat{\varphi}(\omega + 2\pi\, k)|^2 \right) \left(\sum_k (\omega_\beta + 2\pi k_\beta) |\hat{\varphi}(\omega + 2\pi\, k)|^2 \right).$$

En termes de dérivées partielles, le Lemme 3.3 signifie que

$$\forall \alpha \neq \beta, \quad \forall f \in V, \quad \pi\, \partial_\alpha\, \partial_\beta\, f = (\pi\, \partial_\alpha\, \pi^*)\,(\pi\, \partial_\beta\, f).$$

En revanche, quand $\alpha = \beta$, l'identité du Lemme 3.3 est fausse. Il en subsiste cependant une forme plus faible.

LEMME 3.4. *L'opérateur $R^V_{\alpha,\alpha} - (R^V_\alpha)^2$ est positif.*

En effet, de (10) et par Cauchy–Schwarz, on a l'inégalité

$$k^V_\alpha(\omega)^2 \leq k^V_{\alpha,\alpha}(\omega).$$

On peut donc poser

$$S^V_\alpha = \left(R^V_{\alpha,\alpha} - (R^V_\alpha)^2\right)^{1/2},$$

qui est le multiplicateur de symbole

$$\left(k^V_{\alpha,\alpha}(\omega) - k^V_\alpha(\omega)^2\right)^{1/2}.$$

On a alors le lemme suivant.

LEMME 3.5. *Pour toute $f \in V$,*

$$\int \sum_\alpha |R^V_\alpha\, f|^2 + \sum_\alpha |S^V_\alpha\, f|^2 = \int |f|^2.$$

C'est une conséquence directe de la définition de S^V_α, chacun des deux membres de l'égalité étant égal à $\int \sum_\alpha |R^V_{\alpha,\alpha}\, f|^2$.[3]

[3] (P. Auscher) Dans le cas où l'AMR n'a pas une structure de produit tensoriel, la construction précédente doit être modifiée en définissant la matrice des opérateurs $(S^V_{\alpha,\beta})$ comme étant la racine carrée de la matrice $(R^V_{\alpha,\beta})$. Tous les résultats que nous prouverons sur les R^V_α seront également vrais pour les $S^V_{\alpha,\beta}$, et ce jusqu'à la preuve du Théorème 3.2. Nous laissons au lecteur l'écriture du cas général, ne traitant en détail que le cas tensoriel.

3.2.2. Images des ondelettes. Après l'étude algébrique précédente, venons-en à l'analyse des opérateurs adaptés à V. Ils sont les restrictions à V d'opérateurs de convolution, dont les symboles sont des fonctions bornées $2\pi\,\mathbb{Z}^n$-périodiques. Faire leur analyse signifie pour nous en calculer les images des ondelettes ψ_λ, avec $j(\lambda) \leq -1$, qui forment une base de V: ces opérateurs présentent alors trois types de comportement.

L'opérateur Λ^V, de symbole

$$\ell^V(\omega) = \left(\sum_k |\omega + 2\pi k|^2\, |\hat{\varphi}(\omega + 2\pi k|^2\right)^{1/2},$$

s'apparente à l'opérateur de Calderón $\Lambda = \sqrt{-\Delta}$, de symbole $|\omega|$, parce que $\ell^V(\omega)$ est équivalent à $|\omega|$ au voisinage de 0.

Les opérateurs R^V_α et $R^V_{\alpha,\beta}$ s'apparentent aux transformées de Riesz, parce que leurs symboles vérifient

$$k^V_\alpha(\omega) \sim -i\frac{\omega_\alpha}{|\omega|}$$

et

$$k^V_{\alpha,\beta}(\omega) \sim \frac{\omega_\alpha\,\omega_\beta}{|\omega|^2}$$

au voisinage de 0. Quant aux opérateurs S^V_α, leur symbole n'est pas singulier, comme on le verra, et leur action est triviale.

Les opérateurs Λ ou R_α agissent d'une façon bien connue sur les ondelettes, et nous allons démontrer que les opérateurs adaptés à V agissent de façon équivalente. La difficulté provient du caractère périodique de leur symbole: par exemple, k^V_α est singulier en tout point de $2\pi\mathbb{Z}^n$, et pas seulement en l'origine. C'est pourquoi il faut comparer Λ^V à $\pi\Lambda\pi^*$ et R^V_α à $\pi\, R_\alpha\, \pi^*$: la projection π a pour effet de périodiser les symboles de Λ et de R_α et ainsi d'imiter les symboles ℓ^V et k^V_α.

A cet effet, nous définissons les opérateurs T^V_0 et T^V_α, $\alpha = 1, \ldots, n$, agissant dans V, par les relations

$$\Lambda^V = T^V_0\, \pi\Lambda\pi^*, \tag{14}$$

où, rappelons-le, $\Lambda = \sqrt{-\Delta}$, et

$$R^V_\alpha = T^V_\alpha + \pi\, R_\alpha\, \pi^*. \tag{14'}$$

Ce sont des opérateurs de convolution, de symboles respectifs

$$\tau_0(\omega) = \frac{\left(\sum_k |\omega + 2\pi\, k|^2\, |\hat{\varphi}(\omega + 2\pi k)|^2\right)^{1/2}}{\sum_k |\omega + 2\pi\, k|\, |\hat{\varphi}(\omega + 2\pi k)|^2}$$

et

$$\tau_\alpha(\omega) = k^V_\alpha(\omega) - \sum_k \frac{\omega_\alpha + 2\pi k_\alpha}{|\omega + 2\pi k|}\, |\hat{\varphi}(\omega + 2\pi k)|^2.$$

L'opérateur S_α^V a un symbole du même type, noté $\sigma_\alpha(\omega)$, égal à $\left(k_{\alpha,\alpha}^V(\omega) - k_\alpha^V(\omega)^2\right)^{1/2}$.

Tous ces symboles sont des fonctions $2\pi\mathbb{Z}^n$-périodiques de classe au moins C^{2N-1}, où N est défini par les propriétés d'oscillation des ondelettes: voir (4). On démontre en fait le résultat suivant.

LEMME 3.6. *Pour chacun des symboles τ_0, τ_α, σ_α, notés génériquement τ, il existe une constante C telle que*

$$\forall k \in \mathbb{Z}^n, \quad |\hat{\tau}\,(k)| \le \frac{C}{1+|k|^{n+2N-1}}.$$

Nous ne faisons qu'en esquisser la démonstration, de nature technique sans être difficile.

Il suffit de prouver que $\partial^\alpha \tau$ est intégrable sur $[-\pi, \pi]^n$ pour tout α de longueur inférieure ou égale à $n + 2N - 1$, et pour cela, il est également suffisant de montrer que

$$\tau(\omega) = \tau(0) + O(|\omega|^{2N}) \tag{15}$$

au voisinage de 0, puisque τ est C^∞ en–dehors de 0. Finalement, tout se ramène à connaître le comportement de $|\hat{\varphi}(\omega + 2\pi k)|^2$ au voisinage de 0, pour tout $k \in \mathbb{Z}^n$.

On écrit alors

$$\hat{\varphi}\,(\omega) = \hat{\phi}(\omega_1)\cdots\hat{\phi}(\omega_n),$$

et il est connu que

$$\hat{\phi}(\xi) = \prod_{j=1}^{+\infty} m_0(2^{-j}\xi),$$

pour une certaine fonction C^∞ 2π–périodique m_0, satisfaisant entre autres à l'identité

$$|m_0(\xi)|^2 + |m_0(\xi+\pi)|^2 = 1.$$

Par définition de ϕ et de N, on a

$$m_0(\xi+\pi) = O(\xi^{N+1})$$

au voisinage de 0, d'où on déduit que

$$|\hat{\phi}(\xi)|^2 = 1 + O(\xi^{2N+2})$$

et que

$$|\hat{\varphi}(\omega)|^2 = 1 + O(|\omega|^{2N+2}).$$

Si maintenant $k \in \mathbb{Z}^n$, $k \neq 0$, il existe $m \ge 0$ tel que $k \in 2^m\,\mathbb{Z}^n$ et $k \notin 2^{m+1}\,\mathbb{Z}^n$. Le lecteur intéressé pourra montrer qu'on a alors

$$|\hat{\varphi}(\omega + 2\pi k)|^2 \le C\,2^{-(2N+2)m}\,|\omega|^{2N+2}\,|\hat{\varphi}(2^{-m-1}\,[\omega + 2\pi k])|^2,$$

uniformément par rapport à $k \in \mathbb{Z}^n$ et ω dans un voisinage de 0. Cette formule implique que, si $\sigma(\omega)$ est une fonction homogène de ω, d'ordre strictement inférieur à $2N + 2$, et telle que

$$\sum_k |\sigma(\omega + 2\pi k)|\,|\hat{\varphi}\,(\omega + 2\pi k)|^2 \in L^\infty,$$

alors, au voisinage de 0, on a

$$\sum_k \sigma(\omega + 2\pi k)\,|\hat{\varphi}\,(\omega + 2\pi k)|^2 = \sigma(\omega) + 0(|\omega|^{2N+2}).$$

Puisque, dans l'application au Lemme 3.6, σ n'est jamais de degré supérieur à 2, on en déduit l'estimation (15).

L'action des opérateurs S^V_α, T^V_0, et T^V_α s'analyse alors simplement grâce au lemme suivant.

LEMME 3.7. *Si f_λ est une fonction satisfaisant les estimations standard* (6) *et* (7), *avec un paramètre s, une constante C_0, et $j(\lambda) \leq -1$, et si (c_k) est une suite de scalaires tels que*

$$\forall k \in \mathbb{Z}^n, \quad |c_k| \leq \frac{1}{1 + |k|^{n+s}},$$

alors la fonction

$$g_\lambda(x) = \sum_k c_k\, f_\lambda(x - k)$$

satisfait également à (6) *et* (7), *avec le même paramètre s, et une constante C ne dépendant que de C_0.*

Notant $j = j(\lambda)$, on décompose g_λ en

$$g_\lambda = \sum_{|k| \leq 2^{-j}} + \sum_{|k| > 2^{-j}} .$$

Le premier terme est majoré par

$$\begin{aligned} &C_0 \sum_{|k| \leq 2^{-j}} |c_k|\, 2^{jn/2}\,(1 + 2^j |x - k - \lambda|)^{-n-s} \\ &\leq C(n+s)\, C_0\, \|c_k\|_1\, 2^{jn/2}\,(1 + 2^j\,|x - \lambda|)^{-n-s}. \end{aligned}$$

Le second se majore par

$$\begin{aligned} &2C_0 \sum_{|k| > 2^{-j}} \frac{2^{j(n+s)}}{1 + |2^j\, k|^{n+s}}\, 2^{jn/2}\,(1 + 2^j |x - k - \lambda|)^{-n-s} \\ &\leq C'(s)\, C_0\, 2^{js}\, 2^{jn/2}\,(1 + 2^j\,|x - \lambda|)^{-n-s}. \end{aligned}$$

Puisque $j \leq -1$, ce terme est du même ordre que le premier, voire négligeable devant lui.

Les estimations sur $\nabla\, g_\lambda$ se prouvent de la même façon.

Il nous faut maintenant étudier l'action de $\pi\Lambda\pi^*$ et de $\pi\, R_\alpha\, \pi^*$ sur une ondelette ψ_λ appartenant à V. Rappelons d'abord comment agissent Λ et R_α, en partant du résultat classique suivant [6].

LEMME 3.8. *Soit $G(x)$ une fonction C^∞ sur $\mathbb{R}^n \backslash \{0\}$, telle que, pour tout multi-indice γ, il existe une constante C_γ vérifiant*

$$\forall x \in \mathbb{R}^n \backslash \{0\}, \quad |\partial^\gamma G(x)| \leq C_\gamma\, |x|^{-n+m-|\gamma|}, \tag{16}$$

pour un certain entier $m \geq 1$. Il existe alors, si $N \geq m$, une constante C telle que, pour tout $\lambda \in D$, on ait

$$\forall x \in \mathbb{R}^n, \quad |G * \psi_\lambda(x)| \leq C\, 2^{-jm}\, 2^{jn/2}\,(1 + 2^j\,|x - \lambda|)^{-n+m-N-1},$$

et plus généralement, pour tout multi-indice γ, $|\gamma| \leq m + r - 1$,

$$\forall x \in \mathbb{R}^n, \quad |\partial^\gamma G * \psi_\lambda(x)| \leq C'_\gamma\, 2^{-jm+j|\gamma|}\, 2^{jn/2}(1 + 2^j\,|x - \lambda|)^{-n+m-N-1-|\gamma|}.$$

Pour la commodité du lecteur, nous en donnons une démonstration, puis nous en déduirons les estimations sur $R_\alpha\,\psi_\lambda$ et $\Lambda\psi_\lambda$.

On commence par se ramener au cas $j(\lambda) = k(\lambda) = 0$, par homogénéité et invariance par translation, de sorte qu'il s'agit de décrire la décroissance à l'infini de $G * \psi_\varepsilon$.

Celle-ci provient des oscillations de l'ondelette ψ_ε.

En effet, on se rappelle que ψ_ε est un produit tensoriel $\Psi_{\varepsilon_1}(x_1)\cdots\Psi_{\varepsilon_n}(x_n)$, où $\varepsilon_i = 1$ ou 0, $\Psi_1 = \Psi$ et $\Psi_0 = \phi$, l'un au moins des ε_i étant non nul.

Traitons en détail le cas $\varepsilon = (1, 0, \dots, 0)$, les autres cas se traitant de même. Par définition de N, il existe $\Psi^{-(N+1)}$, primitive d'ordre $N+1$ de Ψ, qui soit à décroissance rapide. Soit θ la fonction

$$\theta(x) = \Psi^{-(N+1)}(x_1)\,\phi(x_2)\cdots\phi(x_n).$$

Par construction, $\partial_1^{N+1}\,\theta = \psi_\varepsilon$, et θ est à décroissance rapide.

On se donne maintenant une fonction $\mathcal{X}$ régulière engendrant une partition de l'unité: $\mathcal{X}$ est C^∞, Supp $\mathcal{X} \subset [-10, 10]^n$, et $\sum_{k \in \mathbb{Z}^n} \mathcal{X}(x - k) = 1$ pour tout $x \in \mathbb{R}^n$. On découpe θ en

$$\theta = \sum_{k \in \mathbb{Z}^n} \theta_k,$$

où $\theta_k(x) = \theta(x)\,\mathcal{X}(x - k)$. Ainsi, on a

$$G * \psi_\varepsilon = \sum_{k \in \mathbb{Z}^n} (\partial_1^{N+1}\,\theta_k) * G.$$

Soient $x \in \mathbb{R}^n$, $k \in \mathbb{Z}^n$. Si $|x - k| \leq 20$, on écrit

$$\begin{aligned}|\partial_1^{N+1}\,\theta_k * G(x)| &\leq \int |\partial_1^{N+1}\,\theta_k(y)|\,|G(x - y)|\,dy \\ &\leq C_p\,(1 + |k|)^p\end{aligned}$$

pour tout $p \in \mathbb{N}$, puisque G est localement intégrable.

Si $|x - k| > 20$, alors $|x - y| > \frac{1}{2}\,|x - k|$ quand $y \in$ Supp θ_k. On en déduit

$$\begin{aligned}|\partial_1^{N+1}\,\theta_k * G(x)| &\leq \int |\theta_k(y)|\,|\partial_1^{N+1}\,G(x - y)|\,dy \\ &\leq C_p\,(1 + |k|)^p\,|x - k|^{-n+m-N-1}.\end{aligned}$$

Finalement, on obtient pour tout x

$$|G * \psi_\varepsilon(x)| \leq C_p \sum_{k \in \mathbb{Z}^n} (1 + |k|)^{-p}\,(1 + |x - k|)^{-n+m-N-1}.$$

Choisissant $p \geq n - m + N + 1$, et pourvu que $N \geq m$, il vient

$$|G * \psi_\varepsilon(x)| \leq C\,(1+|x|)^{-n+m-N-1}.$$

Les estimations sur les dérivées de $G * \psi_\varepsilon$ s'obtiennent de façon analogue, quitte à remplacer G par l'une de ses dérivées partielles. Un examen attentif de l'argument montre que l'ordre maximal autorisé de dérivation est égal à $m + r - 1$. Ceci achève la preuve du lemme.

On l'applique alors à $G(x) = |x|^{-n+1}$, ce qui permet d'estimer $\Lambda^{-1}\,\psi_\lambda$. Puisque $R_\alpha = -\partial_\alpha\,\Lambda^{-1}$, et $\Lambda = \sum_\alpha \partial_\alpha\,R_\alpha = -\Delta\,\Lambda^{-1}$, on retrouve les estimations bien connues sur $R_\alpha\,\psi_\lambda$ et $\Lambda\,\psi_\lambda$, qui s'écrivent

(17) si $|\gamma| \leq r - 1$, il existe C_γ telle que

$$\forall\, x \in \mathrm{I\!R}^n, \quad |\partial^\gamma\, R_\alpha\,\psi_\lambda\,(x)| \leq C_\gamma\, 2^{jn/2}\, 2^{j|\gamma|}\,(1+2^j|x-\lambda|)^{-n-N-1-|\gamma|},$$

(18) si $|\gamma| \leq r - 2$, il existe C_γ telle que

$$\forall x \in \mathrm{I\!R}^n, \quad |\partial^\gamma\,\Lambda\,\psi_\lambda\,(x)| \leq C_\gamma\, 2^{jn/2}\, 2^{j(1+|\gamma|)}\,(1+2^j|x-\lambda|)^{-n-N-2-|\gamma|}.$$

Pour finalement obtenir les estimations sur $\pi(R_\alpha\,\psi_\lambda)$ et $\pi(\Lambda\,\psi_\lambda)$, on utilise le lemme suivant, qui décrit l'action de la projection π sur des fonctions "basses fréquences."

LEMME 3.9. *Soient f_λ, où $\lambda \in D$ et $j(\lambda) \leq -1$, une famille de fonctions telle que pour tout $|\gamma| \leq q$, où q est un entier $\leq r$, il existe une constante C_γ et un réel $s_\gamma >$ 0 vérifiant les estimations:*

$$\forall x \in \mathrm{I\!R}^n, \quad |\partial^\gamma\, f_\lambda\,(x)| \leq C_\gamma\, 2^{jn/2}\, 2^{j|\gamma|}\,(1+2^j|x-\lambda|)^{-n-s_\gamma}.$$

Alors, la famille des projections $\pi(f_\lambda)$ vérifient des estimations analogues, où seules les constantes C_γ sont changées.

On commence par montrer l'estimation sur $\pi(f_\lambda)$, qu'on écrit

$$\pi(f_\lambda)(x) = \sum_k \int f_\lambda(y)\,\overline{\varphi(y-k)}\,dy \quad \varphi(x-k).$$

Le noyau intégral de la projection, $\sum_k \overline{\varphi(y-k)}\,\varphi(x-k)$, est majoré en module par $C_p\,(1+|x-y|)^{-p}$, quel que soit $p > 0$. On a donc, choisissant $p = n + s_0$,

$$\begin{aligned}|\pi(f_\lambda)(x)| &\leq C\,2^{jn/2} \int (1+2^j|y-\lambda|)^{-n-s_0}\,(1+|x-y|)^{-n-s_0}\,dy\\ &\leq C\,2^{jn/2}\,(1+2^j|x-\lambda|)^{-n-s_0},\end{aligned}$$

puisque $j \leq -1$.

Pour montrer l'estimation voulue sur $\partial^\gamma\,\pi(f_\lambda)$, où $|\gamma| \leq q$, on utilise le résultat de commutation de Lemarié [4], qui s'écrit

(19) $$\partial^\gamma \pi = \pi_\gamma\,\partial^\gamma,$$

où π_γ est une projection sur un sous–espace V_γ, non orthogonale. Plus précisément, il existe φ_γ et $\widetilde{\varphi}_\gamma$, deux fonctions à décroissance rapide (et qui engendrent deux AMR biorthogonales), telles que

$$\pi_\gamma\, f(x) = \sum_k \int f(y)\overline{\widetilde{\varphi}_\gamma(y-k)}\,dy \quad \varphi_\gamma(x-k).$$

Ceci permet de reproduire les estimations précédentes, et ainsi de terminer la preuve du Lemme 3.9.

Partant des décompositions (14) et (14'), et appliquant les Lemmes 2.6, 2.7, 2.8, et 2.9, on obtient le lemme suivant.

LEMME 3.10. *Les familles de fonctions* $R^V_\alpha\,\psi_\lambda$ *et* $\Lambda^V\,\psi_\lambda$, *avec* $j(\lambda) \le -1$, *vérifient les estimations* (17) *et* (18), *respectivement. La famille des fonctions* $S^V_\alpha\,\psi_\lambda$, *avec* $j(\lambda) \le -1$, *vérifie les estimations suivantes:*
(20)
pour tout γ, $|\gamma| \le r$, *il existe une constante* C_γ *telle que, pour presque tout* $x \in \mathbb{R}^n$,

$$|\partial^\gamma\, S^V_\alpha\,\psi_\lambda(x)| \le C_\gamma\, 2^{jn/2}\, 2^{j|\gamma|}\,(1 + 2^j|x-\lambda|)^{-n-2N+1}.$$

On obtiendrait un résultat analogue sur les $R^V_{\alpha,\beta}\,\psi_\lambda$, que nous n'utiliserons pas. Remarquons, pour terminer cette section, que les opérateurs étudiés étant des convolutions, les fonctions images des ondelettes ψ_λ annulent les mêmes moments, sous la condition que ces moments existent.

3.2.3. Commutateurs. Les commutateurs dont nous avons besoin sont $[R^V_\alpha,\ \pi\, b\, \pi^*]$ et $[S^V_\alpha,\ \pi\, b\, \pi^*]$. Comme ce sont des opérateurs de V dans lui-même et que V est inclus dans H^1 (et même H^2), affirmer qu'ils sont régularisants d'ordre 1 n'a pas un sens très clair. C'est pourquoi l'analogue du théorème de Calderón prend la forme suivante.

LEMME 3.11. *Les opérateurs* $[R^V_\alpha,\ \pi\, b\, \pi^*]\,\Lambda^V$ *et* $[S^V_\alpha,\ \pi\, b\, \pi^*]\,\Lambda^V$, *agissant de* V *dans* V, *ont une norme majorée par* $C\,\|\nabla b\|_\infty$, *où* C *est une constante ne dépendant que de* V.

(Noter que, par scaling, ce lemme se transpose à V_j, pour tout j, avec la même constante C.)

La stratégie de la preuve est la même que pour étudier le premier commutateur de Calderón. On munit V de sa base d'ondelettes (ψ_λ), où $j(\lambda) \le -1$. Si $\Gamma = [T,\ \pi\, b\, \pi^*]\,\Lambda^V$ est le commutateur considéré, $T = R^V_\alpha$ ou S^V_α, on calcule l'image de chaque ψ_λ:

$$T(\pi\, b\,\Lambda^V(\psi_\lambda)) - \pi\, b\, T\,\Lambda^V(\psi_\lambda).$$

Remplaçant b par $b - b(\lambda)$, on obtient

$$\begin{aligned}\Gamma(\psi_\lambda) &= T\pi([b-b(\lambda)]\,\Lambda^V\,(\psi_\lambda))\\ &\quad - \pi[b-b(\lambda)]\,T\,\Lambda^V\,(\psi_\lambda).\end{aligned}$$

Soient alors les opérateurs U et U_T définis par

$$U(\psi_\lambda) = [b-b(\lambda)]\,\Lambda^V\,(\psi_\lambda)$$

et

$$U_T(\psi_\lambda) = [b-b(\lambda)]\,T\,\Lambda^V\,(\psi_\lambda).$$

L'estimation sur la norme de Γ résultera de l'estimation analogue sur les normes de U et de U_T, obtenue en appliquant le lemme fondamental.

Commençons par vérifier les estimations standard.

Grâce à la majoration

$$|b(x) - b(\lambda)| \leq \|\nabla b\|_\infty \, |x - \lambda|, \tag{21}$$

celles-ci découlent, pour la famille des $U(\psi_\lambda)$, des inégalités (18) sur les $\Lambda^V(\psi_\lambda)$, démontrées au Lemme 3.10.

Si $T = R^V_\alpha$, on a $T\,\Lambda^V = -i\,\pi\,\partial_\alpha\,\pi^*$, et les estimations standard sur les $U_T\,(\psi_\lambda)$ proviennent de (21) et du Lemme 3.9.

Si enfin $T = S^V_\alpha$, on applique les Lemmes 3.6 et 3.7 aux fonctions $\Lambda^V\,(\psi_\lambda)$.

Il faut maintenant estimer les intégrales des fonctions $U(\psi_\lambda)$ et $U_T(\psi_\lambda)$, et les constantes de Carleson associées.

Avec $T = Id$, R^V_α ou S^V_α, on calcule donc

$$\int b\,T\,\Lambda^V\,(\psi_\lambda).$$

Utilisant (14) et le fait que T, T^V_0 et $\pi\,\Lambda\,\pi^*$ commutent, cette intégrale se réécrit

$$\int \Lambda(\pi b)\,T^V_0\,T(\psi_\lambda). \tag{22}$$

Or, des relations (19) de Lemarié, on déduit que πb est lipschitzienne, de norme lipschitz contrôlée par celle de b. Comme on a

$$\Lambda(\pi b) = \sum_\alpha R_\alpha\,\partial_\alpha\,\pi b,$$

$\Lambda(\pi b)$ appartient à BMO, de norme majorée par $C\|\nabla b\|_\infty$, où C ne dépend que de V.

D'autre part, grâce aux Lemmes 3.6, 3.7, et 3.10, les fonctions $T^V_0\,T(\psi_\lambda)$ vérifient les estimations standard, et sont de plus d'intégrale nulle.

Il est alors très classique de montrer que les intégrales (22) vérifient l'inégalité de Carleson requise par le lemme fondamental, avec une constante contrôlée par la norme de $\Lambda(\pi b)$ dans BMO (voir [5]).

Ceci permet d'achever la preuve du Lemme 3.11, et conclut notre étude des transformées de Riesz adaptées à V.

3.3. Preuve du Théorème 3.2. Considérons maintenant $f \in V\,(= V_0)$, et estimons la norme de $\pi\,b_{\alpha\beta}\,\partial_\alpha\,\partial_\beta\,f$. Elle se laisse trivialement dominer par $C\|\Delta f\|$, mais nous voulons comparer à $\pi\Delta f$ au lieu de Δf d'une part, et d'autre part obtenir la meilleure constante C, c'est-à-dire $\|B\|$, quitte à ajouter au majorant un terme d'ordre 1.

On commence par échanger la projection et la multiplication, en écrivant

$$\begin{aligned}\pi\,b_{\alpha\beta}\,\partial_\alpha\,\partial_\beta\,f = {}& \pi\,b_{\alpha\beta}\,\pi^*\,\pi\,\partial_\alpha\,\partial_\beta\,f \\ & - [b_{\alpha\beta},\,\pi^*\,\pi]\,\partial_\alpha\,\partial_\beta\,f.\end{aligned}$$

Le second terme se traite facilement.

LEMME 3.12. *Pour toute fonction b lipschitzienne et pour tout α, l'opérateur $[b,\,\pi^*\pi]\,\partial_\alpha$ est continu, de norme contrôlée par $C\,\|\nabla b\|_\infty$, où C ne dépend que de V.*

Ce résultat se démontre exactement comme le Lemme 3.11, avec d'ailleurs un peu moins de complications techniques. Nous en laissons donc l'écriture au lecteur.

On a ainsi

$$\|[b_{\alpha\beta},\pi^*\pi]\,\partial_\alpha\,\partial_\beta\,f\| \leq C\,\|B\|_{\mathrm{lip}}\,\|\nabla f\|.$$

Le terme principal $\pi\, b_{\alpha\beta}\, \pi^*\pi\, \partial_\alpha\, \partial_\beta\, f$ s'écrit en utilisant les transformées de Riesz adaptées à V:

$$\begin{aligned}\pi\, b_{\alpha\beta}\, \pi^*\pi\, \partial_\alpha\, \partial_\beta\, f &= \pi\, b_{\alpha\beta}\, \pi^*\, R^V_{\alpha\beta}\, (\pi\Delta f)\\ &= \pi\, b_{\alpha\beta}\, \pi^*\, R^V_\alpha\, R^V_\beta\, (\pi\Delta f)\\ &\quad + \pi\, b_{\alpha\alpha}\, \pi^*\, (S^V_\alpha)^2\, (\pi\Delta f).^4\end{aligned}$$

Comme dans la preuve du Lemme 3.1, on a maintenant

$$\begin{aligned}\pi\, b_{\alpha\beta}\, \pi^*\pi\, \partial_\alpha\, \partial_\beta\, f &= R^V_\alpha\, (\pi\, b_{\alpha\beta}\, \pi^*)\, R^V_\beta\, (\pi\Delta f)\\ &\quad + S^V_\alpha\, (\pi\; b_{\alpha\alpha}\, \pi^*)\, S^V_\beta\, (\pi\Delta f)\\ &\quad + [\pi\, b_{\alpha\beta}\, \pi^*,\; R^V_\alpha]\, R^V_\beta (\pi\Delta f)\\ &\quad + [\pi\, b_{\alpha\alpha}\, \pi^*,\; S^V_\alpha]\, S^V_\alpha (\pi\Delta f).\end{aligned}$$

D'après le Lemme 3.11, les deux derniers termes sont majorés en norme par $C\|B\|_{\text{lip}}\, \|\nabla f\|$.

Pour estimer les deux premiers, on définit la matrice $\widetilde{B}$, dans $M_{2n}(\mathbb{C})$, par

$$\widetilde{B} = \left(\begin{array}{cccccc} b_{11} & & 0 & \vdots & & \\ & \ddots & & \vdots & & 0 \\ 0 & & b_{nn} & \vdots & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & & & \vdots & & B \\ & & & \vdots & & \end{array}\right).$$

On démontre, exactement comme dans (8), que l'opérateur

$$R^V_\alpha\, (\pi\, b_{\alpha\beta}\, \pi^*)\, R^V_\beta + S^V_\alpha\, (\pi\, b_{\alpha\alpha}\, \pi^*)\, S^V_\alpha$$

est de norme majorée par $\|\widetilde{B}\|$: on utilise pour cela le Lemme 3.5, et le fait que la projection π, étant orthogonale, est de norme 1. A ce stade, nous avons obtenu l'estimation

$$\|\pi\, b_{\alpha\beta}\, \partial_\alpha\, \partial_\beta\, f\| \leq \|\widetilde{B}\|\, \|\pi\Delta f\| + C\, \|B\|_{\text{lip}}\, \|\nabla f\|.$$

Pour conclure, il ne reste plus qu'à remarquer le résultat amusant suivant.

LEMME 3.13.

$$\|\widetilde{B}\| = \|B\|.$$

Le Théorème 3.2 est maintenant complètement démontré. Remarquons qu'il se généralise directement au cas où les coefficients $b_{\alpha\beta}$ sont de régularité $\varepsilon > 0$. On obtient alors

$$\|\pi\, b_{\alpha\beta}\, \partial_\alpha\, \partial_\beta\, f\| \leq \|B\|\, \|\pi\Delta f\| + C\, \|B\|_{C^\varepsilon}\, \|f\|_{\dot{H}^{2-\varepsilon}},$$

où $\dot{H}^{2-\varepsilon}$ désigne l'espace de Sobolev homogène d'ordre $2-\varepsilon$.

[4] Dans le cas des AMR non tensorielles, l'algèbre est légèrement plus compliquée, mais sans que la preuve change de nature.

4. Inversion des opérateurs strictement accrétifs du second ordre à coefficients dominants lipschitziens. Revenons à l'opérateur $L = -\mathrm{div}A\nabla + b$, satisfaisant toutes les hypothèses exposées à la première partie: nous devons prouver que sont également satisfaites (H2) et (H3), afin d'appliquer le schéma d'inversion du Théorème 1.1, et démontrer les Théorèmes 2.2 et 2.3.

Rappelons que la stratégie générale d'inversion de L que nous suivons consiste à découpler les basses et les hautes fréquences, en écrivant

$$L^2 = V_{j_0} \oplus \bigoplus_{j \geq j_0} W_j,$$

et en approchant l'opérateur L par la méthode de Galerkin sur V_{j_0}, par une parametrix sur $V_{j_0}^{\perp}$.

Nous étudions d'abord les approximations Galerkin, prouvant notamment le Théorème 2.3, puis nous construisons la parametrix. Nous supposons donnée une AMR (V_j) satisfaisant aux hypothèses du Théorème 2.2.

4.1. Action de L sur V_j et opérateurs de Galerkin. Reprenant les notations du Théorème 2.1, nous désignons par $\pi_j : L^2 \to V_j$ et $\pi_j^* : V_j \to L^2$ les opérateurs de restriction (par projection orthogonale) et de prolongement associés à V_j. Par hypothèse, l'opérateur $\pi_j L \pi_j^*$, de V_j dans lui-même, est inversible: nous notons Γ_j l'extension naturelle à L^2 de son inverse $(\pi_j L \pi_j^*)^{-1}$. Γ_j est l'opérateur de Galerkin associé à L et à V_j.

On a classiquement

$$\forall f \in L^2, \quad \lim_{j \to +\infty} \|L^{-1}f - \Gamma_j f\|_{H^1} = 0,$$

et le Théorème 2.3 affirme que la même relation est vraie pour la norme H^2. Ce résultat est plus fort que l'hypothèse (H2), qui affirme seulement que les opérateurs $L\Gamma_j$ sont uniformément bornés.

Ces propriétés vont résulter du corollaire suivant du Théorème 3.2.

THÉORÈME 4.1. *Il existe une constante C, ne dépendant que de L et de l'AMR (V_j), telle que pour tout $j \in \mathbb{Z}$ et pour toute $f \in V_j$, on ait*

$$\|Lf\| \leq C\,(\|\pi_j L f\| + \|f\|_{H^1}) \tag{23}$$

En termes plus intuitifs, cette inégalité signifie que la régularité des coefficients est suffisante pour que l'action de L sur V_j n'induise pas un "éclatement" incontrôlé par $\pi_j L f$, c'est-à-dire sa projection basses fréquences.

Pour démontrer ce résultat, on commence par l'établir dans le cas où les coefficients sont constants, qui se réduit à l'observation suivante.

LEMME 4.2. *Si A est constante, $j \in \mathbb{Z}$ et $f \in V_j$, alors*

$$\|L_0 f\| \leq C\|\pi_j L_0 f\|,$$

où $L_0 = -\mathrm{div}A\,\nabla$, et où C ne dépend que de A et de l'AMR.

Il suffit de faire le calcul avec la transformée de Fourier, en utilisant les résultats du paragraphe 3.2.1, pour prouver ce lemme.

Dans le cas général, on commence par renormaliser A: si

$$t = \frac{\delta}{\|A\|^2}, \qquad \|I - tA\| \leq \left(1 - \frac{\delta^2}{\|A\|^2}\right)^{1/2},$$

où δ est la constante d'accrétivité de A. On peut donc se ramener au cas où $A = I - B$, avec $\|B\| < 1$.

Notant toujours L_0 l'opérateur $-\mathrm{div} A\,\nabla$, le Théorème 3.2 donne l'inégalité

$$\|\pi_j\, L_0\, f\| \geq (1 - \|B\|)\, \|\pi_j\, \Delta\, f\| - C\, \|B\|_{\text{lip}}\, \|\nabla\, f\|.$$

Comme on a d'autre part

$$\|L_0 f\| \leq C\, \|\Delta f\| \leq C'\, \|\pi_j\, \Delta\, f\|,$$

on obtient le Théorème 3.2 pour L_0, et donc pour L. Une version précisée en est d'ailleurs

$$\|L_0\, f\| \leq C\, \frac{\|A\|^2}{\delta^2}\, (\|\pi_j\, L_0\, f\| + \|A\|_{\text{lip}} \|\nabla\, f\|),$$

où C ne dépend que de l'AMR. Si A est auto-adjointe, $\frac{\|A\|}{\delta}$ n'est autre que le conditionnement de A, $K(A) = \|A\|\|A^{-1}\|$, à condition bien sûr que δ soit la meilleure constante d'accrétivité, et dans le cas général, on a seulement l'inégalité $\frac{\|A\|}{\delta} \geq K(A)$.

Enfin, comme le Théorème 3.2, le Théorème 4.1 se généralise au cas où L s'écrit

$$L = -a_{\alpha\beta}\, \partial_\alpha\, \partial_\beta + b_\alpha\, \partial_\alpha + b,$$

avec $a_{\alpha\beta} \in C^\varepsilon$, $\varepsilon > 0$.

Le Théorème 4.1 implique immédiatement que les opérateurs $L\,\Gamma_j$ sont uniformément bornés. Si $f \in V_j$, on applique l'inégalité (23) à $g_j = \Gamma_j\, f$: on sait que $\|g_j\|_{H^1} \leq C\, \|f\|$, où C ne dépend que de L, et on a d'autre part $\pi_j\, L\, g_j = f$, d'où finalement $\|L\, g_j\| \leq C\, \|f\|$, qui est le résultat désiré.

La démonstration du Théorème 2.3 est à peine moins difficile. On se donne $f \in L^2$, on pose $g = L^{-1}\, f$, $g_j = \Gamma_j\, f$ et il s'agit de montrer que

$$\lim_{j \to +\infty} \|g - g_j\|_{H^2} = 0.$$

Décomposons $g - g_j$ en $\pi_j^\perp\, g + \pi_j g - g_j$, où $\pi_j^\perp$ est la projection orthogonale sur $V_j^\perp$.

Puisque $g \in H^2$ et que les ondelettes choisies sont base inconditionnelle de H^2, on a

$$\lim_{j \to +\infty} \|\pi_j^\perp g\|_{H^2} = 0. \tag{24}$$

Il reste à estimer la norme H^2 de $\pi_j g - g_j$, qui est équivalente à la norme L^2 de $L(\pi_j g - g_j)$ (nous laissons au lecteur la preuve de cette remarque). Mais l'inégalité (23) nous donne

$$\|L(\pi_j g - g_j)\| \leq C\, (\|\pi_j L(\pi_j g - g_j)\| + \|\pi_j g - g_j\|_{H^1}).$$

On a

$$\lim_{j \to +\infty} \|\pi_j g - g_j\|_{H^1} = 0.$$

D'autre part, par construction, $\pi_j\, L\, g_j = \pi_j\, f = \pi_j\, L\, g$, d'où

$$\pi_j\, L\, (\pi_j g - g_j) = -\pi_j\, L\, \pi_j^\perp\, g.$$

On déduit alors de (24) que

$$\lim_{j \to +\infty} \|\pi_j\, L\, (\pi_j g - g_j)\| = 0,$$

ce qui termine la preuve du Théorème 2.3.

4.2. Construction d'une parametrix. Pour conclure, c'est-à-dire démontrer le Théorème 2.2, il faut établir (H3), sous une forme suffisamment précise pour pouvoir en déduire l'inégalité (5).

La parametrix que nous allons construire est définie sur $V_{j_0}^{\perp}$, où j_0 est fixé, en associant à chaque ondelette de base ψ_λ, $j(\lambda) \geq j_0$, la fonction θ_λ, définie par

$$-\mathrm{div} A(\lambda) \nabla \theta_\lambda = \psi_\lambda.$$

THÉORÈME 4.3. *Si j_0 est fixé, l'opérateur P_{j_0} (dépendant de j_0), défini par*

$$P_{j_0}(\psi_\lambda) = \theta_\lambda \quad si\; j(\lambda) \geq j_0,$$
$$P_{j_0}(f) = 0 \quad si\; f \in V_{j_0},$$

est une parametrix à droite de L, au sens suivant.

(i) *L'opérateur P_{j_0} est continu de L^2 dans H^2, et sa norme est uniformément majorée par rapport à j_0;*

(ii) *l'opérateur $L\,P_{j_0}$ se décompose en*

$$L\,P_{j_0} = \pi_{j_0}^{\perp} - R_{j_0},$$

où $\pi_{j_0}^{\perp}$ est la projection orthonormale sur $V_{j_0}^{\perp}$, et où l'opérateur R_{j_0} est borné sur L^2 en norme par $C\,2^{-j_0}$, la constante C ne dépendant pas de j_0.

Ce théorème implique que L vérifie l'hypothèse (H3), l'inégalité (5) provenant de l'estimation sur la norme de R_{j_0}, et de la preuve du Théorème 2.1.

La définition de P_{j_0} est fondée sur l'idée classique du gel des coefficients dominants. On tire ici partie de la localisation des ψ_λ, qui permet de bien adapter le point de collocation des coefficients.

Il est cependant important de comprendre que la seule localisation des ondelettes dans l'espace physique $\mathbb{R}^n$ n'est pas suffisante. En effet, pour que le Théorème 4.3 puisse être vrai, il est au moins nécessaire que chaque θ_λ approche la fonction $L^{-1}(\psi_\lambda)$: c'est le cas, comme on va le voir, grâce aux propriétés d'oscillation des ψ_λ. (En d'autres termes, le même procédé appliqué aux fonctions $\varphi_{j_0 k}$ ne donnerait pas du tout le même résultat.)

Enfin, soulignons que P_{j_0} est une parametrix en un sens plutôt faible. En l'absence de régularité des coefficients $\partial_\alpha\, a_{\alpha\beta}$ et b, le reste R_{j_0} n'est pas régularisant. Néanmoins, nous verrons qu'il se décompose en

$$R_{j_0} = R_{j_0,0} + \partial_\alpha(a_{\alpha\beta}) R_{j_0,\beta} + bP_{j_0},$$

où les opérateurs $R_{j_0,0}$ et $R_{j_0,\beta}$ sont régularisants d'ordre 1, et P_{j_0}, d'après le théorème, est régularisant d'ordre 2.

Démontrons maintenant le Théorème 4.3 et, pour commencer, décrivons les θ_λ (soulignons que, dans cette notation, λ a la signification usuelle d'un indice, et ne signifie pas qu'il existe θ_ε telle que $\theta_\lambda(x) = 2^{jn/2}\,\theta_\varepsilon(2^j x - k)$).

LEMME 4.4. *Il existe une constante C telle que, pour tout λ et pour tout x,*

$$|\theta_\lambda(x)| \leq C\,2^{jn/2}\,4^{-j}\,(1 + 2^j|x-\lambda|)^{-n-N+1},$$

et plus généralement, pour tout multi-indice γ, $|\gamma| \leq r+1$, il existe C_γ telle que

$$|\partial^\gamma \theta_\lambda(x)| \leq C_\gamma\,2^{jn/2}\,2^{j(|\gamma|-2)}\,(1 + 2^j|x-\lambda|)^{-n-N+1-|\gamma|}.$$

Ce n'est qu'une conséquence directe du Lemme 3.8, puisque $\theta_\lambda = G_\lambda * \psi_\lambda$, où G_λ est la fonction de Green de l'opérateur $-\mathrm{div} A(\lambda)\,\nabla$, qui vérifie, uniformément en λ, les estimations (16), avec $m = 2$.

En utilisant le lemme fondamental, on en déduit sans difficulté la continuité de l'opérateur P_{j_0} de L^2 dans H^2, uniformément par rapport à j_0.

Calculons alors $L\,P_{j_0}$. Si ψ_λ est une ondelette, telle que $j(\lambda) \geq j_0$ (parce que sinon $P_{j_0}\,\psi_\lambda = 0$), on a

$$\begin{aligned} L\,P_{j_0}\,\psi_\lambda &= L\,\theta_\lambda \\ &= \psi_\lambda - R_{j_0}(\psi_\lambda), \end{aligned}$$

avec

$$\begin{aligned} R_{j_0}\,(\psi_\lambda) = [a_{\alpha\beta} - a_{\alpha\beta}(\lambda)]\,\partial_\alpha\,\partial_\beta\,\theta_\lambda \\ - \partial_\alpha(a_{\alpha\beta})\,\partial_\beta\,\theta_\lambda + b\,\theta_\lambda. \end{aligned}$$

Les opérateurs qui à ψ_λ associent, respectivement, $\partial_\beta\,\theta_\lambda$ et θ_λ sont continus sur L^2 et majorés en norme par $C\,2^{-j_0}$ et $C\,4^{-j_0}$ (appliquer le lemme fondamental et les estimations du Lemme 4.4).

Quant à l'opérateur qui à ψ_λ associe $[a_{\alpha\beta}(\lambda) - a_{\alpha\beta}]\,\partial_\alpha\,\partial_\beta\,\theta_\lambda$, il est régularisant d'ordre 1, et continu sur L^2 de norme majorée par $C\,2^{-j_0}$: ceci se démontre à nouveau grâce au lemme fondamental, en suivant la même démarche que pour l'étude des commutateurs. Nous laissons au lecteur l'adaptation des arguments utilisés à ce cas particulier.

Finalement, l'opérateur R_{j_0} est borné, et de norme majorée par $C\,2^{-j_0}$, ce qui achève de démontrer le Théorème 4.3, donc le Théorème 2.2. Le Théorème 2.1 s'applique, et permet d'inverser L, au sens de la meilleure topologie possible.

5. Conclusion. Quelques considérations orientées vers les applications numériques pour conclure.

Les méthodes spectrales, très précises, sont d'un emploi mal commode dès qu'il y a des coefficients variables à contenu spectral assez riche dans les équations à résoudre. Même s'il est possible, en théorie, d'utiliser une méthode de Galerkin, le calcul effectif de la matrice associé à l'opérateur $\pi_j L \pi_j^*$ pose de sérieux problèmes.

C'est pourquoi nous pensons que, remplaçant les fonctions spectrales par les ondelettes, il est peut-être possible de tirer partie de la double localisation des ondelettes d'une façon numériquement efficace.

La formule d'inversion du Théorème 2.1 peut servir de base à un algorithme, de nature itérative à cause de la série de Neumann qui intervient. Les résultats d'une première tentative dans ce sens seront ultérieurement publiés. Nous voudrions souligner ici quelques points qui justifient a priori l'intérêt numérique de cette formule:

(i) le seuil de séparation j_0 est indépendant du second membre de l'équation à résoudre et de la précision cherchée;

(ii) de plus, le nombre d'itérations nécessaires à atteindre un gain donné en précision ne dépend pas de la taille de la grille;

(iii) à cause de la forme de la paramétrix P_{j_0}, une version adaptative d'un algorithme d'inversion est possible en principe;

(iv) enfin, la formule d'inversion étant fondée sur un calcul perturbatif, il ne sera pas nécessaire de calculer avec une grande précision les opérateurs Γ_{j_0} et P_{j_0}: l'opérateur

$$L\,(\widetilde{\Gamma}_{j_0} + \widetilde{P}_{j_0}),$$

où le tilde désigne les opérateurs effectivement calculés, reste une perturbation de l'identité.

Malgré d'évidentes différences géométriques, ces quatre propriétés rapprochent notre méthode des méthodes multi-grilles: nous espérons pouvoir prochainement développer la comparaison.

REFERENCES

[1] A. Benassi, S. Jaffard, and D. Roux, *Analyse multi-echelle des processus Gaussiens Markoviens d'ordre p indexés par* (0,1), C. R. Acad. Sci. Paris Sér. I Math., 313 (1991), pp. 403–406.

[2] G. Beylkin, R. Coifman, and V. Rokhlin, *Fast wavelet transforms and numerical algorithms* I, Comm. Pure Appl. Math., XLIV (1991), pp. 147–183.

[3] I. Daubechies, *Ten Lectures on Wavelets*, CBMS–NSF Regional Conference Series in Applied Mathematics 61, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[4] P. G. Lemarié-Rieusset, *Ondelettes vecteurs à divergence nulle*, Rev. Mat. Iberoamericana, 8 (1992), pp. 91–107.

[5] Y. Meyer, *Ondelettes et opérateurs, tome* 1: *Ondelettes*, Hermann, Paris, 1990.

[6] ———, *Ondelettes et opérateurs, tome* 2: *Opérateurs de Calderón–Zygmund*, Hermann, Paris, 1990.

[7] Y. Meyer and R. R. Coifman, *Ondelettes et opérateurs, tome* 3: *Opérateurs multi linéaires*, Hermann, Paris, 1991.

[8] P. Tchamitchian, *Bases d'ondelettes et intégrales singulières: Analyse des fonctions et calcul sur les opérateurs*, Habilitation à diriger des recherches, Faculté de Luminy, Université d'Aix-Marseille II, Marseille, France, 1989.

SIAM J. MATH. ANAL.
Vol. 27, No. 6, pp. 1704–1737, November 1996

ON QUASI-PERIODIC PERTURBATIONS OF ELLIPTIC EQUILIBRIUM POINTS*

ÀNGEL JORBA[†] AND CARLES SIMÓ[‡]

Abstract. This work focuses on quasi-periodic time-dependent perturbations of ordinary differential equations near elliptic equilibrium points. This means studying

$$\dot{x} = (A + \varepsilon Q(t, \varepsilon))x + \varepsilon g(t, \varepsilon) + h(x, t, \varepsilon),$$

where A is elliptic and h is $\mathcal{O}(x^2)$. It is shown that, under suitable hypothesis of analyticity, nonresonance and nondegeneracy with respect to ε, there exists a Cantorian set $\mathcal{E}$ such that for all $\varepsilon \in \mathcal{E}$ there exists a quasi-periodic solution such that it goes to zero when ε does. This quasi-periodic solution has the same set of basic frequencies as the perturbation. Moreover, the relative measure of the set $[0, \varepsilon_0] \setminus \mathcal{E}$ in $[0, \varepsilon_0]$ is exponentially small in ε_0. The case $g \equiv 0$, $h \equiv 0$ (quasi-periodic Floquet theorem) is also considered.

Finally, the Hamiltonian case is studied. In this situation, most of the invariant tori that are near the equilibrium point are not destroyed but only slightly deformed and "shaken" in a quasi-periodic way. This quasi-periodic "shaking" has the same basic frequencies as the perturbation.

Key words. quasi-periodic perturbations, elliptic points, quasi-periodic solutions, small divisors, quasi-periodic Floquet theorem, Kolmogorov–Arnold–Moser (KAM) theory

AMS subject classifications. 34C27, 34C50, 58F27, 58F30

1. Introduction.

In this work, we will consider autonomous differential equations under quasi-periodic time-dependent perturbations near an elliptic equilibrium point. The kind of equation with which we shall deal is

$$\dot{x} = (A + \varepsilon Q(t, \varepsilon))x + \varepsilon g(t, \varepsilon) + h(x, t, \varepsilon),$$

where A is assumed to be elliptic (that is, all the eigenvalues are purely imaginary and nonzero), h is of second order in x, and the system is autonomous when $\varepsilon = 0$. At this point, we recall the definition of a quasi-periodic function.

DEFINITION 1.1. *A function f is a quasi-periodic function with basic frequencies $\omega_1, \ldots, \omega_r$ if $f(t) = F(\theta_1, \ldots, \theta_r)$, where F is 2π periodic in all its arguments and $\theta_j = \omega_j t$ for $j = 1, \ldots, r$.*

We assume that the quasi-periodic functions appearing in our equations are analytical. For definiteness we give the following definition.

DEFINITION 1.2. *A function f is analytic quasi-periodic on a strip of width ρ if it is quasi-periodic and F (see Definition 1.1) is analytical for $|\mathrm{Im}\ \theta_j| \le \rho$ for $j = 1, \ldots, r$. In this case, we denote by $\|f\|_\rho$ the norm*

$$\sup\{|F(\theta_1, \ldots, \theta_r)| \text{ with } |\mathrm{Im}\ \theta_j| \le \rho,\ 1 \le j \le r\}.$$

This kind of equation appears in many problems. As an example, we can consider the equations of the motion near the equilateral libration points of the earth–moon system, including (quasi-periodic) perturbations coming from the noncircular motion

* Received by the editors November 9, 1994; accepted for publication (in revised form) August 3, 1995. This research was supported by DGICYT (Spain) grant PB94-0215, EC grant ERBCHRXCT-940460, and CIRIT (Catalunya) grant GRQ93-1135.

[†] Departament de Matemàtica Aplicada I, ETSEIB, Universitat Politècnica de Catalunya, Diagonal 647, 08028 Barcelona, Spain (jorba@ma1.upc.es).

[‡] Departament de Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Gran Via 585, 08007 Barcelona, Spain (carles@maia.ub.es).

of the moon and the effect of the sun. See [13], [5], [11], [12], [10], and [15]. In these works, some seminumerical methods were applied to compute a quasi-periodic orbit replacing the equilateral relative equilibrium point (which means that, when the perturbation tends to zero, quasi-periodic orbit tends to the libration point), but there is a lack of theoretical support to ensure that the methods used are really convergent and the computed quasi-periodic orbit really exists. In §2, the existence of that dynamical equivalent is shown for a Cantorian set (of positive measure) of values of ε. Another problem related to this is the study of the stability of that quasi-periodic solution. In order to do this, a kind of Floquet theory is available (see [16]) which now can be obtained as a result of the more general study presented here (see §2).

We also want to note that the Floquet theorem for the quasi-periodic case has already been considered in many papers. An approach similar to ours (based on Kolmogorov–Arnold–Moser (KAM) techniques) can be found in [3]. There the reducibility to constant coefficients is studied for the case in which A is a hyperbolic matrix. For the case in which A is elliptic, some bounds are given on the measure of the set of matrices Q for which the system can be reduced to constant coefficients. The bounds on that measure, however, are not as good as those that can be derived from the work presented here.

Another approach to the reducibility of quasi-periodic linear equations can be found in [14]. The methods used there are not based on KAM techniques, and the results can be applied to systems that are not close to constant coefficients. The main drawback is that the hypotheses used are quite restrictive[1] and are very difficult to check in a practical example.

Finally, it is interesting to consider the Hamiltonian case. In §3, we show that most of the KAM tori of the autonomous system still persist when the quasi-periodic time-dependent perturbation is added.

Studies of this kind for the case of the one-dimensional Schrödinger equation can be found in the literature—see, for instance, [6], [18], [4], or [8]—and some of the methods and ideas used here are already contained in these papers. Note that since the "unperturbed" problem is a harmonic oscillator, it is possible to obtain better results (see, e.g., [8]). Some of the ideas of this paper could be already found in [17] and [7], although they deal with slightly different problems.

2. A dynamical equivalent to elliptic equilibrium points.

In this section, we focus on the equation

$$\dot{x} = (A + \varepsilon Q(t, \varepsilon))x + \varepsilon g(t, \varepsilon) + h(x, t, \varepsilon), \tag{1}$$

where the time dependence is quasi-periodic with vector of basic frequencies $\omega = (\omega_1, \ldots, \omega_r)$ and analytic on a strip of width $\rho_0 > 0$. The reader should recall that h is of second order in x. We want to stress that the equations are not required to be Hamiltonian. (The Hamiltonian case will be considered in §3.)

2.1. The inductive scheme.

To study (1), let us perform some changes to simplify it. First of all, we shall try to eliminate the independent term $g(t)$ by means of quasi-periodic changes of variables. To do this, we shall need a scheme with quadratic convergence. (Otherwise, the small-divisors effect would make the method divergent.) This kind of scheme is based on the Newton method, that is, to linearize the problem

[1] For instance, the system $\dot{x} = (A + \varepsilon Q(t, \varepsilon))x$, where A is elliptic and ε is small, does not satisfy the required hypothesis.

in a known approximation of the solution, solve this linear problem, and take this solution as a new (better) approximation to the solution for which we are looking. These algorithms can overcome the effect of the small divisors and ensure convergence on certain regions. To apply this method to our problem, we must consider the following linearized problem. (We take as an initial guess the zero solution, and we linearize around this point.)

$$\dot{x} = (A + \varepsilon Q(t,\varepsilon))x + \varepsilon g(t,\varepsilon).$$

We are looking for a quasi-periodic solution $\underline{x}(t,\varepsilon)$ whose basic frequencies are those of g and Q such that $\lim_{\varepsilon\to 0} \underline{x}(t,\varepsilon) = 0$. At this point, we note that we do not need to know $\underline{x}(t,\varepsilon)$ exactly because an approximation of order ε is enough. This is another property of the Newton method; we do not need to know the Jacobian matrix exactly but just some approximation of it, and it is enough that this approximation be of the order of the independent term that we want to make zero. In our case, this can be easily done by considering the linear system

$$\dot{x} = Ax + \varepsilon g(t,\varepsilon). \tag{2}$$

Here we need a nonresonance condition. The usual one is

$$|(k,\omega)\sqrt{-1} - \lambda_i| > \frac{c}{|k|^{\gamma_0}}, \tag{3}$$

where λ_i are the eigenvalues of A and $|k| = |k_1| + \cdots + |k_r|$. Condition (3) as well as condition (7), an additional diophantine condition needed later, will be discussed in detail throughout §2.2.

Let us call $\underline{x}(t,\varepsilon)$ the solution of (2) that is quasi-periodic with respect to t (and whose basic frequencies are those of g) and of order ε. The existence of this solution will be shown by Lemma 2.10. Now we can perform the change of variables $x = \underline{x}(t,\varepsilon) + y$ to (1) to obtain

$$\dot{y} = (A + \varepsilon Q_1(t,\varepsilon))y + \varepsilon^2 g_1(t,\varepsilon) + h_1(y,t,\varepsilon), \tag{4}$$

where if $\varepsilon \neq 0$,

$$\begin{aligned}
Q_1(t,\varepsilon) &= Q(t,\varepsilon) + \frac{1}{\varepsilon} D_x h(\underline{x}(t,\varepsilon),t,\varepsilon),\\
g_1(t,\varepsilon) &= \frac{1}{\varepsilon^2} h(\underline{x}(t,\varepsilon),t,\varepsilon) + \frac{1}{\varepsilon} Q(t,\varepsilon)\underline{x}(t,\varepsilon),\\
h_1(y,t,\varepsilon) &= h(\underline{x}(t,\varepsilon) + y,t,\varepsilon) - h(\underline{x}(t,\varepsilon),t,\varepsilon) - D_x h(\underline{x}(t,\varepsilon),t,\varepsilon)y.
\end{aligned}$$

Note that this process cannot be (successfully) iterated. Now we need a solution of

$$\dot{y} = (A + \varepsilon Q_1(t,\varepsilon))y + \varepsilon^2 g_1(t,\varepsilon) \tag{5}$$

with an accuracy of order ε^2; and if we take the kind of approximation given by (2) (that is, dropping Q_1), we will have a divergent scheme. This is because, in this way, one obtains linear convergence in ε, which is overcome by the effect of the small divisors.

To deal with this difficulty, we perform a new change of variables to get something like $\varepsilon^2 Q_2$ instead of εQ_1. This can be done as follows. Let us define the average of Q_1 as

$$\overline{Q}_1(\varepsilon) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} Q_1(t,\varepsilon)\,dt.$$

For the existence of the limit, see, for instance, [9]. Now consider now (5) after averaging with respect to t and some rearrangement:

$$\dot{y} = (\overline{A}(\varepsilon) + \varepsilon\widetilde{Q}_1(t,\varepsilon))y + \varepsilon^2 g_1(t,\varepsilon),$$

where $\widetilde{Q}_1(t,\varepsilon) = Q_1(t,\varepsilon) - \overline{Q}_1(\varepsilon)$, $\overline{A}(\varepsilon) = A + \varepsilon\overline{Q}_1(\varepsilon)$. Now we need to find a quasi-periodic solution of

$$\dot{P} = \overline{A}P - P\overline{A} + \widetilde{Q}_1 \tag{6}$$

with the same basic frequencies as $\widetilde{Q}_1$. This can be done if the eigenvalues of $\overline{A}$ satisfy a diophantine condition. The one used in [16] was

$$|(k,\omega)\sqrt{-1} - \overline{\lambda}_i + \overline{\lambda}_j| > \frac{c}{|k|^{\gamma_0}}. \tag{7}$$

Then, making the change of variables $y = (I + \varepsilon P)z$ (I denotes the identity matrix) to (4) (these changes of variables have already been considered in [3], [16], and [15]), we obtain the equation

$$\dot{z} = (\overline{A}(\varepsilon) + \varepsilon^2 Q_2(t,\varepsilon))z + \varepsilon^2 g_2(t,\varepsilon) + h_2(z,t,\varepsilon), \tag{8}$$

where $Q_2(t,\varepsilon) = (I + \varepsilon P(t,\varepsilon))^{-1}\widetilde{Q}_1 P(t,\varepsilon)$, $g_2(t,\varepsilon) = (I + \varepsilon P(t,\varepsilon))^{-1}g_1(t,\varepsilon)$, and $h_2(z,t,\varepsilon) = (I + \varepsilon P(t,\varepsilon))^{-1}h_1((I + \varepsilon P(t,\varepsilon))z,t,\varepsilon)$. Now, using $\dot{z} = \overline{A}z + \varepsilon^2 g_2(t)$, we are able to find an approximate solution of (8) with an accuracy of order ε^2 that allows us to proceed with the Newton method. In this way, after n steps (each step is composed of the two changes of variables as explained above), the equation will look like

$$\dot{x}_n = (A_n(\varepsilon) + \varepsilon^{2^n} Q_n(t,\varepsilon))x_n + \varepsilon^{2^n} g_n(t,\varepsilon) + h_n(x_n,t,\varepsilon).$$

Then if the norms of A_n, Q_n, g_n, and h_n do not grow too fast with n, the scheme will be convergent to an equation like

$$\dot{y} = A_\infty(\varepsilon)y + h_\infty(y,t,\varepsilon).$$

This equation has the trivial solution $y = 0$, and this shows that in the original system of equations, the origin is replaced by a quasi-periodic orbit whose basic frequencies are those of the perturbation. Note that we have also obtained the linearized flow (given by the "Floquet" matrix A_∞) around this quasi-periodic solution.

2.2. The resonances. We recall that the small divisor conditions needed at each step (to compute the changes of variables) are

$$|(k,\omega)\sqrt{-1} - \lambda_i| > \frac{c}{|k|^{\gamma_0}}, \qquad |(k,\omega)\sqrt{-1} - \lambda_i + \lambda_j| > \frac{c}{|k|^{\gamma_0}}. \tag{9}$$

The first condition is needed to solve equations like (2) and the second one to solve equations like (6). Note that the eigenvalues λ_i are changed at each step of the process (because A is changed), and this implies that we do not know in advance if they will satisfy the diophantine conditions for all the steps.

To deal with this problem, we need to have some control on the variation of the eigenvalues at each step. To explain the main idea, let us focus on (2). Since we are assuming that the eigenvalues of A verify condition (3), at the first step, we can solve the equation and proceed. In the second step, when we need to solve the same equation, we find that the matrix has been (slightly) changed. Now it is $\overline{A}(\varepsilon) = A + \varepsilon\overline{Q}_1(\varepsilon)$. Therefore, since the eigenvalues of $\overline{A}$ are different from those of A, we cannot assure that they satisfy condition (3).

To explain how to overcome this difficulty, let us denote by $\overline{\lambda}_i(\varepsilon)$, $i = 1, \dots, d$, the eigenvalues of the matrix $\overline{A}(\varepsilon)$. Let us write $\overline{\lambda}_i(\varepsilon)$ as

$$\overline{\lambda}_i(\varepsilon) = \lambda_i + \lambda_i^{(1)}\varepsilon + \lambda_i^{(2)}\varepsilon^2 + \cdots, \tag{10}$$

where λ_i is an eigenvalue of the unperturbed matrix A. If we look at $\overline{\lambda}_i(\varepsilon)$ as a function of ε, we can avoid the resonant values of $\overline{\lambda}_i(\varepsilon)$ by avoiding the corresponding values of ε. This implies that taking out a (Cantor-like) set of resonant values of $\overline{\lambda}_i(\varepsilon)$ (this set is the usual union of small intervals centered in the values (k, ω)) is equivalent to taking out the corresponding (by (10)) values of ε. To bound the measure of the "resonant" values of ε, we will require that relation (10) be Lipschitz from below with respect to ε. We also want to note that we need to take out values of ε at each step of the inductive process, so we need to have this condition at each step. Let us examine this. At first sight, it seems enough to ask for $\lambda_i^{(1)} \neq 0$ because this value is produced by the first averaging. Therefore,

1. it is left unchanged by all the others steps of the inductive procedure;
2. it can be computed easily at the beginning (it is a verifiable hypothesis).

The problem is that if we take out a Cantor-like set at each step, the dependence of $\overline{\lambda i}(\varepsilon)$ on ε is not differentiable (because $\overline{\lambda}_i(\varepsilon)$ is defined only on a set with empty interior), and we do not even know if it is continuous. Therefore, it is not obvious how to derive the Lipschitz condition that we need.

To deal with the latter difficulty, we will show explicitly that at each step, relation (10) is Lipschitz. (Note that the definition of Lipschitz holds perfectly on sets with empty interior.) This will allow us to control the measure of the set of ε that we are taking out.

Finally, we want to note that this technique has to be applied twice at each step—once for equations like (2) and once for equations like (6).

2.3. The measure of the resonant set. Another important point is to bound the measure of the set of values of ε too close to resonance. To do this, we will assume that ε belongs to an interval $[0, \varepsilon_0]$, where ε_0 is small enough. We will show is that it is possible to bound the measure of the set of resonant values of ε by a quantity exponentially small in ε_0. To simplify the discussion, we will again focus on (2), so the corresponding small denominator is $(k, \omega)\sqrt{-1} - \lambda_i$.

The usual procedure is to use the bound given by (3) because it is good enough to produce convergence and to give a positive measure set of admissible frequencies. (This has been done in [15] and [16].) Note that the size of the set of resonant values of λ is given by the bound of the small divisors.[2] This implies that we should try to choose that value as small as possible. On the other hand, however, this value will appear in the denominators of the Fourier series. Therefore, if it is too small, we will not be able to prove convergence.

[2] For instance, in the case of (3), the set of resonant values of λ is

$$\bigcup_{|k|\neq 0} B\left((k, \omega)\sqrt{-1}, \frac{c}{|k|^{\gamma_0}}\right),$$

where $B(a, r)$ denotes the ball (in the complex plane) centered in a with radius r.

The condition that we have used is

$$|(k,\omega)\sqrt{-1}-\lambda_i| > \frac{c}{|k|^{\gamma_n}}e^{-\nu_n|k|} = D(k,n),$$

where γ_n has been taken to be equal to $\gamma_0 z^n$ $(1 < z < 2)$ and ν_n is $\frac{\nu_0}{(n+1)^2}$. Here n denotes the actual step of the inductive process. To begin the discussion of this expression, let us remark that the measure of the resonant set of λ at each step n is given by $\sum_{k\neq 0} 2D(k,n)$, $k \in \mathbb{Z}^r$, and the total measure is

$$\sum_{n\geq 0}\sum_{k\in\mathbb{Z}^r\setminus\{0\}} 2\frac{c}{|k|^{\gamma_n}}e^{-\nu_n|k|}. \tag{11}$$

Therefore, a first condition that we need is that those sums are convergent.

Before continuing with the discussion of $D(k,n)$, let us first explain where the exponentially small character (of the set of resonant values of ε) comes from. This will (hopefully) make clear the reasons to choose an expression like $D(k,n)$.

As we stated before, the eigenvalues of the matrix A move at each step of the inductive process by an amount of $\mathcal{O}(\varepsilon)$. Let us call $I_i(\varepsilon)$ the interval (with diameter $\mathcal{O}(\varepsilon)$) where the eigenvalue number i moves. This implies that if the eigenvalues of the unperturbed matrix satisfy a condition like (3), the values (k,ω) are outside $I_i(\varepsilon)$ if $|k| < N(\varepsilon)$ for a suitable value $N(\varepsilon)$. (Another way of saying this is that the values (k,ω) cannot approach λ_i too fast because of (3).) For that reason, we do not need to take out resonances with $|k| < N(\varepsilon)$; and this leads to the fact that in (11) it is enough to start the sum in k when $|k| \geq N(\varepsilon)$. This in turn implies that if the expression $D(k,n)$ decays exponentially with $|k|$, we will obtain something exponentially small with ε_0.

This is the reason for putting something like $\exp(-\nu|k|)$ in $D(k,n)$. Since this value will appear in the denominators of the corresponding Fourier series, we will have the factor $\exp(\nu|k|)$ multiplying the coefficients of those series. This will produce a reduction of the analyticity strip of the series; the width will go from ρ to $\rho-\nu$. Of course, after a few steps, the functions will not be analytic. Thus the entire inductive process will be over. To avoid this problem, we have chosen ν depending on the actual step $\nu_n = \frac{\nu_0}{(n+1)^2}$ in such a way that the total reduction on the analyticity strip remains bounded. (Of course, other selections of ν_n are possible, but they do not change the final result.)

The next step is to realize that with this selection of ν_n, the exponential goes to 1 when n goes to infinity. Thus we need to add some factor in front of the exponential to ensure that the sum with respect to n is still exponentially small. For this reason, we have added the factor $c/|k|^{\gamma_n}$. The selection $\gamma_n = \gamma_0 z^n$ is not the only one (one can use, for instance, $\gamma_n = \gamma_0 n^j$ for some j), but the results with the present choice seem to be better than for other choices. Finally, the value z has to be taken between 1 and 2. If it is taken equal to 2, then the divisor is too small and we are not able to guarantee convergence. This is seen more clearly in the proofs.

Finally, this entire procedure is applied (at each step) in the same way for (6) using the same exponential bound for the denominators.

2.4. Some remarks. Before finishing the overview of this paper, it is interesting to remark that since the equations with which we deal are not necessarily Hamiltonian, it is possible that in some step of the inductive process, the eigenvalues of the matrix A leave the imaginary axis. In this case, we do not need to worry about resonances from this step onward. Since we cannot know in advance if this is going to happen, we have considered the worst case during all of the proofs; that is, the eigenvalues are always on the imaginary axis. On the other hand, if the initial matrix A is partially elliptic and partially hyperbolic, the results are still valid. In the hyperbolic case, they are of course much better; that is, they hold for a full interval $[0, \varepsilon_0]$.

In some cases, it is possible that at the first step of the inductive process the eigenvalues leave the imaginary axis. (This is really the general case.) Theorem 2.4 ensures that this case can be detected when averaging the original system and looking for the new equilibrium point of this autonomous system. The linearized equations around that point and the "Floquet" matrix (A_∞) of the quasi-periodic orbit differ in $\mathcal{O}(\varepsilon^2)$.

Another interesting point is to compare what we are doing here with the proof of the KAM theorem. In the proof of the KAM theorem (see, for instance, [1]), we use the action variables as parameters to avoid resonances. Here we use the eigenvalues of the matrix A, but since we cannot move them directly, we move them by means of the single parameter ε. Note that the nondegeneracy condition of the KAM (nonzero Jacobian of the frequencies with respect to the actions) says basically that we can control the frequencies through the actions. Here we want to control the eigenvalues by means of ε, so we ask for a suitable Lipschitz condition. As is well known (see, for instance, [2]), the nondegeneracy condition of the KAM theorem can be relaxed to a second-order condition. Here it is possible to do something similar (instead of asking for $\lambda_i^{(1)} \neq 0$ in (10), we can allow $\lambda_i^{(1)} = 0$ but ask for $\lambda_i^{(2)} \neq 0$ or even a higher-order condition), and the estimates on the measure of the Cantorian set of ε are obtained in a similar way. (In fact, the estimates can be even better.) It is also remarkable that the scheme of the proof that we are using is quite similar to that of the KAM theorem [1].

Finally, note that if the nonlinearity h and the independent term g of the initial equation (1) are both equal to zero, we have a Floquet theorem. Now the result obtained is better than that contained in [16]. There it was shown that the measure of the set of "resonant" $\varepsilon \in [0, \varepsilon_0]$ is $o(\varepsilon_0)$, and here it is proved to be exponentially small with ε_0.

2.5. Theorems. From now on, if $x \in \mathbb{R}^n$, we denote by $\|x\|$ the sup norm of x. If A is a matrix, $\|A\|$ denotes the corresponding sup norm.

THEOREM 2.1. *Consider the differential equation*

$$\dot{x} = (A + \varepsilon Q(t, \varepsilon))x + \varepsilon g(t, \varepsilon) + h(x, t, \varepsilon), \tag{12}$$

where $Q(t, \varepsilon)$, $g(t, \varepsilon)$, and $h(x, t, \varepsilon)$ depend on time in a quasi-periodic way with basic frequencies $(\omega_1, \dots, \omega_r)^t$, $r \geq 2$, and $|\varepsilon| < \varepsilon_0$. We assume that A is a constant $d \times d$ matrix with d different eigenvalues λ_i and $\det A \neq 0$. *Let us suppose that $h(x, t, \varepsilon)$ is analytic with respect to x on the ball $B_r(0)$, $h(0, t, \varepsilon) = 0$ and $D_x h(0, t, \varepsilon) = 0$. Moreover, we assume the following:*

1. *Q, g, and h are analytic with respect to t on a strip of width $\rho_0 > 0$, and they depend on ε in a Lipschitz way.*

2. *$\|D_{xx} h(x, t, \varepsilon)\| \leq K$, where $\|x\| \leq r$, $|\varepsilon| \leq \varepsilon_0$, and t belongs to the strip defined in* 1.

3. *The vector $(\lambda_1, \dots, \lambda_d, \sqrt{-1}\omega_1, \dots, \sqrt{-1}\omega_r)$ satisfies the nonresonance conditions*

$$|(k, \omega)\sqrt{-1} - \lambda_i| > \frac{2c}{|k|^{\gamma_0}}, \qquad |(k, \omega)\sqrt{-1} - \lambda_i + \lambda_j| > \frac{2c}{|k|^{\gamma_0}}$$

for all $1 \le i, j \le d$, $k \in \mathbb{Z}^r \setminus \{0\}$, $c > 0$, *and* $\gamma_0 \ge r - 1$. *As usual,* $|k|$ *is taken as* $|k| = |k_1| + \cdots + |k_r|$.

4. *Let us denote by* $\underline{x}(t, \varepsilon)$ *the unique analytical quasi-periodic solution of* $\dot{x} = Ax + \varepsilon g(t, \varepsilon)$ *such that* $\lim_{\varepsilon \to 0} \underline{x}(t, \varepsilon) = 0$ (*the existence of this solution is shown by Lemma* 2.10), *and let us define*

$$\underline{A}(\varepsilon) = A + \varepsilon \overline{Q}(\varepsilon) + \overline{D_x h(\underline{x}(t, \varepsilon), t, \varepsilon)}.$$

Let $\lambda_j^0(\varepsilon)$, $j = 1, \dots, d$, *be the eigenvalues of* $\underline{A}$. *We require the existence of* $\overline{\delta}$, $\delta > 0$ *such that*

$$\frac{\overline{\delta}}{2} |\varepsilon_1 - \varepsilon_2| > |\lambda_i^0(\varepsilon_1) - \lambda_j^0(\varepsilon_1) - (\lambda_i^0(\varepsilon_2) - \lambda_j^0(\varepsilon_2))| > 2\delta |\varepsilon_1 - \varepsilon_2| > 0,$$

$$\frac{\overline{\delta}}{2} |\varepsilon_1 - \varepsilon_2| > |\lambda_k^0(\varepsilon_1) - \lambda_k^0(\varepsilon_2)| > 2\delta |\varepsilon_1 - \varepsilon_2| > 0$$

for all i, j, and k satisfying $1 \le i < j \le d$ *and* $1 \le k \le d$ *and provided that* $|\varepsilon_1|$ *and* $|\varepsilon_2|$ *are less than some small value* ε_0.

Then there exists a Cantorian set $\mathcal{E} \subset (0, \varepsilon_0)$ *with positive Lebesgue measure such that* (12) *can be transformed by means of a change of variables into*

$$\dot{y} = A_\infty(\varepsilon) y + h_\infty(y, t, \varepsilon),$$

where A_∞ *is a constant matrix and* $h_\infty(y, t, \varepsilon)$ *is of second order in* y. *If* ε_0 *is small enough, the relative measure of* $(0, \varepsilon_0) \setminus \mathcal{E}$ *in* $(0, \varepsilon_0)$ *is less than* $\exp(-c_1/\varepsilon_0^{c_2})$ *for* $c_1 > 0$ *and* $c_2 > 0$ (*independent of* ε_0), *where* c_2 *is any number such that* $c_2 < \frac{1}{\gamma_0}$. *Furthermore, the quasi-periodic change of variables that performs this transformation is analytic with respect to t, and it has the same basic frequencies as Q, g, and h.*

Remark 1. In hypothesis 3, we use $2c$ instead of the usual c in the diophantine condition to simplify the notation inside the proofs.

Remark 2. During the proof of this theorem, we will suppose that $\rho_0 \ge 1 + \frac{\pi^2}{6}$. This condition can be achieved by introducing a new time $\tau = st$, where

$$s = \max \left\{ \frac{1 + \frac{\pi^2}{6}}{\rho_0}, 1 \right\}.$$

This scaling may change the constant c; therefore, the set $\mathcal{E}$ is scaled by the same factor.

Remark 3. For fixed values of λ_i, $i = 1, \dots, d$, $\lambda_i \ne \lambda_j$ if $i \ne j$, hypothesis 3 is not satisfied for any c only for a set of values of ω of zero measure if $\gamma_0 > r - 1$.

Remark 4. If $r = 1$, that is, if the perturbation is periodic, no small divisors appear if ε is small enough and the results hold for all $\varepsilon \in (0, \varepsilon_0)$. The proof is classical. We shall assume without explicit mention that $r \ge 2$ in what follows.

COROLLARY 2.2. *Under the hypothesis of Theorem* 2.1, *there exists a Cantorian set* $\mathcal{E} \subset (0, \varepsilon_0)$ *with positive Lebesgue measure* (*and with the complementary being exponentially small*) *such that* (12) *has a quasi-periodic solution* $x_\varepsilon(t)$ *with basic frequencies* $(\omega_1, \dots, \omega_r)$ *such that*

$$\lim_{\substack{\varepsilon \to 0 \\ \varepsilon \in \mathcal{E}}} \|x_\varepsilon\| = 0.$$

COROLLARY 2.3 (a Floquet theorem). *Consider the linear differential equation*

$$\dot{x} = (A + \varepsilon Q(t,\varepsilon))x, \tag{13}$$

where $Q(t,\varepsilon)$ depends quasi-periodically on time with basic frequencies $(\omega_1, \dots, \omega_r)^t$, $r \geq 2$, and $|\varepsilon| < \varepsilon_0$. We assume that A is a constant $d \times d$ matrix with d different eigenvalues λ_i and $\det A \neq 0$. Moreover, we assume the following:

1. *Q is analytic with respect to t on a strip of width $\rho_0 > 0$ and depends on ε in a Lipschitz way.*

2. *The vector $(\lambda_1, \dots, \lambda_d, \sqrt{-1}\omega_1, \dots, \sqrt{-1}\omega_r)$ satisfies the nonresonance condition*

$$|(k,\omega)\sqrt{-1} - \lambda_i + \lambda_j| > \frac{2c}{|k|^{\gamma_0}}$$

for all $1 \leq i,\ j \leq d$, $k \in \mathbb{Z}^r \setminus \{0\}$, $c > 0$, and $\gamma_0 \geq r - 1$.

3. *Let us define*

$$\underline{A}(\varepsilon) = A + \varepsilon\overline{Q}(\varepsilon).$$

Let $\lambda_j^0(\varepsilon)$, $j = 1, \dots, d$, be the eigenvalues of $\underline{A}$. We require the existence of $\overline{\delta}$, $\delta > 0$ such that

$$\frac{\overline{\delta}}{2}|\varepsilon_1 - \varepsilon_2| > |\lambda_i^0(\varepsilon_1) - \lambda_j^0(\varepsilon_1) - (\lambda_i^0(\varepsilon_2) - \lambda_j^0(\varepsilon_2))| > 2\delta|\varepsilon_1 - \varepsilon_2| > 0$$

for all i, j, and k satisfying $1 \leq i < j \leq d$ and $1 \leq k \leq d$ and provided that $|\varepsilon_1|$ and $|\varepsilon_2|$ are less than some small value ε_0.

Then there exists a Cantorian set $\mathcal{E} \subset (0, \varepsilon_0)$ with positive Lebesgue measure such that (13) *can be reduced to a system with constant coefficients*

$$\dot{y} = A_\infty(\varepsilon)y$$

by means of a change of variables $x = (I + \varepsilon P(t,\varepsilon))y$, where I is the identity matrix and P is analytic and quasi-periodic with respect to t with ω as a vector of basic frequencies. If ε_0 is small enough, the relative measure of $(0,\varepsilon_0) \setminus \mathcal{E}$ in $(0,\varepsilon_0)$ is less than $\exp(\frac{-c_1}{\varepsilon_0^{c_2}})$ for $c_1 > 0$ and $c_2 > 0$ (independent of ε_0), where c_2 is any number such that $c_2 < \frac{1}{\gamma_0}$.

Remark. This corollary is the result of taking $g \equiv h \equiv 0$ in Theorem 2.1. We have also weakened the nonresonance condition. This fact becomes clear by looking into the proof for that theorem.

THEOREM 2.4. *Let us consider* (12), *and let us assume that all the hypotheses of Theorem* 2.1 *hold. Moreover, let us assume that the nonlinear part $h(x,t,\varepsilon)$ is of class C^2 with respect to ε and $h(x,t,0) \equiv h(x)$. Then if ε is sufficiently small, the averaged system*

$$\dot{y} = (A + \varepsilon\overline{Q})y + \varepsilon\overline{g} + \overline{h}(y,\varepsilon)$$

has an equilibrium point $x_0(\varepsilon)$ such that

1. $\lim_{\varepsilon\to 0} \|x_0(\varepsilon)\| = 0$;

2. *the matrix A_{x_0} of the linearized system around $x_0(\varepsilon)$ and the matrix A_∞ obtained in Theorem* 2.1 *satisfy $\|A_{x_0} - A_\infty\| = \mathcal{O}(\varepsilon^2)$.*

COROLLARY 2.5. *Let us define $\lambda_i^{x_0}$, $1 \le i \le d$, as the eigenvalues of the matrix A_{x_0} defined in Theorem 2.4. Then, under the hypothesis of Theorem 2.4, an equivalent version of hypothesis 4 in Theorem 2.1 is obtained if λ_i^0 are replaced by $\lambda_i^{x_0}$.*

The proofs of the results above have been split into several parts to simplify the reading. Section 2.6 contains lemmas needed to show the convergence of the iterative scheme used to obtain Theorem 2.1. Section 2.7 presents the convergence proof. Up to this point, we do not worry about the measure of the set of values of ε that must be taken out. Section 2.8 includes the lemmas used to prove that matrix A depends on ε in a Lipschitz way at each step of the procedure. The lemma used to bound the measure of the Cantorian set where Theorem 2.1 holds is given in §2.9. Section 2.10 actually states the bounds for that measure, and, finally, §2.11 is devoted to Theorem 2.4.

2.6. Convergence lemmas. In what follows, we will use the fact that an analytic quasi-periodic function $f(t)$ on a strip of width ρ with $\omega = (\omega_1, \dots, \omega_r)$ as a vector of basic frequencies has Fourier coefficients defined by

$$f_k = \frac{1}{(2\pi)^r} \int_{\mathbb{T}^r} F(\theta_1, \dots, \theta_r) e^{-(k,\theta)\sqrt{-1}}\, d\theta$$

(F is defined in Definition 1.1) such that f can be expanded as

$$f(t) = \sum_{k \in \mathbb{Z}^r} f_k e^{(k,\omega)\sqrt{-1}t}$$

for all t such that $|\mathrm{Im}\ t| < \frac{\rho}{\|\omega\|}$. Moreover, the analyticity of f implies that

$$|f_k| \le \|f\|_\rho e^{-\rho|k|}.$$

LEMMA 2.6. *Let $\delta \in]0, 1]$, $\alpha \ge 1$. Let us define the function*

$$\chi(s) = \left(\frac{s-1}{e}\right)^{s-1} \sqrt{s-1}.$$

Then

$$\sum_{k \in \mathbb{Z}^r} |k|^\alpha e^{-\delta|k|} \le \frac{20r}{3\delta^{r+\alpha}} \chi(r+\alpha).$$

Proof. We shall use the fact that $\#\{k \in \mathbb{Z}^r \ / \ |k| = m\} \le 2rm^{r-1}$. This is checked immediately for $m = 1$ or for $r \le 3$. Then we use induction with respect to r for $m \ge 2$. We then obtain

$$\sum_{k \in \mathbb{Z}^r} |k|^\alpha e^{-\delta|k|} \le 2r \sum_{m=0}^{\infty} m^{r+\alpha-1} e^{-\delta m} = (\triangle).$$

Since the unique maximum of $g(x) = x^{r+\alpha-1}e^{-\delta x}$ is reached when $x = \frac{r+\alpha-1}{\delta}$, we can bound the sum above by this maximum plus the integral as

$$(\triangle) \le 2r \left[\left(\frac{r+\alpha-1}{\delta e}\right)^{r+\alpha-1} + \frac{1}{\delta^{r+\alpha}} \Gamma(r+\alpha) \right]$$

$$= \frac{2r}{\delta^{r+\alpha}}\left(\frac{r+\alpha-1}{e}\right)^{r+\alpha-1}\left[\delta+\left(\frac{e}{r+\alpha-1}\right)^{r+\alpha-1}\Gamma(r+\alpha)\right]$$

$$< \frac{2r}{\delta^{r+\alpha}}\left(\frac{r+\alpha-1}{e}\right)^{r+\alpha-1}\frac{10}{3}\sqrt{r+\alpha-1} = \frac{20r}{3}\frac{\chi(r+\alpha)}{\delta^{r+\alpha}}. \qquad \square$$

LEMMA 2.7. *Let $h : U \subset \mathbb{R}^d \longrightarrow \mathbb{R}^d$ be a function of class $\mathcal{C}^2$ on a ball $B_r(0)$ that satisfies $h(0) = 0$, $D_x h(0) = 0$, and $\|D_{xx}h(x)\| \le K$, where $x \in B_r(0)$. Then $\|h(x)\| \le \frac{K}{2}\|x\|^2$ and $\|D_x h(x)\| \le K\|x\|$.*

The proof follows from Taylor's formula.

LEMMA 2.8. *Let M be a diagonal matrix with d different nonzero eigenvalues μ_j, $j = 1,\dots,d$, and $\alpha = \min\{\min_{i,j;\, i\ne j}|\mu_i-\mu_j|, \min_i|\mu_i|\}$. Let N be a matrix such that $(d+1)\|N\| < \alpha$. Let σ_j, $j=1,\dots,d$, be the eigenvalues of $M+N$, and let B a suitable matrix such that $B^{-1}(M+N)B = D = \operatorname{diag}(\sigma_j)$ with condition number $C(B)$. Then the following hold:*

1. $\beta = \min\{\min_{i,j;\, i\ne j}|\sigma_i-\sigma_j|, \min_i|\sigma_i|\} \ge \alpha - 2\|N\|$.
2. $C(B) \le \frac{\alpha+(d-3)\|N\|}{\alpha-(d+1)\|N\|}$. *In particular, if* $\|N\| < \frac{\alpha}{3d-1}$, *then* $C(B) < 2$.

Proof. The proof can be found in [16] or [15]. □

LEMMA 2.9. *Let A_0 be a $d\times d$ matrix such that* $\operatorname{Spec}(A_0) = \{\lambda_1^0,\dots,\lambda_d^0\}$, *$|\lambda_i^0| > 2\mu$, $|\lambda_i^0-\lambda_j^0| > 2\mu$, $i\ne j$, where $\mu > 0$. Let B_0 be a regular matrix such that $B_0^{-1}A_0B_0 = D_0 = \operatorname{diag}(\lambda_1^0,\dots,\lambda_d^0)$. Let us define $\beta_0 = \max\{\|B_0\|, \|B_0^{-1}\|\}$, and let α be a value such that $0 < \alpha < \frac{2\mu}{(3d-1)\beta_0^2}$. Then if A verifies $\|A-A_0\| < \alpha$, the following conditions hold:*

1. $\operatorname{Spec}(A) = \{\lambda_1,\dots,\lambda_d\}$, *and* $|\lambda_i| > \mu$, $|\lambda_i-\lambda_j| > \mu$, $i\ne j$.
2. *There exists a nonsingular matrix B such that $B^{-1}AB = \operatorname{diag}(\lambda_1,\dots,\lambda_d)$, which satisfies $\|B\|\le\beta$ and $\|B^{-1}\|\le\beta$, with $\beta = 2\beta_0$.*

Proof. Let A be a matrix, and we write $A = A_0 + (A-A_0)$. Then $B_0^{-1}AB_0 = D_0 + N$, where $N = B_0^{-1}(A-A_0)B_0$. Here we can apply Lemma 2.8 to obtain 1 if $\|A-A_0\| \le \frac{2\mu}{(d+1)\beta_o^2}$. Note that Lemma 2.8 states that the condition number of the matrix C that diagonalizes D_0+N is less than 2 provided that $\|A-A_0\| < \frac{2\mu}{(3d-1)\beta_0^2}$. In this case, the matrix that diagonalizes A can be obtained by multiplying B_0 by C. Hence its norm can be bounded by $2\beta_0$. □

In the next lemmas, the parameters γ and ν are assumed to be positive.

LEMMA 2.10. *Let us consider the equation $\dot{x} = Ax + \varepsilon g(t)$, where A is a $d\times d$ matrix belonging to the ball $B_\alpha(A_0) \subset \mathcal{L}(\mathbb{R}^d,\mathbb{R}^d)$ with α as given by Lemma 2.9, $g(t) = (g_i(t))_{1\le i\le d}$, and $g_i(t)$ is an analytic quasi-periodic function on a strip of width ρ_1 and is expressed as*

$$g_i(t) = \sum_{k\in\mathbb{Z}^r} g_i^k e^{(k,\omega)\sqrt{-1}t}.$$

Let us assume that $|(k,\omega)\sqrt{-1}-\lambda_i| > \frac{c}{|k|^\gamma}e^{-\nu|k|}$ $\forall\lambda_i \in \operatorname{Spec}(A)$. Let ρ_2 be such that $0 < \rho_2 < \rho_1 - \nu$ and $\delta = \rho_1-\rho_2-\nu \le 1$. Then there exists a unique quasi-periodic solution of $\dot{x} = Ax+\varepsilon g(t)$ that has the same basic frequencies as g and that satisfies

$$\|x\|_{\rho_2} \le \varepsilon\|g\|_{\rho_1}L_1,$$

where $L_1 = 4\beta_0^2[\frac{1}{\mu} + \frac{20r}{3c}\frac{\chi(r+\gamma)}{\delta^{r+\gamma}}]$ and μ and β_0 are defined in Lemma 2.9.

Remark. In this and forthcoming lemmas, we consider A, Q, g, and h depending also on ε (see Theorem 2.1), but, for simplicity, we do not write this explicitly.

Proof. Let B be the matrix found in Lemma 2.9. Making the change of variables $x = By$ and defining $h(t) = B^{-1}g$, the equation becomes

$$\dot{y} = Dy + \varepsilon h(t).$$

Since D is a diagonal matrix, we can handle this equation as d unidimensional equations, which can easily be solved. If $y = (y_i)_{1\leq i \leq d}$ and

$$y_i(t) = \sum_{k\in\mathbb{Z}^r} y_i^k e^{(k,\omega)\sqrt{-1}t},$$

the coefficients must be $y_i^k = \varepsilon \frac{h_i^k}{(k,\omega)\sqrt{-1}-\lambda_i}$, and they can be bounded by

$$|y_i^k| \leq \begin{cases} \varepsilon\frac{\|h\|_{\rho_1}}{\mu} & \text{if } k=0,\\ \varepsilon\|h\|_{\rho_1}\frac{|k|^\gamma}{c}e^{-(\rho_1-\nu)|k|} & \text{if } k\neq 0.\end{cases}$$

Now we need to bound the norm $\|y\|_{\rho_2}$. Let t be a complex value such that $|\text{Im } \omega_i t| \leq \rho_2$ (for all i). Then

$$|y_i(t)| \leq \sum_{k\in\mathbb{Z}^r} |y_i^k|\,|e^{(k,\omega t)\sqrt{-1}}| \leq \varepsilon\frac{\|h\|_{\rho_1}}{\mu} + \sum_{k\neq 0}\varepsilon\|h\|_{\rho_1}\frac{|k|^\gamma}{c}e^{-(\rho_1-\nu)|k|}e^{\rho_2|k|}.$$

Setting $\delta = \rho_1 - \rho_2 - \nu$, we can use Lemma 2.6 to bound the sum above as

$$|y_i(t)| \leq \varepsilon\|h\|_{\rho_1}\left[\frac{1}{\mu} + \frac{20r\chi(r+\gamma)}{3c\delta^{r+\gamma}}\right].$$

Since $\|h\|_{\rho_1} \leq \|B^{-1}\|\|g\|_{\rho_1}$ and $\|x\|_{\rho_2} \leq \|B\|\|y\|_{\rho_2}$, the result follows. □

LEMMA 2.11. *Let us consider the equation $\dot{P} = AP - PA + Q$, where $A \in B_\alpha(A_0)$ and $Q = (q_{ij})$, where $q_{ij}(t)$ are analytic quasi-periodic functions on a strip of width ρ_1 and are expressed as*

$$q_{ij}(t) = \sum_{k\in\mathbb{Z}^r} q_{ij}^k e^{(k,\omega)\sqrt{-1}t}.$$

We also assume that Q has average equal to zero and that $|(k,\omega)\sqrt{-1} - \lambda_i + \lambda_j| > \frac{c}{|k|^\gamma}e^{-\nu|k|}$ $\forall\lambda_i \in \text{Spec}(A)$. Let ρ_2 be such that $0 < \rho_2 < \rho_1 - \nu$ and $\delta = \rho_1 - \rho_2 - \nu \leq 1$. Then there exists a unique quasi-periodic solution of $\dot{P} = AP - PA + Q$ that has the same basic frequencies as Q and that satisfies

$$\|P\|_{\rho_2} \leq \|Q\|_{\rho_1} L_2,$$

where $L_2 = 16\beta_0^2\frac{20r\chi(r+\gamma)}{3c\delta^{r+\gamma}}$ and β_0 is defined in Lemma 2.9.

Proof. Let B be the matrix found in Lemma 2.9. Making the change of variables $P = BSB^{-1}$ and defining $R = B^{-1}QB$, the equation becomes

$$\dot{S} = DS - SD + R,$$

where the matrix R has zero average. Since D is a diagonal matrix, we can handle this equation as d^2 unidimensional equations, which can be solved easily. If $S = (s_{ij})$ and

$$s_{ij}(t) = \sum_{k\in\mathbb{Z}^r\setminus\{0\}} s_{ij}^k e^{(k,\omega)\sqrt{-1}t},$$

the coefficients must be $s_{ij}^k = \frac{r_{ij}^k}{(k,\omega)\sqrt{-1}-\lambda_i+\lambda_j}$, and they can be bounded by

$$|s_{ij}^k| \leq \|r_{ij}\| \frac{|k|^\gamma}{c} e^{-(\rho_1-\nu)|k|}.$$

Now we need to bound the norm $\|S\|_{\rho_2}$. Let t be a complex value such that $|\mathrm{Im}\ \omega_i t| \leq \rho_2$ (for all i). Then

$$|s_{ij}(t)| \leq \sum_{k\in\mathbb{Z}^r} |s_{ij}^k|\ |e^{(k,\omega t)\sqrt{-1}}| \leq \sum_{k\neq 0} \|r_{ij}\|_{\rho_1} \frac{|k|^\gamma}{c} e^{-(\rho_1-\nu)|k|} e^{\rho_2|k|}.$$

Now we can use Lemma 2.6, setting $\delta = \rho_1 - \rho_2 - \nu$, to bound the sum above as

$$|s_{ij}(t)| \leq \frac{\|r_{ij}\|_{\rho_1}}{c} \left[\frac{20r\chi(r+\gamma)}{3\delta^{r+\gamma}}\right].$$

Since $\|P\|_{\rho_2} \leq \|B\|\|S\|_{\rho_2}\|B^{-1}\|$, we can use $\|R\|_{\rho_1} \leq \|B^{-1}\|\|Q\|_{\rho_1}\|B\|$ to obtain the result. □

LEMMA 2.12. *Let us consider $\dot{x} = (A + \varepsilon Q(t))x + \varepsilon g(t) + h(x,t)$, where the time dependence is assumed to be analytic quasi-periodic on a strip of width ρ_1. We also assume that $h(x,t)$ is analytic with respect to x on the ball $B_\tau(0)$ and satisfies $\|D_{xx}h(x,t)\|_{\rho_1} \leq K$ $\forall x \in B_\tau(0)$. Moreover, $A \in B_\alpha(A_0)$ and $|(k,\omega)\sqrt{-1} - \lambda_i| > \frac{c}{|k|^\gamma}e^{-\nu|k|}$ $\forall \lambda_i \in \mathrm{Spec}(A)$. Let ρ_2 be such that $0 < \rho_2 < \rho_1 - \nu$ and $\delta = \rho_1 - \rho_2 - \nu \leq 1$. Then there exists a change of variables $x = y + \underline{x}(t)$ that transforms the initial equation into*

$$\dot{y} = (\overline{A} + \varepsilon\widetilde{Q}_1)y + \varepsilon^2 g_1(t) + h_1(x,t),$$

where $\widetilde{Q}$ has zero average and the following bounds hold.

1. $\|\widetilde{Q}_1\|_{\rho_2} \leq 2\|Q\|_{\rho_1} + 2KL_1\|g\|_{\rho_1}$, *where L_1 was defined in Lemma* 2.10.
2. $\|g_1\|_{\rho_2} \leq KL_1^2\|g\|_{\rho_1}^2/2 + L_1\|Q\|_{\rho_1}\|g\|_{\rho_1}$.
3. $\|\overline{A}\| \leq \|A\| + \varepsilon(\|g\|_{\rho_1}KL_1 + \|Q\|_{\rho_1})$.
4. $\|D_{yy}h_1(y,t)\|_{\rho_2} \leq K$.
5. $\|\underline{x}\|_{\rho_2} \leq \varepsilon\|g\|_{\rho_1}L_1$.

Here $y \in B_{\tau_1}(0)$, $\tau_1 = \tau - \|\underline{x}\|_{\rho_2}$, and ε is small enough.

Proof. Let $\underline{x}$ be such that $\dot{\underline{x}} = A\underline{x} + \varepsilon g$. In Lemma 2.10, we obtained

$$\|\underline{x}\|_{\rho_2} \leq \varepsilon\|g\|_{\rho_1}L_1.$$

Making the change of variables $x = y + \underline{x}(t)$, we get

$$\dot{y} = (A + \varepsilon Q + D_x h(\underline{x}(t),t))y + h(\underline{x}(t),t) + \varepsilon Q\underline{x}(t) + h_1(y,t),$$

where $h_1(y,t) = h(\underline{x}(t) + y,t) - h(\underline{x}(t),t) - D_x h(\underline{x}(t),t)y$. Defining $Q_1 = Q + \frac{1}{\varepsilon}D_x h(\underline{x}(t),t)$ and $g_1 = \frac{1}{\varepsilon^2}h(\underline{x}(t),t) + \frac{1}{\varepsilon}Q\underline{x}(t)$ ($\varepsilon \neq 0$), the equation is then

$$\dot{y} = (A + \varepsilon Q_1(t))y + \varepsilon^2 g_1(t) + h_1(y,t).$$

To finish, the terms of this equation must be bounded. Let us start with Q_1. Using Lemma 2.7, we get

$$\|Q_1\|_{\rho_2} \leq \|Q\|_{\rho_2} + \frac{1}{\varepsilon}K\|\underline{x}\|_{\rho_2} \leq \|Q\|_{\rho_1} + \|g\|_{\rho_1}KL_1.$$

Now let us bound $\|g_1\|_{\rho_2}$, again by means of Lemma 2.7, as

$$\|g_1\|_{\rho_2} \leq \frac{1}{\varepsilon^2}\frac{K}{2}\|\underline{x}\|_{\rho_2}^2 + \frac{1}{\varepsilon}\|Q\|_{\rho_1}\|\underline{x}\|_{\rho_2} \leq \frac{KL_1^2\|g\|_{\rho_1}^2}{2} + \|Q\|_{\rho_1}\|g\|_{\rho_1}L_1.$$

Now it is $D_{yy}h_1(y,t)$'s turn, that is,

$$\|D_{yy}h_1\|_{\rho_2} = \|D_{xx}h(\underline{x}(t)+y,t)\| \leq K.$$

To do this, we must require that $y \in B_{\tau_1}(0)$, where $\tau_1 = \tau - \|\underline{x}\|_{\rho_2}$. ($\varepsilon$ is assumed to be small enough.) Now, using that $Q_1(t) = \overline{Q}_1 + \widetilde{Q}_1(t)$ and defining $\overline{A} = A + \varepsilon\overline{Q}_1$, we obtain

$$\dot{y} = (\overline{A} + \varepsilon\widetilde{Q}_1(t))y + \varepsilon^2 g_1(t) + h_1(y,t).$$

Finally,

$$\|\overline{A}\| \leq \|A\| + \varepsilon\|\overline{Q}_1\|_{\rho_2},$$

and taking into account that $\|\overline{Q}_1\|_{\rho_2} \leq \|Q_1\|_{\rho_2}$ and that $\|\widetilde{Q}_1(t)\|_{\rho_2} \leq 2\|Q_1\|_{\rho_2}$ the proof is finished. □

LEMMA 2.13. *Let us consider $\dot{x} = (A + \varepsilon Q(t))x + \varepsilon^2 g(t) + h(x,t)$, where the time dependence is assumed to be analytic quasi-periodic on a strip of width ρ_1 and Q has zero average. We also assume that $h(x,t)$ is analytic with respect to x on the ball $B_\tau(0)$ and that it satisfies $\|D_{xx}h(x,t)\|_{\rho_1} \leq K$ $\forall x \in B_\tau(0)$. Moreover, $A \in B_\alpha(A_0)$ and $|(k,\omega)\sqrt{-1} - \lambda_i + \lambda_j| > \frac{c}{|k|^\gamma}e^{-\nu|k|}$ $\forall \lambda_i$, $\lambda_j \in$ Spec(A). Let ρ_2 be such that $0 < \rho_2 < \rho_1 - \nu$ and $\delta = \rho_1 - \rho_2 - \nu \leq 1$. Then there exists a change of variables $x = (I + \varepsilon P(t))y$, where I is the identity $d \times d$ matrix and $P(t)$ is analytic quasi-periodic on a strip of width ρ_2, which transforms the initial equation into*

$$\dot{y} = (\overline{A} + \varepsilon^2\widetilde{Q}_1)y + \varepsilon^2 g_1(t) + h_1(y,t),$$

where $\widetilde{Q}_1$ has zero average and the following bounds hold.

1. $\|\widetilde{Q}_1\|_{\rho_2} \leq \frac{2\|P\|_{\rho_2}}{1-\varepsilon\|P\|_{\rho_2}}\|Q\|_{\rho_1}$, *where $\|P\|_{\rho_2} \leq \|Q\|_{\rho_1}L_2$ and L_2 was defined in Lemma* 2.11.
2. $\|g_1\|_{\rho_2} \leq \frac{1}{1-\varepsilon\|P\|_{\rho_2}}\|g\|_{\rho_1}$.
3. $\|D_{yy}h_1\|_{\rho_2} \leq K\frac{(1+\varepsilon\|P\|_{\rho_2})^2}{1-\varepsilon\|P\|_{\rho_2}}$.
4. $\|\overline{A}\| \leq \|A\| + \varepsilon^2\frac{\|P\|_{\rho_2}}{1-\varepsilon\|P\|_{\rho_2}}\|Q\|_{\rho_1}$.

Here $y \in B_{\tau_1}(0)$, $\tau_1 = \frac{\tau}{1+\varepsilon\|P\|_{\rho_2}}$, and ε is small enough.

Proof. Using Lemma 2.11 we can solve $\dot{P} = AP - PA + Q$. The solution that we have found verifies

$$\|P\|_{\rho_2} \leq \|Q\|_{\rho_1}L_2.$$

Now, by means of the change of variables $x = (I+\varepsilon P)y$ and introducing the notation $Q_1 = (I+\varepsilon P)^{-1}QP$, $g_1 = (I+\varepsilon P)^{-1}g$, and $h_1(y,t) = (I+\varepsilon P)^{-1}h((I+\varepsilon P)y,t)$, we obtain the equation

$$\dot{y} = (A + \varepsilon^2 Q_1(t))y + \varepsilon^2 g_1(t) + h_1(y,t).$$

Next, we are going to bound the terms of this equation. For this purpose, we need the bound of $\|P\|_{\rho_2}$ provided by Lemma 2.11 and displayed above, that is,

$$\|Q_1\|_{\rho_2} \le \left(\sum_{i=0}^{\infty} \varepsilon^i \|P\|_{\rho_2}^i\right) \|Q\|_{\rho_2}\|P\|_{\rho_2} \le \frac{\|P\|_{\rho_2}}{1-\varepsilon\|P\|_{\rho_2}}\|Q\|_{\rho_1},$$

$$\|g_1\|_{\rho_2} \le \frac{1}{1-\varepsilon\|P\|_{\rho_2}}\|g\|_{\rho_1},$$

$$\|D_{yy}h_1\|_{\rho_2} \le \frac{1}{1-\varepsilon\|P\|_{\rho_2}}K\|I+\varepsilon P\|_{\rho_2}^2 \le K\frac{(1+\varepsilon\|P\|_{\rho_2})^2}{1-\varepsilon\|P\|_{\rho_2}}.$$

Of course, we require $y \in B_{r_1}(0)$, where $r_1 = r/(1+\varepsilon\|P\|_{\rho_2})$, and ε is small enough. To finish this, we rewrite the equation using $Q_1(t) = \overline{Q}_1 + \widetilde{Q}_1(t)$, and $\overline{A} = A + \varepsilon^2\overline{Q}_1$, and we obtain

$$\dot{y} = (\overline{A} + \varepsilon^2\widetilde{Q}_1)y + \varepsilon^2 g_1(t) + h_1(y,t),$$

and we only need to bound $\overline{A}$ as

$$\|\overline{A}\| \le \|A\| + \varepsilon^2\|Q_1\|_{\rho_2}. \qquad \square$$

Thus far, we have the main tools to carry out one step of the inductive process. Now we present a lemma that will be used to show the convergence.

LEMMA 2.14. *Let η_n be a sequence of real positive numbers such that*

$$\eta_{n+1} \le (\overline{\gamma}z^n)^{\overline{\gamma}z^n}\eta_n^2$$

for all $n \ge 0$, where $\overline{\gamma} > 0$, $1 < z < 2$. Then

$$\eta_n \le \left[\left(\overline{\gamma}z^{\frac{z}{2-z}}\right)^{\frac{\overline{\gamma}}{2-z}}\eta_0\right]^{2^n}.$$

Proof. Taking logarithms, we have

$$\begin{aligned}\log\eta_{n+1} &\le (\overline{\gamma}z^n)\log(\overline{\gamma}z^n) + 2\log\eta_n\\ &\le (\overline{\gamma}z^n)\log(\overline{\gamma}z^n) + 2\overline{\gamma}z^{n-1}\log(\overline{\gamma}z^{n-1}) + 4\log\eta_{n-1} \le \cdots\\ &\le \overline{\gamma}\sum_{j=0}^{n} 2^j z^{n-j}(\log\overline{\gamma} + (n-j)\log z) + 2^{n+1}\log\eta_0\\ &= \overline{\gamma}\,2^{n+1}\log\overline{\gamma}\sum_{l=0}^{n}\frac{z^l}{2^{l+1}} + \overline{\gamma}\,2^{n+1}\log z\sum_{l=1}^{n} l\frac{z^l}{2^{l+1}} + 2^{n+1}\log\eta_0\\ &\le \overline{\gamma}\,2^{n+1}\frac{1}{2-z}\log\overline{\gamma} + \overline{\gamma}\,2^{n+1}\frac{z}{(2-z)^2}\log z + 2^{n+1}\log\eta_0.\end{aligned}$$

The result follows by exponentiation. $\square$

LEMMA 2.15. *Let $\{a_n\}_n$ be a sequence of positive real numbers that satisfies $a_n \in\ ]0,1]$, $\prod_{n=0}^{\infty} a_n = a \in\ ]0,1]$. Let $\{b_n\}_n$ be another sequence of positive real numbers that satisfies $\sum_{n=0}^{\infty} b_n = b < +\infty$. Consider the new sequence $\{r_n\}_n$ defined* *by $\tau_{n+1} = a_n\tau_n - b_n$. Then the sequence $\{\tau_n\}_n$ converges to a limit value τ_∞ that satisfies $\tau_\infty \geq a\tau_0 - b$.*

Proof. It is easy to see that

$$\tau_{n+1} = \left(\prod_{i=0}^{n} a_i\right)\tau_0 - \sum_{i=0}^{n-1}\left[\left(\prod_{j=i+1}^{n} a_j\right) b_i\right] - b_n.$$

As all the terms appearing in this expression converge, so does τ_n. Moreover, using that

$$\prod_{j=i+1}^{n} a_j \leq 1$$

for all n, the result follows. □

2.7. Proof of Theorem 2.1 (part I). Here we will present the proof without worrying about resonances, and then in §2.10, we will take out the values of ε for which the proof fails.

First of all, let us denote by A_0 the initial matrix A (see Theorem 2.1) corresponding to the averaged linear part of the differential system. Let μ be a real value such that if $\mathrm{Spec}(A_0) = \{\lambda_1^0, \dots, \lambda_d^0\}$, then $|\lambda_i^0| > 2\mu$ and $|\lambda_i^0 - \lambda_j^0| > 2\mu$ for all $i \neq j$. Then Lemma 2.9 can be applied to obtain values α and β such that all the matrices contained inside the ball $B_\alpha(A_0) = \{\frac{A}{\|A-A_0\|} < \alpha\}$ can be diagonalized. Moreover, the matrix B of the diagonalizing change of variables satisfies $\|B\| < \beta$ and $\|B^{-1}\| < \beta$. During the proof, we shall see that if ε is small enough, all the matrices A_n that appear during the inductive process are inside that ball.

Since we assume that the dependence of Q, g, and h with respect to ε is Lipschitz, every time we compute some norm, we mean without explicit mention that we look for the maximum not only with respect to t in the suitable strip but also with respect to ε in the allowed range.

To begin the proof, we suppose that we have applied the method presented previously up to step n, and we will see that we can apply it again to get the $n+1$ step. In this way, we shall obtain bounds for the quasi-periodic part at the nth step and for the transformation at this step, and this allows us to prove the convergence.

We note that in the first step (that is, when the current data are the initial ones) the index n is equal to 0.

Now suppose that we are at the nth step. This means that the equation we have is

$$\dot{x}_n = (A_n(\varepsilon) + \varepsilon^{2^n} Q_n(t,\varepsilon))x_n + \varepsilon^{2^n} g_n(t,\varepsilon) + h_n(x_n,t,\varepsilon), \tag{14}$$

where A_n belongs to $B_\alpha(A_0)$; its eigenvalues λ_i verify the nonresonance condition

$$|(k,\omega)\sqrt{-1} - \lambda_i| > \frac{c}{|k|^{\gamma_n}} e^{-\nu_n|k|},$$

where $\gamma_n = \gamma_0 z^n$ $(1 < z < 2)$ and $\nu_n = \frac{\nu_0}{(n+1)^2}$, with $0 < \nu_0 < \frac{1}{4}$. Since we need to reduce the width of the analyticity strip of the quasi-periodic functions, we define $\rho_{n+1} = \rho_n - \frac{1}{(n+1)^2}$ and $\sigma_n = \rho_n - \frac{1}{2(n+1)^2}$, with $\rho_0 = 1 + \frac{\pi^2}{6}$. During the proof, we shall see that the analyticity ball (with respect to x) of $h_n(x,t)$ must be reduced at each step of the inductive process; and we shall find that by selecting ε small enough, the limit radius of this ball is positive. Let us define τ_n as this radius at step n. Now we can apply Lemma 2.12 to transform (14) into

$$\dot{y}_n = (\widehat{A}_n(\varepsilon) + \varepsilon^{2^n}\widehat{Q}_n(t,\varepsilon))y_n + \varepsilon^{2^{n+1}}\widehat{g}_n(t,\varepsilon) + \widehat{h}_n(y_n,t,\varepsilon), \tag{15}$$

where the width of the analyticity strip has been reduced to σ_n. Now, assuming that the nonresonance condition

$$|(k,\omega)\sqrt{-1} - \lambda_i + \lambda_j| > \frac{c}{|k|^{\gamma_n}}e^{-\nu_n|k|}$$

holds for all λ_i, $\lambda_j \in \operatorname{Spec}(\widehat{A}_n(\varepsilon))$, we can apply Lemma 2.13 to (15) and get

$$\dot{x}_{n+1} = (A_{n+1}(\varepsilon) + \varepsilon^{2^{n+1}}Q_{n+1}(t,\varepsilon))x_{n+1} + \varepsilon^{2^{n+1}}g_{n+1}(t,\varepsilon) + h_{n+1}(x_{n+1},t,\varepsilon). \tag{16}$$

Now the width of the analyticity strip has been reduced to ρ_{n+1}. The next step of the proof is to obtain bounds of the terms appearing in (16) depending on the bounds of the terms of (14).

In what follows, $L_{1,n}$ and $L_{2,n}$ denote the values of L_1 and L_2 as introduced in Lemmas 2.10 and 2.11, where γ, ν, and δ are replaced by γ_n, ν_n, and $\frac{1/2-\nu_0}{(n+1)^2}$, respectively.

Using Lemma 2.13 and the condition $\varepsilon^{2^n}\|P_n\|_{\rho_{n+1}} \leq \frac{1}{2}$ (see below), we get

$$\|Q_{n+1}\|_{\rho_{n+1}} \leq 4L_{2,n}\|\widehat{Q}_n\|^2_{\sigma_n}.$$

Here we need Lemma 2.12 to bound the expression above, but the bound provided by this lemma has a "still unknown" term, that is, the bound of the second derivative of h_n. Let us call this value K_n. Note that it is "modified" at each step by Lemma 2.13. In order to bound it, we shall assume that ε is small enough to ensure that $\varepsilon^{2^n}\|P_n\|_{\rho_{n+1}}$ is less than $\frac{1}{2}$. This implies that the value of ε will be reduced at each step, if necessary, to guarantee that condition. We will see that this condition is achieved from a certain step onward, without modifying ε anymore. Therefore, we assume that $K_n \leq (\frac{9}{2})^n K_0$ (when the convergence is proved, we shall give a more realistic bound of K_n that converges to a real number), and Lemma 2.12 states that

$$\|\widehat{Q}_n\|_{\sigma_n} \leq 2\|Q_n\|_{\rho_n} + 2K_nL_{1,n}\|g_n\|_{\rho_n}.$$

Now we bound the norm of g_{n+1} as

$$\|g_{n+1}\|_{\rho_{n+1}} \leq 2\|\widehat{g}_n\|_{\sigma_n},$$

and, from Lemma 2.12,

$$\|g_{n+1}\|_{\rho_{n+1}} \leq K_nL^2_{1,n}\|g_n\|^2_{\rho_n} + 2L_{1,n}\|Q_n\|_{\rho_n}\|g_n\|_{\rho_n}.$$

For simplicity, let us denote $\alpha_n = \|Q_n\|_{\rho_n}$ and $\beta_n = \|g_n\|_{\rho_n}$. This means that we have obtained the bounds

$$\alpha_{n+1} \leq 16L_{2,n}\left(\alpha_n + \left(\frac{9}{2}\right)^n L_{1,n}\beta_n\right)^2,$$

$$\beta_{n+1} \leq \left(\frac{9}{2}\right)^n L^2_{1,n}\beta^2_n + 2L_{1,n}\alpha_n\beta_n.$$

To bound α_n and β_n, we define, $\eta_n = \max\{\alpha_n, \beta_n\}$. Since $L_{2,n} < 4L_{1,n}$, after some rearranging, we get

$$\alpha_{n+1} \leq 64 L_{1,n} \left(1 + \left(\frac{9}{2}\right)^n L_{1,n}\right)^2 \eta_n^2,$$
$$\beta_{n+1} \leq \left(\left(\frac{9}{2}\right)^n L_{1,n}^2 + 2L_{1,n}\right) \eta_n^2.$$

Since we can assume $c \leq 1$ without adding any additional constraint on the small divisors, we have $L_{1,n} > 1$. Hence

$$\eta_{n+1} < 128 \left(\frac{9}{2}\right)^n L_{1,n}^3 \eta_n^2.$$

It is immediate to check that there exists $\overline{\gamma}$ (depending on γ_0, r, β_0, c, ν_0, and z) such that

$$128 \left(\frac{9}{2}\right)^n L_{1,n}^3 < (\overline{\gamma} z^n)^{\overline{\gamma} z^n} \quad \forall n \geq 0.$$

Using Lemma 2.14, we have $\eta_n \leq M_1^{2^n}$, where $M_1 = (\overline{\gamma} z^{\frac{z}{2-z}})^{\frac{\overline{\gamma}}{2-z}} \eta_0$. With this, we have proved that

$$\|Q_n\|_{\rho_n} \leq M_1^{2^n}, \qquad \|g_n\|_{\rho_n} \leq M_1^{2^n}.$$

Note that this bound allows us to ensure that if $\varepsilon < \varepsilon_1 = M_1^{-1}$, then

$$\lim_{n \to \infty} \varepsilon^{2^n} \|Q_n\|_{\rho_n} = \lim_{n \to \infty} \varepsilon^{2^n} \|g_n\|_{\rho_n} = 0.$$

The next step is to bound $\|P_n\|_{\rho_{n+1}}$. For this purpose, we first use Lemma 2.13 and then Lemma 2.12 to obtain

$$\|P_n\|_{\rho_{n+1}} \leq 2L_{2,n}(\|Q_n\|_{\rho_n} + K_n L_{1,n} \|g_n\|_{\rho_n}) \leq 4 \left(\frac{9}{2}\right)^n L_{1,n} L_{2,n} \eta_n.$$

Now it is not difficult to prove that $L_{1,n} \leq M_2^{2^n}$ and $L_{2,n} \leq M_2^{2^n}$ for a suitable constant M_2. (This is easily shown by taking logarithms.) Hence we can derive

$$\|P_n\|_{\rho_{n+1}} \leq M_3^{2^n}$$

for a suitable constant M_3. This means that if $\varepsilon < \varepsilon_1 = \min\{M_1^{-1}, M_3^{-1}\}$, we have

$$\lim_{n \to \infty} \varepsilon^{2^n} \|P_n\|_{\rho_{n+1}} = 0.$$

This allows the condition $\varepsilon^{2^n} \|P_n\|_{\rho_{n+1}} < \frac{1}{2}$ without reducing the value of ε at each step. Now we will bound $\|\underline{x}_n\|_{\sigma_n}$ as

$$\|\underline{x}_n\|_{\sigma_n} \leq \varepsilon^{2^n} L_{1,n} \|g_n\|_{\rho_n} < \varepsilon^{2^n} M_4^{2^n}$$

for a suitable M_4. When the changes of coordinates have been bounded, we can estimate the decrease of the radius τ_n of the ball where h_n is analytic with respect to x. It has been shown that

$$\tau_{n+1} = \frac{\widehat{\tau}_n}{1 + \varepsilon^{2^n} \|P_n\|_{\rho_{n+1}}} = \frac{1}{1 + \varepsilon^{2^n} \|P_n\|_{\rho_{n+1}}} \tau_n - \frac{\|\underline{x}_n\|_{\rho_n}}{1 + \varepsilon^{2^n} \|P_n\|_{\rho_{n+1}}}.$$

Now we define

$$a_n = \frac{1}{1+\varepsilon^{2^n}\|P_n\|_{\rho_{n+1}}}, \qquad b_n = \frac{\|\underline{x}_n\|_{\rho_n}}{1+\varepsilon^{2^n}\|P_n\|_{\rho_{n+1}}}.$$

It is easy to prove that $\prod_{n=0}^{\infty} a_n$ converges by

$$\left|\ln \prod_{n=0}^{N} a_n\right| \le \sum_{n=0}^{N} |\ln a_n| < \sum_{n=0}^{N} \varepsilon^{2^n}\|P_n\|_{\rho_{n+1}} < \infty \quad \forall N \in \mathbb{N}.$$

Because $\sum b_n$ is also convergent, we can apply Lemma 2.15 to get $r_\infty \ge a r_0 - b$, which is positive if ε is taken small enough.

Now let us bound $\|A_n\|$ as

$$\begin{aligned}\|A_{n+1}\| &\le \|\widehat{A}_n\| + \varepsilon^{2^n}\frac{\|P_n\|_{\rho_{n+1}}}{1-\varepsilon^{2^n}\|P_n\|_{\rho_{n+1}}} \\ &\le \|\widehat{A}_n\| + \varepsilon^{2^n}\|Q_n\|_{\rho_n} + \varepsilon^{2^n}\frac{\|P_n\|_{\rho_{n+1}}}{1-\varepsilon^{2^n}\|P_n\|_{\rho_{n+1}}}.\end{aligned}$$

Using the bounds found above we can write that

$$\|A_{n+1}\| \le \|A_n\| + \kappa_n,$$

where $\kappa_n \le \varepsilon^{2^n} M_5^{2^n}$ for a suitable M_5. Because $\sum \kappa_n$ is convergent we can ensure that if ε is selected small enough, the matrices A_n are always inside the ball $B_\alpha(A_0)$ defined before.

Now consider the value K_n. Above we have used the pessimistic bound $K_n \le (\frac{9}{2})^n K_0$. Note that this bound does not allow us to guarantee the convergence of the functions $h_n(x_n, t)$ to an analytic function $h_\infty(x_\infty, t)$ with respect to x. Now we can use a more accurate bound of that value to get this. From Lemma 2.13, we know that

$$K_{n+1} \le K_n \frac{(1+\varepsilon^{2^n}\|P_n\|_{\rho_{n+1}})^2}{1-\varepsilon^{2^n}\|P_n\|_{\rho_{n+1}}};$$

by means of the inequality $\frac{1}{1-x} \le 1 + 2x$ if $0 \le x \le \frac{1}{2}$, we get

$$K_{n+1} \le \left(1 + 2\varepsilon^{2^n}\|P_n\|_{\rho_{n+1}}\right)^3 K_n.$$

And, using the bounds of $\|P_n\|_{\rho_{n+1}}$ that we already know, it is easy to see that the (bound of the) value K_n converges.

Hence we have obtained the convergence proof for all $|\varepsilon| < \varepsilon_0$ for a suitable ε_0 without taking into account the "bad set" of values of ε for which the diophantine conditions at some step n are not satisfied.

2.8. Lipschitz lemmas. In this section, we present the lemmas needed to show that at each step of the inductive process, the dependence on ε of the eigenvalues of the matrix A_n is Lipschitz.

LEMMA 2.16. *Let $f : [-\varepsilon, \varepsilon] \to \mathbb{C}$ be a Lipschitz function from above (with constant L) and from below (with constant l), that is,*

$$|f(x) - f(y)| \le L|x-y|, \qquad |f(x) - f(y)| \ge l|x-y|.$$

Let $g : [-\varepsilon, \varepsilon] \to \mathbb{C}$ be another Lipschitz function from above with constant $\alpha < l$, that is,

$$|g(x) - g(y)| \leq \alpha|x - y|.$$

Then $h = f + g$ is Lipschitz from above with constant $L + \alpha$ and from below with constant $l - \alpha$, that is,

$$|h(x) - h(y)| \leq (L + \alpha)|x - y|, \qquad |h(x) - h(y)| \geq (l - \alpha)|x - y|.$$

Proof. The proof is elementary. □

Remark. Henceforth, all Lipschitz functions appearing in the text will be Lipschitz from above unless otherwise stated. Moreover, we will sometimes use $\mathcal{L}(f)$ to denote the Lipschitz constant (always with respect to ε) of a Lipschitz function f. The set on which this constant is taken should be clear from the context. For instance, if $f(t, \varepsilon)$ is known to be defined for $|\text{Im } t| \leq \rho$ and $\varepsilon \in E \subset \mathbb{R}$ and is Lipschitz with respect to ε in E, then $|f(t, \varepsilon_2) - f(t, \varepsilon_1)| \leq \mathcal{L}(f)|\varepsilon_2 - \varepsilon_1|$ for all t, ε_1, and ε_2 in the allowed domain.

In what follows, we shall denote by $\overline{\mathbb{N}}$ the set of nonnegative integers, that is, $\mathbb{N} \cup \{0\}$.

LEMMA 2.17. *Let us define*

$$f(z, \varepsilon) = \sum_{|k| \geq 2} a_k(\varepsilon) z^k, \quad k \in \overline{\mathbb{N}}^d,$$

and assume that the sum is convergent $\forall z \in D = D_1 \times \cdots \times D_d \subset \mathbb{C}^d$, where D_j are fixed disks of $\mathbb{C}$. Moreover, we suppose that f depends on ε in a Lipschitz way with Lipschitz constant L. Let us take $\widehat{D} \subset D$ such that $\widehat{D} = \widehat{D}_1 \times \cdots \times \widehat{D}_d$ and satisfying $\text{radius}(\widehat{D}_j) \leq \alpha \, \text{radius}(D_j) = \alpha r_j$, $0 < \alpha < 1$. *Then if $z \in \widehat{D}$, it holds that*

1. $|f(z, \varepsilon_1) - f(z, \varepsilon_2)| \leq K_2(\alpha) L |\varepsilon_1 - \varepsilon_2| \alpha^2$,
2. $\|D_z f(z, \varepsilon_1) - D_z f(z, \varepsilon_2)\| \leq K_1(\alpha) L |\varepsilon_1 - \varepsilon_2| \alpha$,

where both $K_i(\alpha)$, $i = 1, 2$, defined for $\alpha < 1$, are continuous and increasing functions.

Proof Let $\partial_0 D$ be $\partial D_1 \times \cdots \times \partial D_d$, where ∂ stands for the boundary of the corresponding sets. Since

$$a_k(\varepsilon) = \frac{1}{(2\pi\sqrt{-1})^d} \int_{\partial_0 D} \frac{f(z, \varepsilon)}{z_1^{k_1+1} \cdots z_d^{k_d+1}} \, dz_1 \cdots dz_d,$$

we have that

$$\begin{aligned} |a_k(\varepsilon_1) - a_k(\varepsilon_2)| &\leq \frac{1}{(2\pi)^d} \int_{\partial_0 D} \frac{|f(z, \varepsilon_1) - f(z, \varepsilon_2)|}{|z_1|^{k_1+1} \cdots |z_d|^{k_d+1}} |dz_1 \cdots dz_d| \\ &\leq \frac{L|\varepsilon_1 - \varepsilon_2|}{(2\pi)^d} \int_{\partial_0 D} \frac{|dz_1| \cdots |dz_d|}{r_1^{k_1+1} \cdots r_d^{k_d+1}} = \frac{L}{r^k} |\varepsilon_1 - \varepsilon_2|. \end{aligned}$$

On the other hand,

$$\begin{aligned} |f(z, \varepsilon_1) - f(z, \varepsilon_2)| &\leq \sum_{|k| \geq 2} |a_k(\varepsilon_1) - a_k(\varepsilon_2)| |z^k| \\ &\leq L|\varepsilon_1 - \varepsilon_2| \sum_{|k| \geq 2} \frac{|z_1|^{k_1} \cdots |z_d|^{k_d}}{r_1^{k_1} \cdots r_d^{k_d}} = (\diamond). \end{aligned}$$

Now, using $z \in \widehat{D}$ (that is, $|z_j| \leq \alpha r_j$) and $\#\{k \in \overline{\mathbb{N}}^d \;/\; |k| = m\} \leq dm^{d-1}$ if $m \geq 1$ (which can be obtained by induction with respect to d), we obtain

$$(\diamond) \leq L|\varepsilon_1 - \varepsilon_2| \sum_{|k|\geq 2} \alpha^{|k|} \leq L|\varepsilon_1 - \varepsilon_2| \sum_{m=2}^{\infty} dm^{d-1}\alpha^m = K_2(\alpha)L|\varepsilon_1 - \varepsilon_2|\alpha^2,$$

where $K_2(\alpha) = d\sum_{m=2}^{\infty} m^{d-1}\alpha^{m-2}$. Finally, it is not difficult to see that $K_2(\alpha)$ is convergent if $|\alpha| < 1$. This completes the proof of 1.

Since

$$\frac{\partial f}{\partial z_j}(z, \varepsilon) = \sum_{|k|\geq 2} k_j a_k(\varepsilon) z^{k-e_j},$$

we can proceed in the same way as before, that is,

$$\begin{aligned}
\left| \frac{\partial f}{\partial z_j}(z, \varepsilon_1) - \frac{\partial f}{\partial z_j}(z, \varepsilon_2) \right| &\leq \sum_{|k|\geq 2} k_j |a_k(\varepsilon_1) - a_k(\varepsilon_2)||z|^{k-e_j} \\
&\leq \sum_{|k|\geq 2} k_j \frac{L}{r^k} |\varepsilon_1 - \varepsilon_2||z|^{k-e_j} \leq \frac{L|\varepsilon_1 - \varepsilon_2|}{r_j} \sum_{|k|\geq 2} k_j \alpha^{k-e_j} \\
&\leq \frac{L|\varepsilon_1 - \varepsilon_2|}{r_j} \sum_{m=2}^{\infty} dm^d \alpha^{m-1} \leq K_1(\alpha)L|\varepsilon_1 - \varepsilon_2|\alpha,
\end{aligned}$$

where $K_1(\alpha) = \frac{d}{r_\infty} \sum_{m=2}^{\infty} m^d \alpha^{m-2}$ and r_∞ is a lower bound of the values r_j (see §2.7). Here, we note that $K_1(\alpha)$ is convergent if $|\alpha| < 1$. To complete the proof, we only need to take the sup norm of the vector of components

$$\frac{\partial f}{\partial z_j}(z, \varepsilon_1) - \frac{\partial f}{\partial z_j}(z, \varepsilon_2). \qquad \square$$

LEMMA 2.18. *Let us suppose that $P(t, \varepsilon)$ is a matrix depending on ε in a Lipschitz way with constant L. If $\|P\| \leq \frac{1}{2}$, then $(I + P(t, \varepsilon))^{-1}$ is Lipschitz with respect to ε with constant $4L$.*

Proof. It is known that

$$(I + P)^{-1} = I - P + P^2 - P^3 + \cdots,$$

and then it is easy to see that

$$\begin{aligned}
\mathcal{L}((I + P)^{-1}) &\leq \mathcal{L}(P) + \mathcal{L}(P^2) + \mathcal{L}(P^3) + \cdots \leq \sum_{n=1}^{\infty} \left[n\|P\|^{n-1}L \right] \\
&= \frac{1}{(1 - \|P\|)^2} L \leq 4L. \qquad \square
\end{aligned}$$

LEMMA 2.19. *Let $q(t, \varepsilon)$ be an analytic quasi-periodic function on a strip of width ρ_1. We write*

$$q(t, \varepsilon) = \sum_{k \in \mathbb{Z}^r} q^k(\varepsilon) e^{(k,\omega)\sqrt{-1}t},$$

and we assume that all the coefficients $q^k(\varepsilon)$ depend on ε in a Lipschitz way with constant L_k. Moreover, we suppose that $L_k \leq L|k|^\alpha e^{-\rho_1|k|}$ if $k \neq 0$, where L is a positive constant. Let us take $\rho_2 \in]0, \rho_1[$. Then if $q(t,\varepsilon)$ is restricted to a strip of width ρ_2, it depends on ε in a Lipschitz way with constant

$$L' = L_0 + L\frac{20r}{3\delta^{r+\alpha}}\chi(r+\alpha),$$

where $L_0 = L_{k=0}$, $\delta = \rho_1 - \rho_2$, and χ is as defined in Lemma 2.6.

Proof.

$$\begin{aligned} |q(t,\varepsilon_1) - q(t,\varepsilon_2)| &\leq \sum_{k\in\mathbb{Z}^r} |q^k(\varepsilon_1) - q^k(\varepsilon_2)||e^{(k,\omega)\sqrt{-1}t}| \\ &\leq \left[L_0 + L\sum_{k\neq 0} |k|^\alpha e^{-\delta|k|} \right] |\varepsilon_1 - \varepsilon_2|. \end{aligned}$$

Here we can apply Lemma 2.6 to obtain the desired result. □

LEMMA 2.20. *Let $q(t,\varepsilon)$ be an analytic quasi-periodic function on a strip of width ρ,*

$$q(t,\varepsilon) = \sum_{k\in\mathbb{Z}^r} q^k(\varepsilon) e^{(k,\omega)\sqrt{-1}t}.$$

Let us assume that $q(t,\varepsilon)$ depends on ε in a Lipschitz way with constant L. Then the coefficients $q^k(\varepsilon)$ depend on ε in a Lipschitz way since

$$|q^k(\varepsilon_1) - q^k(\varepsilon_2)| \leq L_k|\varepsilon_1 - \varepsilon_2|,$$

where $L_k = Le^{-\rho|k|}$.

Proof. Let us fix ε_1 and ε_2 and define $p(t) = q(t,\varepsilon_1) - q(t,\varepsilon_2)$. Since $\|p\|_\rho \leq L|\varepsilon_1 - \varepsilon_2|$, the Fourier coefficients of p satisfy

$$|p^k| \leq L|\varepsilon_1 - \varepsilon_2|e^{-\rho|k|},$$

and using the fact that $|p^k| = |q^k(\varepsilon_1) - q^k(\varepsilon_2)|$, the result follows. □

LEMMA 2.21. *Let us define*

$$f(\varepsilon) = \frac{g(\varepsilon)}{\sigma - \lambda(\varepsilon)},$$

where $|\sigma - \lambda(\varepsilon)| \geq u$ and g and λ are Lipschitz functions with constants L_g and L_λ, respectively. Then f is Lipschitz with constant

$$L_f = \frac{L_g}{u} + \|g\|_\infty \frac{L_\lambda}{u^2}.$$

Proof. The proof is straightforward. □

LEMMA 2.22. *Let A_0 be a $d \times d$ matrix such that* $\operatorname{Spec}(A_0) = \{\lambda_1^0, \ldots, \lambda_d^0\}$, *$|\lambda_i^0| > 2\mu$, $|\lambda_i^0 - \lambda_j^0| > 2\mu$, $i \neq j$, where $\mu > 0$. Let $A(\varepsilon)$ be a matrix-valued function such that $\|A(\varepsilon) - A_0\| < \alpha$ if $|\varepsilon| < \varepsilon_0$ and dependent on ε in a Lipschitz way, with constant L_A.*

Let $B(\varepsilon)$ be the change of variables that diagonalizes $A(\varepsilon)$ (see Lemma 2.9). Then there exist $\tau_1 = \tau_1(A_0, \alpha, \beta)$ and $\tau_2 = \tau_2(A_0, \alpha, \beta)$ such that

$$\begin{aligned} \|B(\varepsilon_1) - B(\varepsilon_2)\| &\le \tau_1 L_A |\varepsilon_1 - \varepsilon_2|, \\ \|B^{-1}(\varepsilon_1) - B^{-1}(\varepsilon_2)\| &\le \tau_1 L_A |\varepsilon_1 - \varepsilon_2|, \\ |\lambda_j(\varepsilon_1) - \lambda_j(\varepsilon_2)| &\le \tau_2 L_A |\varepsilon_1 - \varepsilon_2|, \end{aligned}$$

where $\lambda_j(\varepsilon)$ are the eigenvalues of $A(\varepsilon)$ and the definition of values α and β can be found in Lemma 2.9.

Proof. This result is essentially contained in [19, pp. 66–67], but for an analytic dependence on ε. The result for a Lipschitz dependence on ε can be obtained as follows.

1. Let us consider the matrix A as a function of all its elements a_{ij}. This implies that if the elements are close enough to those of A_0, the eigenvalues and eigenvectors depend on a_{ij} in an analytic way. Hence in any compact set inside the domain of analyticity, they also depend in a Lipschitz way.

2. The elements $a_{ij}(\varepsilon)$ of $A(\varepsilon)$ also depend on ε in Lipschitz way with the same constant since

$$\begin{aligned} |a_{ij}(\varepsilon_1) - a_{ij}(\varepsilon_2)| &\le \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}(\varepsilon_1) - a_{ij}(\varepsilon_2)| \\ &= \|A(\varepsilon_1) - A(\varepsilon_2)\| \le L|\varepsilon_1 - \varepsilon_2|. \end{aligned}$$

3. Finally, we compose the Lipschitz dependence (of the eigenvalues and eigenvectors) on a_{ij} with the Lipschitz dependence of a_{ij} on ε. □

LEMMA 2.23. *Let us consider the equation*

$$\dot{x} = A(\varepsilon)x + g(t, \varepsilon)$$

under the same hypothesis as in Lemma 2.10. Let ρ_2 be such that $0 < \rho_2 < \rho_1 - 2\nu$ and $\delta = \rho_1 - \rho_2 - 2\nu \le 1$. Moreover, we assume that $A(\varepsilon)$ and $g(t, \varepsilon)$ depend on ε in a Lipschitz way with constants L_A and L_g, respectively. Then the solution $x(t, \varepsilon)$ of this equation (see Lemma 2.10) depends on ε in a Lipschitz way for t belonging to the strip of width ρ_2 with constant

$$L_x \le \frac{\chi(r + 2\gamma)}{\delta^{r+2\gamma}} (E_1 L_A \|g\|_{\rho_1} + E_2 L_g),$$

where E_1 and E_2 are positive constants that do not depend on the actual step of the inductive process of §2.7.

Proof. First of all, let us make the change of variables $x = B(\varepsilon)y$ (the matrix $B(\varepsilon)$ is given by Lemma 2.22) in order to diagonalize the matrix $A(\varepsilon)$. With this, the equation becomes

$$\dot{y} = D(\varepsilon)y + h(t, \varepsilon),$$

where $D(\varepsilon)$ is a diagonal matrix and $h(t, \varepsilon) = B^{-1}(\varepsilon)g(t, \varepsilon)$. Lemma 2.22 ensures that $L_D \equiv \mathcal{L}(D) \equiv \max_i L_{\lambda_i} = \tau_2 L_A$ and $L_h \equiv \mathcal{L}(h) = \tau_1 L_A \|g\|_{\rho_1} + \beta L_g$. Moreover, since

$$h(t, \varepsilon) = \sum_{k \in \mathbb{Z}^r} h^k(\varepsilon) e^{(k,\omega)\sqrt{-1}t},$$

we have by Lemma 2.20 that $L_{h^k} \equiv \mathcal{L}(h^k) = L_h e^{-\rho_1|k|}$.

As shown in Lemma 2.10, the solution in which we are interested is given by

$$y_i^k(\varepsilon) = \frac{h_i^k(\varepsilon)}{(k,\omega)\sqrt{-1} - \lambda_i(\varepsilon)}.$$

Now let us compute $L_{y_i^k} \equiv \mathcal{L}(y_i^k)$. We distinguish two cases and use Lemma 2.21 in both.

Case 1: $k = 0$.

$$\begin{aligned} L_{y_i^0} &= \frac{L_{h^0}}{\mu} + |h^0|\frac{L_{\lambda_i}}{\mu^2} \le \frac{\tau_1 L_A\|g\|_{\rho_1} + \beta L_g}{\mu} + \|g\|_{\rho_1}\beta\frac{\tau_2 L_A}{\mu^2} \\ &= \left(\frac{\tau_1}{\mu} + \frac{2\beta_0\tau_2}{\mu^2}\right) L_A\|g\|_{\rho_1} + \frac{2\beta_0}{\mu}L_g \equiv C_1 L_A\|g\|_{\rho_1} + C_2 L_g, \end{aligned}$$

where μ has been defined in Lemma 2.9 and C_1 and C_2 do not depend on the step of the iterative process.

Case 2: $k \neq 0$.

$$\begin{aligned} L_{y_i^k} &= \frac{L_{h^k}}{\frac{c}{|k|^\gamma}e^{-\nu|k|}} + |h^k|\frac{L_{\lambda_i}}{\left(\frac{c}{|k|^\gamma}e^{-\nu|k|}\right)^2} \\ &\le \frac{|k|^\gamma e^{\nu|k|}}{c} L_h e^{-\rho_1|k|} + \frac{|k|^{2\gamma}e^{2\nu|k|}}{c^2}\|h\|_{\rho_1} e^{-\rho_1}\tau_2 L_A \\ &\le |k|^{2\gamma} e^{-(\rho_1 - 2\nu)|k|}\left[\frac{\tau_1}{c}\|g\|_{\rho_1} L_A + \frac{\beta}{c}L_g + \frac{\beta\tau_2}{c^2}\|g\|_{\rho_1} L_A\right] \\ &\equiv |k|^{2\gamma} e^{-(\rho_1 - 2\nu)|k|}\left[C_3\|g\|_{\rho_1} L_A + C_4 L_g\right], \end{aligned}$$

where now C_3 and C_4 do not depend on the step of the iterative process.

Now we can apply this to bound the Lipschitz constant L_y corresponding to $y(t,\varepsilon) = \sum_k y_i^k(\varepsilon)e^{(k,\omega)\sqrt{-1}t}$. From Lemma 2.19, we obtain

$$L_y = C_1 L_A\|g\|_{\rho_1} + C_2 L_g + (C_3 L_A\|g\|_{\rho_1} + C_4 L_g)\frac{20r}{3\delta^{r+2\gamma}}\chi(r+2\gamma),$$

where $0 < \rho_2 < \rho_1 - 2\nu$ such that $\delta = \rho_1 - \rho_2 - 2\nu \le 1$. To simplify the following steps, we note that

$$L_y \le \frac{\chi(r+2\gamma)}{\delta^{r+2\gamma}}(C_5 L_A\|g\|_{\rho_1} + C_6 L_g)$$

for suitable constants C_5 and C_6, both independent on the actual step of the inductive process.

Since $x = B(\varepsilon)y$, we have $L_x \equiv \mathcal{L}(x) \le \tau_1 L_A\|y\|_{\rho_2} + \beta L_y$, which allows us (using the bound on $\|y\|_{\rho_2}$ given inside the proof of Lemma 2.10) to establish the bound

$$L_x \le \tau_1 L_A\left[\frac{1}{\mu} + \frac{20r\chi(r+\gamma)}{3c\delta^{r+\gamma}}\right]\beta\|g\|_{\rho_1} + \beta\frac{\chi(r+2\gamma)}{\delta^{r+2\gamma}}(C_5 L_A\|g\|_{\rho_1} + C_6 L_g).$$

This can easily be rearranged to

$$L_x \le \frac{\chi(r+2\gamma)}{\delta^{r+2\gamma}}(E_1 L_A\|g\|_{\rho_1} + E_2 L_g),$$

where E_1 and E_2 are suitable constants not dependent on the actual step of the inductive process. □

2.9. Measure lemma. Here we give the basic lemma used to bound the measure of the resonances.

LEMMA 2.24. *Let $\omega \in \mathbb{R}^r$ and $v \in \sqrt{-1}\mathbb{R}$ such that*

$$|v - \sqrt{-1}(k,\omega)| \geq \frac{2c}{|k|^{\gamma_0}}$$

for all $k \in \mathbb{Z}^r \setminus \{0\}$, where $c > 0$ and $\gamma_0 > 0$. We define the nth resonant subset $\mathcal{R}_\mu^{(n)} = \mathcal{R}_\mu^{(n)}(v)$ as

$$\mathcal{R}_\mu^{(n)} = \Big\{\varphi \in \sqrt{-1}\mathbb{R},\ |\varphi| < \mu \ / \ \exists k' \in \mathbb{Z}^r \setminus \{0\}$$

$$\text{such that } |\varphi + v - \sqrt{-1}(k',\omega)| < \frac{c}{|k'|^{\gamma_n}} e^{-\nu_n |k'|}\Big\},$$

where $\nu_n = \frac{\nu_0}{(n+1)^2}$, $0 < \nu_0 < \frac{1}{4}$, $\gamma_n = \gamma_0 z^n$, $1 < z < 2$, and $\gamma_0 \geq r - 1$. Let $\mathcal{R}_\mu = \cup_{n\geq 0}\mathcal{R}_\mu^{(n)}$ and $\psi(\mu) = \frac{m(\mathcal{R}_\mu)}{2\mu}$, where m denotes the Lebesgue measure. Then $\psi(\mu) \leq \exp(-c_1/\mu^{c_2})$ for some positive constants c_1 and c_2, where $c_2 < \frac{1}{\gamma_0}$, provided μ is small enough.

Proof. Let k' and φ be such that

$$|\varphi + v - \sqrt{-1}(k',\omega)| < \frac{c}{|k'|^{\gamma_n}} e^{-\nu_n |k'|}.$$

Since

$$|v - \sqrt{-1}(k',\omega)| \geq \frac{2c}{|k'|^{\gamma_0}},$$

we have

$$\mu > |\varphi| > \frac{2c}{|k'|^{\gamma_0}} - \frac{c}{|k'|^{\gamma_n} \exp(\nu_n |k'|)} > \frac{c}{|k'|^{\gamma_0}},$$

and hence $|k'| \geq \lceil (\frac{c}{\mu})^{1/\gamma_0} \rceil \equiv M(\mu)$, where for $\alpha \in \mathbb{R}$, $\lceil \alpha \rceil$ denotes the lowest integer greater than or equal to α. Now let us add for all $|k'| \geq M(\mu)$ and all $n \geq 0$ to have an upper bound on $m(\mathcal{R}_\mu)$. Adding for all $|k'| \geq M(\mu)$ and for a fixed n, we obtain

$$\sum_{|k'| \geq M(\mu)} \frac{2c}{|k'|^{\gamma_n} \exp(\nu_n |k'|)} 2c \sum_{j \geq M(\mu)} \frac{2rj^{r-1}}{j^{\gamma_n}} e^{-\nu_n j}$$

$$< 4crM(\mu)^{r-1-\gamma_n} \frac{e^{-\nu_n M(\mu)}}{1 - e^{-\nu_n}} < \frac{5cr}{\nu_0} M(\mu)^{r-1-\gamma_n} (n+1)^2 e^{-\nu_n M(\mu)}$$

because $r - 1 - \gamma_n \leq 0$ and $1 - e^{-\alpha} > 0.8\alpha$ if $0 \leq \alpha \leq \frac{1}{4}$. Adding for all n, we have

$$m(\mathcal{R}_\mu) \leq \frac{5cr}{\nu_0} M(\mu)^{r-1} \sum_{n\geq 0} (n+1)^2 M(\mu)^{-\gamma_0 z^n} \exp\left(-\frac{\nu_0}{(n+1)^2} M(\mu)\right).$$

For our purposes, a rough bound is enough. Let $n_* = \log(\nu_0 M(\mu))/\log z$. We split $\sum_{n\geq 0}$ as $\sum_{n=0}^{n_*} + \sum_{n > n_*}$ and assume μ small enough so that

$$\frac{(n+2)^2 M(\mu)^{-\gamma_0 z^{n+1}}}{(n+1)^2 M(\mu)^{-\gamma_0 z^n}} \leq \frac{1}{2} \quad \forall n \geq 0.$$

Then

$$\sum_{n\geq 0}(n+1)^2 M(\mu)^{-\gamma_0 z^n}\exp\left(-\frac{\nu_0}{(n+1)^2}M(\mu)\right)$$
$$\leq 2M(\mu)^{-\gamma_0}\exp\left(-\frac{\nu_0}{(n_*+1)^2}M(\mu)\right)+(n_*+1)^2 M(\mu)^{-\gamma_0 z^{n_*}}.$$

To finish the proof, after selecting any value of c_1 and $c_2 < \frac{1}{\gamma_0}$, we want to show that each term is less than $\frac{1}{2}\exp(-\frac{c_1}{\mu^{c_2}})$. To this end, we take logarithms. We have to prove

$$\log A - \log\left(\frac{c}{\mu}\right) - \frac{\nu_0}{\left[\frac{\log\left(\nu_0\left(\frac{c}{\mu}\right)^{\frac{1}{\gamma_0}}\right)}{\log z}+1\right]^2}\left(\frac{c}{\mu}\right)^{\frac{1}{\gamma_0}} < -\log 4 - \frac{c_1}{\mu^{c_2}},$$

$$\log A + 2\log\left[\frac{\log\left(\nu_0\left(\frac{c}{\mu}\right)^{\frac{1}{\gamma_0}}\right)}{\log z}+1\right] - \nu_0\left(\frac{c}{\mu}\right)^{\frac{1}{\gamma_0}}\log\left(\frac{c}{\mu}\right) < -\log 2 - \frac{c_1}{\mu^{c_2}},$$

where $A = \frac{5cr}{\nu_0}(\frac{c}{\mu})^{(r-1)/\gamma_0}$. Both inequalities are true if μ is small enough or, equivalently, for μ in a fixed range if c_1 is big enough. □

2.10. Proof of Theorem 2.1 (part II). Thus far, we have shown the convergence of the iterative scheme provided that some nonresonance conditions hold at each step n (see §2.7). Now our purpose is to show that all the matrices $A_n(\varepsilon)$ are Lipschitz (with respect to ε) from above and below and that their Lipschitz constants are bounded (from above and below, respectively) by constants that do not depend on n. As we shall see later, this allows us to take out a dense set (with small relative measure) of values of ε for which the resonance conditions assumed during §2.7 might not hold at some step (i.e., for some n) of the proof.

To prove that $A_n(\varepsilon)$ is Lipschitz from below, we shall proceed in the following way. Since $A_0(\varepsilon)$ is Lipschitz from above and below (by hypothesis), it is enough to show that $A_n(\varepsilon)$ is Lipschitz from above and $\mathcal{L}(A_0) - \mathcal{L}(A_n) = \mathcal{O}(\varepsilon)$ since Lemma 2.16 implies that if ε is sufficiently small, $A_n(\varepsilon)$ is also Lipschitz from below. For this reason, we will focus on Lipschitz constants from above, which for simplicity will simply be called Lipschitz constants. The notation used will be

$$\mathcal{L}(A_n(\varepsilon)) = L_{A_n}, \quad \mathcal{L}(\widehat{A}_n(\varepsilon)) = L_{\widehat{A}_n}, \quad \mathcal{L}(\varepsilon^{2^n}Q_n(t,\varepsilon)) = L_{Q_n},$$
$$\mathcal{L}(\varepsilon^{2^n}\widehat{Q}_n(t,\varepsilon)) = L_{\widehat{Q}_n}, \quad \mathcal{L}(\varepsilon^{2^n}g_n(t,\varepsilon)) = L_{g_n}, \quad \mathcal{L}(\varepsilon^{2^n}\widehat{g}_n(t,\varepsilon)) = L_{\widehat{g}_n},$$
$$\mathcal{L}(\varepsilon^{2^n}P_n(t,\varepsilon)) = L_{P_n}, \quad \mathcal{L}(h_n(x,t,\varepsilon)) = L_{h_n}, \quad \mathcal{L}(\widehat{h}_n(x,t,\varepsilon)) = L_{\widehat{h}_n}.$$

Our purpose now is to bound the Lipschitz constants of the equation terms at step $n+1$ as a function of the Lipschitz constants at step n. Let us assume that the scheme of §2.7 has been applied up to step n. Then the following bounds can be established:

$$\|\widehat{Q}_n\|_{\sigma_n} \leq N_1^{2^n},$$
$$\|P_n\|_{\rho_{n+1}} \leq N_1^{2^n},$$

$$\|\underline{x}_n\|_{\sigma_n} \leq (\varepsilon_0 N_1)^{2^n},$$
$$\|\widehat{g}_n\|_{\sigma_n} \leq N_1^{2^{n+1}},$$

where N_1 is a positive constant and ε_0 is as defined at the end of §2.7 and is also assumed to satisfy $\varepsilon_0 N_1 < 1$. Using the bounds given in §2.7, the proofs are not difficult.

We shall also use the bounds $\|Q_n\|_{\rho_n} \leq N_1^{2^n}$ and $\|g_n\|_{\rho_n} \leq N_1^{2^n}$ (see §2.7) when needed. The constant N_1 can be easily obtained from the constants M_i, $i = 1, \ldots, 4$, introduced in §2.7.

Now let us bound $L_{Q_{n+1}}$. It is easy to obtain

$$\varepsilon^{2^{n+1}} Q_{n+1} = (I + \varepsilon^{2^n} P_n(t,\varepsilon))^{-1} (\varepsilon^{2^n} \widehat{Q}_n(t,\varepsilon))(\varepsilon^{2^n} P_n(t,\varepsilon)),$$

and this implies

$$\begin{aligned} L_{Q_{n+1}} \leq\ & 4L_{P_n} \varepsilon_0^{2^n} \|\widehat{Q}_n\|_{\sigma_n} \varepsilon_0^{2^n} \|P_n\|_{\rho_{n+1}} + \|(I + \varepsilon^{2^n} P_n)^{-1}\|_{\rho_{n+1}} L_{\widehat{Q}_n} \varepsilon_0^{2^n} \|P_n\|_{\rho_{n+1}} \\ & + \|(I + \varepsilon^{2^n} P_n)^{-1}\|_{\rho_{n+1}} \varepsilon_0^{2^n} \|\widehat{Q}_n\|_{\sigma_n} L_{P_n}. \end{aligned}$$

Using the fact that $\|(I + \varepsilon^{2^n} P_n)^{-1}\|_{\rho_{n+1}} \leq 2$ (see §2.7), we obtain

$$\begin{aligned} L_{Q_{n+1}} &\leq 4L_{P_n} (\varepsilon_0 N_1)^{2^n} (\varepsilon_0 N_1)^{2^n} + 2L_{\widehat{Q}_n} (\varepsilon_0 N_1)^{2^n} + 2(\varepsilon_0 N_1)^{2^n} L_{P_n} \\ &< 6(\varepsilon_0 N_1)^{2^n} L_{P_n} + 2(\varepsilon_0 N_1)^{2^n} L_{\widehat{Q}_n}. \end{aligned}$$

Now we consider $L_{\underline{x}_n}$. From Lemma 2.23, it is not difficult to obtain

$$L_{\underline{x}_n} \leq \frac{\chi(r + 2\gamma_n)}{\delta_n^{r+2\gamma_n}} (E_1 L_{A_n} \varepsilon_0^{2^n} \|g_n\|_{\rho_n} + E_2 L_{g_n}),$$

where δ_n is now $\frac{1/2 - 2\nu_0}{(n+1)^2}$. Introducing $L_{3,n} = \frac{\chi(r+2\gamma_n)}{\delta_n^{r+2\gamma_n}} \max\{E_1, E_2\}$, we can write

$$L_{\underline{x}_n} \leq L_{3,n} \left(L_{A_n} (\varepsilon_0 N_1)^{2^n} + L_{g_n} \right). \tag{17}$$

Let us consider now $L_{\widehat{Q}_n}$. We recall that $\widehat{Q}_n$ was defined as

$$\varepsilon^{2^n} \widehat{Q}_n = \varepsilon^{2^n} Q_n(t,\varepsilon) + D_x h_n(\underline{x}_n(t,\varepsilon), t, \varepsilon).$$

This allows us to write

$$L_{\widehat{Q}_n} \leq L_{Q_n} + \|D_{xx} h_n(\underline{x}_n(t,\varepsilon), t, \varepsilon)\| L_{\underline{x}_n} + \mathcal{L}(D_x h_n(\underline{x}, t, \varepsilon)).$$

Here we can use the fact that $\|D_{xx} h_n(\underline{x}(t,\varepsilon), t, \varepsilon)\| \leq K_\infty$ (see §2.7), Lemma 2.17, and (17) to write

$$L_{\widehat{Q}_n} \leq L_{Q_n} + K_\infty L_{3,n} \left(L_{A_n} (\varepsilon_0 N_1)^{2^n} + L_{g_n} \right) + K_1(\alpha) L_{h_n} (\varepsilon_0 N_1)^{2^n},$$

where $\alpha = \frac{\|\underline{x}_n\|_{\sigma_n}}{r_\infty}$. Moreover, note that α goes to zero when $\|\underline{x}_n\|_{\sigma_n}$ does. This implies that if ε is small enough, we can assume that $K_1(\alpha)$ is less than, for instance, $K_1(\frac{1}{2})$.

Now we focus on L_{P_n}. The definition of P_n is

$$\varepsilon^{2^n}\dot{P}_n(t,\varepsilon) = \widehat{A}_n(\varepsilon^{2^n}P_n(t,\varepsilon)) - (\varepsilon^{2^n}P_n(t,\varepsilon))\widehat{A}_n + \varepsilon^{2^n}\widehat{Q}_n(t,\varepsilon).$$

Since this is a linear system of differential equations, we can apply a lemma that is essentially like Lemma 2.23 but for the actual system of equations and with new constants $\overline{E}_1$ and $\overline{E}_2$ to get

$$\begin{aligned} L_{P_n} &\leq \frac{\chi(r+2\gamma_n)}{\delta_n^{r+2\gamma_n}}\left(\overline{E}_1\|\varepsilon_0^{2^n}\widehat{Q}_n\|_{\sigma_n}L_{A_n} + \overline{E}_2L_{\widehat{Q}_n}\right) \\ &\leq \frac{\chi(r+2\gamma_n)}{\delta_n^{r+2\gamma_n}}\left(\overline{E}_1(\varepsilon_0N_1)^{2^n}L_{\widehat{A}_n} + +\overline{E}_2L_{\widehat{Q}_n}\right) \leq L_{3,n}\left((\varepsilon_0N_1)^{2^n}L_{\widehat{A}_n} + L_{\widehat{Q}_n}\right), \end{aligned}$$

where $L_{3,n}$ has been redefined as $L_{3,n} = \frac{\chi(r+2\gamma_n)}{\delta_n^{r+2\gamma_n}}\max\{E_1, E_2, \overline{E}_1, \overline{E}_2\}$.

Let us consider $L_{g_{n+1}}$. From

$$\varepsilon^{2^{n+1}}g_{n+1}(t,\varepsilon) = (I + \varepsilon^{2^n}P_n(t,\varepsilon))^{-1}(\varepsilon^{2^{n+1}}\widehat{g}_n(t,\varepsilon)),$$

it follows that

$$\begin{aligned} L_{g_{n+1}} &\leq 4L_{P_n}\varepsilon_0^{2^{n+1}}\|\widehat{g}_n\|_{\sigma_n} + \|(I + \varepsilon^{2^n}P_n(t,\varepsilon))^{-1}\|_{\rho_{n+1}}L_{\widehat{g}_n} \\ &\leq 4L_{P_n}(\varepsilon_0N_1)^{2^{n+1}} + 2L_{\widehat{g}_n}. \end{aligned}$$

Now consider now $L_{\widehat{A}_n}$. Since we have

$$\widehat{A}_n(\varepsilon) = A_n(\varepsilon) + \varepsilon^{2^n}\overline{\widehat{Q}}_n(\varepsilon),$$

it follows that

$$L_{\widehat{A}_n} = L_{A_n} + L_{\widehat{Q}_n}.$$

Moreover, since $A_{n+1} = \widehat{A}_n$, we also have

$$L_{A_{n+1}} = L_{\widehat{A}_n}.$$

Now let us bound $L_{\widehat{g}_n}$. Recall that

$$\varepsilon^{2^{n+1}}\widehat{g}_n(t,\varepsilon) = h_n(\underline{x}_n(t,\varepsilon),t,\varepsilon) + \varepsilon^{2^n}Q_n(t,\varepsilon)\underline{x}_n(t,\varepsilon),$$

which implies

$$\begin{aligned} L_{\widehat{g}_n} &\leq \|D_xh_n(\underline{x}_n(t,\varepsilon),t,\varepsilon)\|_{\sigma_n}L_{\underline{x}_n} + L_{h_n(\underline{x}_n,t,\varepsilon)} \\ &\quad + L_{Q_n}\|\underline{x}_n(t,\varepsilon)\|_{\sigma_n} + \|\varepsilon^{2^n}Q_n(t,\varepsilon)\|_{\rho_n}L_{\underline{x}_n} \\ &\leq [K_\infty\|\underline{x}_n(t,\varepsilon)\|_{\sigma_n} + \|\varepsilon^{2^n}Q_n(t,\varepsilon)\|_{\rho_n}]L_{\underline{x}_n} \\ &\quad + K_2\left(\frac{1}{2}\right)\|\underline{x}_n(t,\varepsilon)\|_{\sigma_n}^2L_{h_n} + \|\underline{x}_n(t,\varepsilon)\|_{\sigma_n}L_{Q_n}, \end{aligned}$$

where we have used the fact that $L_{h_n(\underline{x}_n,t,\varepsilon)} \leq K_2(\alpha)L_{h_n}\|\underline{x}_n\|^2$ and, as before, that $K_2(\alpha) \leq K_2(\frac{1}{2})$ if ε_0 is small enough. On the other hand,

$$K_\infty\|\underline{x}_n(t,\varepsilon)\|_{\sigma_n} + \|\varepsilon^{2^n}Q_n(t,\varepsilon)\|_{\rho_n} \leq (K_\infty + 1)(\varepsilon_0N_1)^{2^n}.$$

This implies

$$L_{\widehat{g}_n} \le K_3 L_{3,n} (\varepsilon_0 N_1)^{2^n} (L_{A_n} + L_{g_n} + L_{h_n} + L_{Q_n}),$$

where $K_3 = \max\{K_\infty + 1,\ K_2(\frac{1}{2})\}$ and we have used $\varepsilon_0 N_1 < 1$ and $L_{3,n} > 1$.

Let us follow with $L_{h_{n+1}}$. Since

$$h_{n+1}(x_{n+1}, t, \varepsilon) = (I + \varepsilon^{2^n} P_n(t,\varepsilon))^{-1} \widehat{h}_n((I + \varepsilon^{2^n} P_n(t,\varepsilon)) x_{n+1}, t, \varepsilon),$$

we have

$$\begin{aligned} L_{h_{n+1}} &\le 4 L_{P_n} \|\widehat{h}_n((I + \varepsilon^{2^n} P_n) x_{n+1}, t, \varepsilon)\|_{\rho_{n+1}} \\ &\quad + \|(I + \varepsilon^{2^n} P_n)^{-1}\|_{\rho_{n+1}} \|D_x \widehat{h}_n\|_{\rho_{n+1}} L_{P_n} + \|(I + \varepsilon^{2^n} P_n)^{-1}\|_{\rho_{n+1}} L_{\widehat{h}_n} \\ &\le 4 L_{P_n} \frac{K_\infty}{2} \tau_\infty + 2 K_\infty \tau_\infty L_{P_n} + 2 L_{\widehat{h}_n}, \end{aligned}$$

which allows us to write

$$L_{h_{n+1}} \le 4 K_\infty \tau_\infty L_{P_n} + 2 L_{\widehat{h}_n}.$$

Finally, let us consider $L_{\widehat{h}_n}$. We recall that

$$\widehat{h}_n(y_n, t, \varepsilon) = h_n(\underline{x}_n(t,\varepsilon) + y_n, t, \varepsilon) - h_n(\underline{x}_n(t,\varepsilon), t, \varepsilon) - D_x h_n(\underline{x}_n(t,\varepsilon), t, \varepsilon) y_n,$$

which implies

$$\begin{aligned} L_{\widehat{h}_n} &\le \|D_x h_n(\underline{x}_n(t,\varepsilon) + y_n, t, \varepsilon)\| L_{\underline{x}_n} + L_{h_n} + \|D_x h_n(\underline{x}_n(t,\varepsilon), t, \varepsilon)\| L_{\underline{x}_n} \\ &\quad + L_{h_n} + \|D_{xx} h_n(\underline{x}_n(t,\varepsilon), t, \varepsilon)\| \|y_n\| L_{\underline{x}_n} + \|y_n\| L_{D_x h_n} \\ &\le K_\infty \tau_\infty L_{\underline{x}_n} + L_{h_n} + K_\infty \tau_\infty L_{\underline{x}_n} + L_{h_n} + K_\infty \tau_\infty L_{\underline{x}_n} + \tau_\infty K_1(\alpha) \alpha L_{h_n}. \end{aligned}$$

Furthermore, if ε is small enough, we have $\tau_\infty K_1(\alpha)\alpha \le 1$, and the following bound can be obtained:

$$L_{\widehat{h}_n} \le 3 K_\infty \tau_\infty L_{\underline{x}_n} + 3 L_{h_n}.$$

Thus far, we have stated some bounds on the Lipschitz constants. The next step is to relate (in closed formulas) the bounds of step $n+1$ with bounds of step n.

Let us define $a_n = L_{A_n}$, $b_n = \max\{L_{Q_n}, L_{g_n}\}$, and $c_n = L_{h_n}$; and let $e_n = (\varepsilon_0 N_1)^{2^n}$. Furthermore, let $L_{4,n} = L_{3,n} \max\{K_\infty, 6, 2K_3, 6K_\infty \tau_\infty\}$. After some rearrangement, we can write the bounds on the recurrences as

$$\begin{aligned} a_{n+1} &\le a_n + b_n + L_{4,n}(a_n e_n + b_n) + K_1(\alpha) e_n c_n, \\ b_{n+1} &\le 5 L_{4,n}^2 e_n a_n + 8 L_{4,n}^2 e_n b_n + (4K_1(\alpha) + 1) L_{4,n} e_n c_n, \\ c_{n+1} &\le 3 L_{4,n}^2 e_n a_n + 4 L_{4,n}^2 b_n + (6 + 2K_1(\alpha) L_{4,n}) c_n. \end{aligned} \tag{18}$$

Let $d_n = \max\{a_n, b_n, c_n\}$. It is immediate (recalling that $e_n < 1$ and $L_{4,n} > 1$) to obtain

$$d_{n+1} \le R L_{4,n}^2 d_n,$$

where $R = 14+4K_1(\alpha)$. As before, it is easy to obtain $RL_{4,n}^2 \leq M_6^{2^n}$ for some suitable M_6 independent of n. Therefore,

$$d_n \leq \prod_{j=0}^{n} M_6^{2^j} d_0 < M_6^{2^{n+1}} d_0.$$

Going back to (18), we have

$$\begin{aligned} b_{n+1} \leq \frac{5}{14} M_6^{2^n} (\varepsilon_0 N_1)^{2^n} M_6^{2^{n+1}} d_0 + \frac{8}{14} M_6^{2^n} (\varepsilon_0 N_1)^{2^n} M_6^{2^{n+1}} d_0 \\ + M_6^{2^n} (\varepsilon_0 N_1)^{2^n} M_6^{2^{n+1}} d_0 \leq 2d_0 \left(N_1 M_6^3\right)^{2^n} \varepsilon_0^{2^n} \leq (\varepsilon_0 S)^{2^n}, \end{aligned}$$

where S is a constant independent of n and ε_0. Taking $\varepsilon_0 < S^{-1}$, we have $b_n \longrightarrow 0$ as $n \longrightarrow 0$. Furthermore, we also obtain

$$a_{n+1} - a_n < (\varepsilon_0 T)^{2^n}$$

for a suitable constant T independent of n and ε_0. We also require $\varepsilon_0 < T^{-1}$. Then

$$a_n - a_0 < \sum_{j=0}^{n-1} (\varepsilon_0 T)^{2^j} < 2T\varepsilon_0$$

for all n provided that $\varepsilon_0 < \min\{S^{-1}, \frac{1}{2}T^{-1}\}$.

This is the bound we were looking for; it shows that $A_n(\varepsilon)$ is a Lipschitz function of ε and that $\mathcal{L}(A_n) - \mathcal{L}(A_0) = \mathcal{O}(\varepsilon)$. This means that, using Lemmas 2.16 and 2.22, the eigenvalues λ_j^n and the differences $\lambda_{j_1}^n - \lambda_{j_2}^n$ are Lipschitz from above and from below if ε is small enough.

To complete the proof, we take into account the resonances. Since we want to skip the possible resonances due to λ_j^n and $\lambda_{j_1}^n - \lambda_{j_2}^n$ at the nth step, we have to apply Lemma 2.24 for each one of the eigenvalues and couples. This amounts to skipping a measure at most d^2 times that we skipped in the frequency space. To go back to the parameter space, that is, to ε, we use the Lipschitz constant from below. In this way, we obtain the Cantorian $\mathcal{E}$ with the desired properties. □

2.11. Proof of Theorem 2.4. Since $\det A \neq 0$, the contraction lemma ensures that if ε is small enough, there exists a function $x_0(\varepsilon)$ such that

$$(A + \varepsilon\overline{Q})x_0(\varepsilon) + \varepsilon\overline{g} + \overline{h}(x_0(\varepsilon), \varepsilon) = 0,$$

and it verifies that $x_0(\varepsilon) = \mathcal{O}(\varepsilon)$. Let us define

$$A_{x_0} = A + \varepsilon\overline{Q} + D_x\overline{h}(x_0(\varepsilon), \varepsilon),$$

and let $\underline{x}(t, \varepsilon)$ be such that

$$\dot{\underline{x}} = A\underline{x} + \varepsilon g(t, \varepsilon). \tag{19}$$

(The existence of this solution was shown in Lemma 2.10, and we recall that it was proved to be $\mathcal{O}(\varepsilon)$.) The terms of order ε of the matrix A_∞ are provided by Lemma 2.12 at the first step of the inductive process. This modified matrix $\widehat{A}$ is

$$\widehat{A} = A + \varepsilon\overline{Q} + \overline{D_x h(\underline{x}, t, \varepsilon)}.$$

Then

$$\|A_{x_0}-\widehat{A}\| = \|D_x\overline{h}(x_0(\varepsilon),\varepsilon) - \overline{D_xh(\underline{x}(t,\varepsilon),t,\varepsilon)}\|.$$

However,

$$\overline{D_xh(\underline{x}(t,\varepsilon),t,\varepsilon)} = \overline{(C+\varepsilon R(t))\underline{x}(t)} + \mathcal{O}(\varepsilon^2) = C\overline{\underline{x}} + \mathcal{O}(\varepsilon^2),$$

where $C = \frac{1}{2}D_{xx}h(0,t,0)$ is a constant matrix by hypothesis. Moreover, it is also easy to obtain that

$$D_x\overline{h}(x_0(\varepsilon),\varepsilon) = (C+\varepsilon\overline{R})x_0(\varepsilon) + \mathcal{O}(\varepsilon^2) = Cx_0(\varepsilon) + \mathcal{O}(\varepsilon^2).$$

We have obtained that

$$\|A_{x_0}-\widehat{A}\| = C(x_0(\varepsilon)-\overline{\underline{x}}) + \mathcal{O}(\varepsilon^2).$$

Now, averaging (19), we get that $A\overline{\underline{x}}+\varepsilon\overline{g}=0$, and using $Ax_0(\varepsilon)+\varepsilon\overline{g} = -(\varepsilon\overline{Q}x_0(\varepsilon)+\overline{h}(x_0(\varepsilon),\varepsilon)) = \mathcal{O}(\varepsilon^2)$, we obtain $\|x_0(\varepsilon)-\overline{\underline{x}}\| = \mathcal{O}(\varepsilon^2)$, which completes the proof. □

3. The neighborhood of an elliptic equilibrium point of a Hamiltonian system. Let us consider the Hamiltonian

$$H^0(p,q,t) = H_0(p) + H_1(p,q,t),$$

where $|H_1|$ is small and depends on time in a quasi-periodic way, with $\varphi = (\varphi_1,\dots,\varphi_r)$ as a vector of basic frequencies. To obtain an autonomous system we define $q_2 = q$, $p_2 = p$, and $q_1 = \varphi t$. Therefore, the Hamiltonian takes the form

$$H(p_1,p_2,q_1,q_2) = (\varphi,p_1) + H_0(p_2) + H_1(p_2,q_2,q_1),$$

where p_1 are the actions corresponding to q_1. (Obviously, they are not relevant in this problem and have only been added to obtain a Hamiltonian form.) We are interested in the invariant tori that the unperturbed system $H = H_0(p_2)$ had. Note that the KAM theorem (see [1]) cannot be applied directly due to the degeneracy of this case.[3]

We have considered this case in Theorem 3.1, and we have found that the proof of the classical KAM theorem (see [1]) still works because the perturbing frequencies are not modified in any step of the inductive process, and we only have to worry about the proper frequencies of the Hamiltonian, which can be controlled provided that the nondegeneracy condition

$$\det\left(\frac{\partial^2 H_0}{\partial(p_2)^2}\right) \neq 0$$

holds. The result obtained is that there exist invariant tori near the origin for ε small enough. The frequencies of these tori are those of the unperturbed tori plus those of the perturbation. This can be described by saying that the unperturbed tori are "quasi-periodically dancing" to the "rhythm" of the perturbation. The tori whose frequencies are in resonance with those of the perturbation are destroyed.

Finally, in the case where the origin is not a fixed point of the perturbed Hamiltonian, we can reduce to this case by performing a change of variables, transforming

[3] However, see the comments in [2, pp. 193–194] for a related result.

the quasi-periodic orbit that replaces the equilibrium point (we recall that this orbit exists for a Cantorian set of values of ε) in a fixed point.

THEOREM 3.1. *Let us consider the Hamiltonian*

$$H(p_1, p_2, q_1, q_2) = (\varphi, p_1) + H_0(p_2) + \varepsilon H_1(p_2, q_1, q_2),$$

where q_1 are the angles of the perturbation, p_1 are the corresponding actions, q_2 and p_2 are the angles and actions of the unperturbed system, and $\varphi = (\varphi_1, \ldots, \varphi_{n_1})$ is a constant vector of frequencies that satisfies the nonresonance condition

$$|(k, \varphi)| > \frac{c}{|k|^\gamma} \quad \forall k \in \mathbb{Z}^{n_1} \setminus \{0\}, \quad \gamma > n_1 - 1.$$

Let G^1 be a compact domain of $\mathbb{R}^{n_1}$, let G^2 be a compact domain of $\mathbb{R}^{n_2}$, and let G be $G^1 \times G^2$. Now suppose that this Hamiltonian function $H(p_1, p_2, q_1, q_2) = H(p, q)$ is analytic on the domain $F = \{(p, q) \,/\, p = (p_1, p_2) \in G, |\mathrm{Im}\, q| \le \rho\}$ and has period 2π with respect to the variables q. Let us assume that in the domain F,

$$\det \left| \frac{\partial^2 H_0}{\partial (p_2)^2} \right| \neq 0.$$

Then if ε is small enough, the motion defined by the Hamiltonian equations

$$\begin{aligned} \dot{p}_1 &= -\frac{\partial H}{\partial q_1}, & \dot{q}_1 &= \varphi, \\ \dot{p}_2 &= -\frac{\partial H}{\partial q_2}, & \dot{q}_2 &= \frac{\partial H}{\partial p_2}, \end{aligned} \tag{20}$$

has the following properties.

1. *There exists a decomposition* $\mathrm{Re}\, F = F_1 + F_2$, *where F_1 is invariant and F_2 is small:* $\mathrm{mes}\, F_2 \le \kappa_1(\varepsilon)\, \mathrm{mes}\, F$, *where $\kappa_1(\varepsilon)$ is $o(\varepsilon)$.*

2. *F_1 is composed of invariant n-dimensional analytic tori I_ϕ, defined parametrically by the equations*

$$p = p_\phi + f_\phi(Q), \qquad q = Q + g_\phi(Q),$$

where f_ϕ and g_ϕ are analytic functions of period 2π in the variables Q and ϕ is a parameter determining the torus I_ϕ. In fact, ϕ consists of all the frequencies, i.e., those of the external excitation and the proper frequencies, $\phi = (\varphi_1, \ldots, \varphi_{n_1}, \omega_1, \ldots, \omega_{n_2})$.

3. *The invariant tori I_ϕ differ little from the tori $p = p_\phi$, namely,*

$$|f_\phi(Q)| < \kappa_2(\varepsilon), \qquad |g_\phi(Q)| < \kappa_2(\varepsilon),$$

where $\kappa_2(\varepsilon)$ is $o(\varepsilon)$.

4. *The motion* (20) *on the invariant torus I_ϕ is quasi-periodic with n frequencies $\varphi_1, \ldots, \varphi_{n_1}, \omega_1, \ldots, \omega_{n_2}$ $(n = n_1 + n_2)$, namely,*

$$\dot{Q} = \phi, \qquad \omega = \left. \frac{\partial H_0}{\partial p_2} \right|_{p_\phi}.$$

3.1. Sketch of the proof of Theorem 3.1. The proof of this theorem is essentially the same as that of the KAM theorem contained in [1], and its technical details can be found in [15]. Here we show the idea of this proof.

Let us define p and q as the vectors p_1, p_2 and q_1, q_2, respectively. Now the Hamiltonian that we have is

$$H^\varepsilon = (\varphi, p_1) + H_0(p_2) + \varepsilon \overline{H}_1(p_2) + \varepsilon \widetilde{H}_1(p_2, q), \tag{21}$$

and let us consider the generating function $S(P,q) = Pq + \sum_{k \neq 0} S^k(P_2) e^{(k,q)\sqrt{-1}}$. If we perform the canonical change of variables

$$\begin{aligned} p_1 &= P_1 + \varepsilon \frac{\partial S}{\partial q_1}, \\ p_2 &= P_2 + \varepsilon \frac{\partial S}{\partial q_2}, \\ Q_1 &= q_1, \\ Q_2 &= q_2 + \varepsilon \frac{\partial S}{\partial P_2} \end{aligned}$$

on (21), we obtain

$$H^\varepsilon = (\varphi, P_1) + H_0(P_2) + \varepsilon \overline{H}_1(P_2) + \varepsilon F + \varepsilon^2 R(P_2, q),$$

where $F = (\varphi, S_{q_1}) + (\omega(P_2), S_{q_2}) + \widetilde{H}_1(P_2, q)$ and $\omega(p_2) = \frac{\partial H_0}{\partial p_2}(p_2)$. Let $\phi(P_2)$ be the vector $\varphi, \omega(P_2)$. We require $F = 0$, namely,

$$(\phi(P_2), S_q) + \widetilde{H}_1(P_2, q) = 0.$$

Now, using the fact that $\widetilde{H}_1(P_2, q) = \sum_{k \neq 0} h_1^k(P_2) e^{(k,q)\sqrt{-1}}$, the coefficients of the Fourier expansion for the generating function S can be obtained easily as

$$S^k(P_2) = \frac{h_1^k}{(\phi(P_2), k)} \sqrt{-1}.$$

To ensure the convergence of this series, it is sufficient to use the usual nonresonance condition

$$|(\phi(P_2), k)| \geq \frac{c}{|k|^\gamma}, \tag{22}$$

which allows us to prove the convergence in a smaller strip than that on which $\widetilde{H}_1$ is analytic. With this, the Hamiltonian takes the form

$$H^\varepsilon = (\varphi, P_1) + H_1^\varepsilon(P_2) + \varepsilon^2 \overline{H}_3(P_2) + \varepsilon^2 \widetilde{H}_3(P_2, Q).$$

This new Hamiltonian is very similar to (21) but with ε^2 instead of ε. Note that the difference between this proof and the one in [1] is condition (22). Due to the fact that the first components of $\phi(P_2)$ are those of φ, *which are constant throughout the inductive process*, we only have to worry about the last ones, $\omega(P_2)$. These components are different at each step of the process, but they can be controlled by the nondegeneracy condition

$$\det\left(\frac{\partial^2 H_0}{\partial (p_2)^2}\right) \neq 0.$$

This is done exactly as shown in [1]. Note that to get a rigorous proof of this theorem, we need only copy the proof contained in [1] and add the "parameter" φ. The unique difference is that now the nonresonance condition is stronger in the sense that we must eliminate a bigger set of (resonant) tori.

Acknowledgments. À. Jorba thanks R. de la Llave and the Department of Mathematics of the University of Texas at Austin for their hospitality. Both authors thank A. Neishtadt for useful comments.

REFERENCES

[1] V. I. Arnol'd, *Proof of a theorem of A. N. Kolmogorov on the invariance of quasi-periodic motions under small perturbations of the Hamiltonian*, Russian Math. Surveys, 18 (1963), pp. 9–36.

[2] V. I. Arnol'd, V. V. Kozlov, and A. I. Neishtadt, *Dynamical Systems* III, Encyclopaedia of Mathematical Sciences, Springer-Verlag, Berlin, 1988.

[3] N. N. Bogoljubov, Ju. A. Mitropoliski, and A. M. Samoilenko, *Methods of Accelerated Convergence in Nonlinear Mechanics*, Springer-Verlag, New York, 1976.

[4] L. Chierchia, *Absolutely continuous spectra of quasiperiodic Schrödinger operators*, J. Math. Phys., 28 (1987), pp. 2891–2898.

[5] C. Dìez, À. Jorba, and C. Simó, *A dynamical equivalent to the equilateral libration points of the earth–moon system*, Celestial Mech. Dynam. Astronom., 50 (1991), pp. 13–29.

[6] E. I. Dinaburg and J. G. Sinai, *The one-dimensional Schrödinger equation with quasiperiodic potential*, Functional Anal. Appl., 9 (1975), pp. 8–21.

[7] L. H. Eliasson, *Perturbations of stable invariant tori for Hamiltonian systems*, Ann. Scuola Norm. Sup. Pisa, 15 (1988), pp. 115–148.

[8] ———, *Floquet solutions for the 1-dimensional quasi-periodic Schrödinger equation*, Comm. Math. Phys., 146 (1992), pp. 447–482.

[9] A. M. Fink, *Almost Periodic Differential Equations*, Lecture Notes in Math. 377, Springer-Verlag, Berlin, 1974.

[10] G. Gómez, À. Jorba, J. Masdemont, and C. Simó, *A quasiperiodic solution as a substitute of L_4 in the earth–moon system*, in Proc. 3rd International Symposium on Spacecraft Flight Dynamics, ESA Publications Division, ESTEC, Noordwijk, the Netherlands, 1991, pp. 35–41.

[11] ———, *Study refinement of semi-analytical halo orbit theory*, final report, contract 8625/89/D/MD(SC), European Space Operations Center, European Space Agency, Paris, 1991.

[12] ———, *Study of Poincaré maps for orbits near Lagrangian points*, final report, contract 9711/91/D/IM(SC), European Space Operations Center, European Space Agency, Paris, 1993.

[13] G. Gómez, J. Llibre, R. Martínez, and C. Simó, *Study on orbits near the triangular libration points in the perturbed restricted three-body problem*, final report, contract 6139/84/D/JS(SC), European Space Opertations Center, European Space Agency, Paris, 1987.

[14] R. A. Johnson and G. R. Sell, *Smoothness of spectral subbundles and reducibility of quasi-periodic linear differential systems*, J. Differential Equations, 41 (1981), pp. 262–288.

[15] À. Jorba, *On quasiperiodic perturbations of ordinary differential equations*, Ph.D. thesis, Universitat de Barcelona, Barcelona, 1991.

[16] À. Jorba and C. Simó, *On the reducibility of linear differential equations with quasiperiodic coefficients*, J. Differential Equations, 98 (1992), pp. 111–124.

[17] J. Moser, *Convergent series expansions for quasiperiodic motions*, Math. Ann., 169 (1967), pp. 136–176.

[18] J. Moser and J. Pöschel, *On the stationary Schrödinger equation with a quasiperiodic potential*, Phys. A, 124 (1984), pp. 535–542.

[19] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Cambridge, UK, 1965.

SIAM J. MATH. ANAL.
Vol. 27, No. 6, pp. 1738–1744, November 1996

PERIODIC MONOTONE SYSTEMS WITH AN INVARIANT FUNCTION[*]

JIANG JI-FA[†]

Abstract. The author studies the periodic time-dependent type-K monotone system

$$\dot{x}_i = F_i(t, x_1, \ldots, x_n) \quad (i = 1, \ldots, n)$$

in the interior of the nonnegative orthant in n-space satisfying the following conditions: (i) if $x \neq y$, $x_i = y_i$, and $x_j \leq y_j$ for $j \neq i$, then $F_i(t,x) \leq F_i(t,y)$; (ii) $F(t,x)$ is periodic in t of period $\tau > 0$; (iii) F possesses an invariant function with positive gradient. It is proved that every solution to such a system either converges to a periodic solution or eventually leaves any compact set. This result gives an affirmative answer to the conjecture recently proposed by B. R. Tang, Y. Kuang, and H. Smith in [*SIAM J. Math. Anal.*, 24 (1993), pp. 1331–1339] for periodic type-K monotone systems.

Key words. type-K monotone system, periodic solution, invariant function, Poincaré map

AMS subject classifications. Primary, 34C25; Secondary, 34D20, 90A16

1. Introduction.

Consider the system of ordinary differential equations

$$\dot{x} = F(t, x), \tag{1}$$

where $F : \mathbf{R} \times P \to \mathbf{R}^n$ is continuous with $P = \{p \in \mathbf{R}^n : p_i > 0 \text{ for all } i,\ 1 \leq i \leq n\}$ and satisfies the following conditions:

(i) $F(t,x)$ is periodic in t of period $\tau > 0$, that is, $F(t+\tau, x) = F(t,x)$ for all $t \in \mathbf{R}$ and all $x \in P$;

(ii) system (1) has the uniqueness property for the initial-value problems on $\mathbf{R} \times P$ and every solution of (1) can be extended into the future;

(iii) for each fixed t, the vector function $F(t,x)$ is of type K in P, i.e., $F_i(t,x) \leq F_i(t,y)$ for any two distinct points $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ in P with $x_i = y_i$ and $x_k \leq y_k (k = 1, 2, \ldots, n;\ k \neq i)$;

(iv) (1) possesses an order-increasing invariant function, i.e., there exists a C^1 function $H : P \to \mathbf{R}$ such that grad $H(x) \gg O$ for each $x \in P$ and

$$\langle \text{grad } H(x), F(t,x) \rangle = 0 \tag{2}$$

for all $t \in \mathbf{R}$ and all $x \in P$. The autonomous system (1) has been studied extensively in the econometria lectures [17, 18] and the papers [4, 13, 19]. The limiting behavior of these systems is well understood. However, if one wishes to build a theory of such systems which reflects changes due to seasonal adjustments, then it is important to study time-dependent systems. In the case $H(x) = \sum_{i=1}^{n} x_i$, (1) is the generalized gross-substitute system which was studied by Nakajima [2] and Sell and Nakajima [3]. Their results show that every compact solution to such a system is asymptotically periodic (almost periodic). When $F(t,x)$ is independent of t and $F_i(x_1, x_2, \ldots, x_n)$ is strictly increasing in x_k for all $k \neq i$, the flow generated by (1) is strongly monotone. Mierczyński [4] proved that every solution either converges to an equilibrium or eventually leaves any compact set. The author's paper [13] investigates the case where $F_i(x_1, x_2, \ldots, x_n)$ is nondecreasing in x_k for all $k \neq i$ and proves the

[*] Received by the editors December 27, 1993; accepted for publication (in revised form) June 27, 1995. This research was supported by National Natural Foundation of China grant 19271001.

[†] Department of Mathematics, University of Science and Technology of China, Hefei, People's Republic of China.

same result. Assume that $F(t,x)$ is continuously differentiable in $x \in P$, $D_xF(t,x)$ is irreducible, and all its off-diagonal terms are nonnegative. If every solution of (1) is compact, then the abstract results of Takáč [6] and Dancer and Hess [9] imply that every solution is asymptotically periodic of period τ. In the recent paper [1], B. R. Tang, Y. Kuang, and H. Smith conjectured that for the periodic (almost periodic) system (1), if conditions (ii)–(iv) hold, then every solution either converges to a periodic solution (an almost periodic solution) or eventually leaves any compact set. Meanwhile, they partly proved this conjecture. More precisely, assuming that H is C^2 with cooperative Hessian matrix, the inequality in (iii) is strict, and some additional conditions for H hold, they cleverly constructed a Liapunov function and proved the above conjecture is true by the theory of skew-product flow.

The object of this paper is to prove that the conjecture proposed by B. R. Tang, Y. Kuang, and H. Smith [1] holds in the case where $F(t,x)$ is periodic in t. We shall prove the following.

THEOREM A. *If system* (1) *satisfies conditions* (i)–(iv), *then every compact solution is asymptotically periodic of period* τ.

This result naturally generalizes that of [2] to the case of a not-necessarily linear-invariant function and drops the strict monotonicity condition in [1, 4] and the irreducibility condition in [6, 9, 14], each of which is needed to guarantee that the system is strongly monotone. Using the ideas in [14], we can prove that the set of fixed points for the Poincaré map T is a curve totally ordered by $\ll$ if the strict monotonicity or irreducibility condition is satisfied. However, if we assume that the system is monotone without strict assumption, then other cases can occur. For example, consider the system

$$\begin{aligned} \dot{x} &= -x + y + z^2, \\ \dot{y} &= x - y + (3 + 2\sin t)z^2, \\ \dot{z} &= -(2 + \sin t)z, \end{aligned} \tag{3}$$

which is defined in $x \geq 0$, $y \geq 0$, and $z \geq 0$ and possesses an invariant function $H = x + y + z^2$. By solving the nonlinear equations in (3), we can easily prove that for any solution $(x(t),\ y(t),\ z(t))$ of (3), there exists a constant $c \geq 0$ such that $(x(t), y(t), z(t)) \to (c, c, 0)$ as $t \to +\infty$ and the set of periodic points for (3) is $\{(c, c, 0) : c \geq 0\}$, which is totally ordered by $\leq$ but not by $\ll$. Moreover, if we add an equation $\dot{w} = 0$ to the system (3), then it becomes a four-dimensional one in which the set of periodic points is $\{(c, c, 0, d) : c \geq 0, d \geq 0\}$. This set is not ordered by $\leq$.

Next, we consider the system

$$\begin{aligned} \dot{x}_1 &= (2 + \cos t)e^{x_1}x_2 - e^{x_1} + 1, \\ \dot{x}_2 &= 1 - e^{-x_1} - (2 + \cos t)x_2 \end{aligned} \tag{4}$$

for $x_1 \geq 0$, $x_2 \geq 0$. System (4) has an invariant function $H(x_1, x_2) = 1 - e^{-x_1} + x_2$. We calculate the corresponding Liapunov function $V(x,y) : \mathbf{R}^2_+ \times \mathbf{R}^2_+ \to \mathbf{R}_+$, which is defined in [1, p. 1334]:

$$V(x,y) = |e^{-x_1} - e^{-y_1}| + |x_2 - y_2|.$$

It is easy to see that $\lim_{|x-y| \to +\infty} V(x,y) = +\infty$ does not hold. Therefore, Theorem 3.1 in [1] cannot be applied to system (4). However, Theorem A can be applied.

2. Definitions and preliminary lemmas. Let x, $y \in \mathbf{R}^n$. There exists a partial order in $\mathbf{R}^n$ given by $x \leq y$ ($x \ll y$) if and only if $x_i \leq y_i$ ($x_i < y_i$) for $i = 1, 2, \ldots, n$. We write $x < y$ to signify that $x \leq y$ and $x \neq y$. If x and y are two vectors with $x \ll y$, let $[[x, y]] = \{z : x \ll z \ll y\}$, $[x, y]] = \{z : x \leq z \ll y\}$, and $[[x, y] = \{z : x \ll z \leq y\}$. If x and y are two vectors with $x < y$, let $[x, y] = \{z : x \leq z \leq y\}$. $[p, \infty]]$ is the set $\{x : x \geq p\}$. If $A \subset P$ is a set, then $a \leq A$ ($A \leq a$) means $a \leq x$ ($x \leq a$) for all $x \in A$; similar notation holds for $a < A$, etc.

We denote by $\varphi(t, x)$ the solution of (1) satisfying $\varphi(0, x) = x$. For $x \in P$, we also write $x(t)$ for $\varphi(t, x)$.

Fundamental to our study is Kamke's theorem, which is stated as follows.

KAMKE'S THEOREM. *Assume* (ii) *and* (iii) *hold. Let $x(t)$ and $y(t)$ be solutions of* (1) *defined for $a \leq t \leq b$ such that $x(a) \ll y(a)$ (resp., $x(a) \leq y(b)$). Then $x(t) \ll y(t)$ (resp., $x(t) \leq y(t)$) for all $t \in [a, b]$.*

W. A. Coppel [10. p. 30] discussed the following question: assuming that $x(a) \leq y(a)$, under what conditions can we have the equality $x_i(b) = y_i(b)$? His argument shows that if (ii) and (iii) hold, then $x_i(b) = y_i(b)$ if and only if $x_i(t) = y_i(t)$ for any $t \in [a, b]$.

Define the Poincaré map $T : P \to P$ by

$$Tx = \varphi(\tau, x).$$

Then T is continuous and $T^k x = \varphi(k\tau, x)$ for any positive integer k. Let $T_i^k x$ denote the ith component of $T^k x$. If the dimension n of Euclidean space $\mathbf{R}^n$ is fixed, we let $N = \{1, 2, \ldots, n\}$. If $I \subset N$, we denote by $C(I)$ the complement of I in N, $C(I) = N - I$. From Kamke's theorem and the argument discussed by Coppel in [10, p. 30], we obtain the following.

LEMMA 2.1. *If $x < y$ with $x_i < y_i$ for $i \in I$, then $T^k x < T^k y$ and $T_i^k x < T_i^k y$ for $i \in I$ and any fixed positive integer k.*

DEFINITION. *Let $x(t)$ be a solution of system* (1) *defined on $[t_o, +\infty)$ for some $t_o \in \mathbf{R}$. $x(t)$ is said to be* compact *if there exist two positive vectors $a, b \in P$ such that $a \leq x(t) \leq b$ for all $t \geq t_o$.*

A compact solution $x(t)$ is said to be *asymptotically periodic of period τ* if there is a periodic solution $y(t)$, $y(t + \tau) = y(t)$, such that $x(t) - y(t) \to O$ as $t \to +\infty$.

Suppose $\varphi(t, x)$ is a compact solution of (1). The positive semiorbit of x is defined by $O^+(x) = \{T^k x : k \in \mathbf{Z}_+\}$, where $\mathbf{Z}_+ = \{0, 1, 2, \ldots\}$ and the ω-limit set of x is defined by $\omega(x) = \{y : T^{n_k} x \to y (k \to \infty)$ for some sequence $n_k \to \infty$ in $\mathbf{Z}_+\}$. Then $\omega(x)$ is fully invariant, i.e., $T\omega(x) = \omega(x)$.

LEMMA 2.2. *Assume that $\varphi(t, x)$ is a compact solution of* (1) *and $\omega(x)$ is the ω-limit set of x. Then for any $y \in \omega(x)$, $\varphi(t, y)$ is also a compact solution of* (1).

Proof. By the definition of $\omega(x)$, there exists some sequence $n_k \to \infty (k \to \infty)$ such that $\lim_{k \to \infty} T^{n_k} x = y$, i.e., $\varphi(n_k \tau, x) \to y$, as $k \to \infty$. Since the solutions of (1) are continuous with respect to initial conditions (see [11, p. 94]), for any fixed $t > 0$, $\varphi(t, y) = \lim_{k \to \infty} \varphi(t, \varphi(n_k \tau, x))$. By condition (i), $F(t, x)$ is periodic in t of period τ. Therefore, $\varphi(t, \varphi(n_k \tau, x)) = \varphi(t + n_k \tau, x)$. The compactness of $\varphi(t, x)$ implies that there exist two positive vectors a, $b \in P$ such that $a \leq \varphi(t, x) \leq b$ for any $t \geq 0$. It follows that

$$a \leq \varphi(t, \varphi(n_k \tau, x)) \leq b.$$

Letting $k \to \infty$, we have

$$a \leq \varphi(t, y) \leq b,$$

i.e., $\varphi(t, y)$ is compact. This completes the proof.

Finally, let $E = \{p \in P : Tp = p\}$ denote the set of all fixed points for T.

3. Proof of Theorem A.

It is convenient to establish some preliminary results before proceeding to the proof of Theorem A. In all of these, we assume the hypotheses of the theorem hold.

LEMMA 3.1. *Suppose $p \in E$. Then for any neighborhood W of p in P, there exists another neighborhood $U \subset W$ of p such that $T^k U \subset U$ for any positive integer k. Therefore, the fixed point p of T is Liapunov stable.*

Proof. Fix a, $b \in W$ with $a \ll p \ll b$ and $[a, b] \subset W$. Then we may choose c, $d \in [[a, b]]$ so that $c \ll p \ll d$. Let $h_1 = H(c)$ and $h_2 = H(d)$. Then we define the sets

$$A = H^{-1}(h_1) \bigcap [[O, p], \qquad B = H^{-1}(h_2) \bigcap [p, \infty]], \quad \text{and} \quad U = \bigcup_{(u,v) \in A \times B} [[u, v]].$$

It is easy to see that A, $B \subset [a, b]$ if $\|p - c\|$ and $\|p - d\|$ are sufficiently small. Clearly, $U \subset [a, b]$ is open.

Since H is an invariant function of system (1), $H(\varphi(t, x)) \equiv$ const for a fixed $x \in P$ and all $t \geq 0$. Thus $H(T^k u) = H(u) = h_1$ and $H(T^k v) = H(v) = h_2$ for any positive integer k. From $p \in E$ and Kamke's theorem, it follows that $T^k u \leq p \leq T^k v$. This proves that $T^k u \in A$ and $T^k v \in B$. Therefore,

$$T^k U = \bigcup_{(u,v) \in A \times B} [[T^k u, T^k v]] \subset U \subset W$$

for each positive integer k. The proof of the lemma is complete.

LEMMA 3.2. *There cannot exist two points p, $q \in E$ such that $p < q$ and $[p, q] \cap E = \{p, q\}$.*

Proof. Suppose the contrary, that is, there exist p, $q \in E$ with $p < q$ such that there is no fixed point between p and q. Lemma 2.1 implies that $Tx < Ty$ for any x, $y \in P$ with $x < y$. Following Dancer and Hess [9], such a map T is strictly order preserving. Applying [9, Prop. 1], we obtain that there is a monotone entire orbit $(x_k)_{k \in \mathbf{Z}}$ in $[p, q]$ connecting p and q, where $\mathbf{Z}$ denotes the set of all integers. Then either $\lim_{k \to -\infty} x_k = p$ and $\lim_{k \to \infty} x_k = q$ or $\lim_{k \to -\infty} x_k = q$ and $\lim_{k \to \infty} x_k = p$. Because H is an invariant function, $H(x_k) = c$ for any $k \in \mathbf{Z}$. The continuity of H implies that $H(p) = H(q) = c$. But from $p < q$ and grad $H(x) \gg O$ for any $x \in P$, it follows that $H(p) < H(q)$, a contradiction. This proves Lemma 3.2.

LEMMA 3.3. *Suppose that $\varphi(t, x)$ is a compact solution of* (1) *and $q \in \omega(x)$ is a stable fixed point of T. Then $\omega(x) = \{q\}$.*

This lemma is adapted from [9, p. 130].

LEMMA 3.4. *Assume that $\varphi(t, x)$ is a compact solution of* (1). *If p and q are the greatest lower bound and the least upper bound of $\omega(x)$, respectively, then p and q are the fixed points of T.*

Proof. Since $\varphi(t, x)$ is a compact solution of (1), there are two positive vectors a, $b \in P$ such that $a \leq \varphi(t, x) \leq b$ for any $t \geq 0$. Hence $a \leq \omega(x) \leq b$. It follows that $a \leq p \leq q \leq b$. Suppose that the conclusion is false. Then, for example, $Tp \neq p$. By the definition of p, $p \leq \omega(x)$. Kamke's theorem and the full invariance of $\omega(x)$ imply that $Tp \leq \omega(x)$. By definition, Tp is also a lower bound for $\omega(x)$, whence $Tp < p$. Because grad $H(x) \gg O$ for each $x \in P$, $H(Tp) < H(p)$. But it follows from (2) that $H(Tp) = H(p)$—a contradiction. This proves the lemma.

LEMMA 3.5. *If $\varphi(t,x)$ is a compact solution of* (1) *and $\omega(x) = \{p\}$, then $\varphi(t,x) - \varphi(t,p) \to O$ as $t \to \infty$.*

This lemma is well known, so its proof is omitted.

Suppose that $\varphi(t,x)$ is a compact solution of (1) and $\omega(x)$ is the ω-limit set of x. Let $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$ be the greatest lower bound and the least upper bound of $\omega(x)$, respectively. For $y = (y_1, y_2, \ldots, y_n)$, we define

$$\sigma_p(y) = \sharp\{i : y_i \neq p_i\}.$$

Here, $\sharp$ denotes the cardinality of the set. Furthermore, we define

$$M(x,p) = \max\{\sigma_p(y) \ : \ y \in \omega(x)\}.$$

From Lemma 2.2, for any $y \in \omega(x)$, $\varphi(t,y)$ is compact. Therefore, we can also define $M(y,u)$, where u is the greatest lower bound of $\omega(y)$.

Proof of Theorem A. Suppose that $\varphi(t,x)$ is a compact solution of (1) and $\omega(x)$ is the ω-limit set of x. In order to prove the theorem, by Lemma 3.5, we have only to prove that $\omega(x)$ is a singleton. Lemma 3.1 tells us that p is Liapunov stable for any $p \in E$. Therefore, by Lemma 3.3, we have only to prove that $\omega(x) \cap E \neq \phi$. Therefore, suppose $\omega(x) \cap E = \phi$, and we shall prove the following two facts:

(I) $M(x,p) \leq n-1$; and

(II) there exists $y \in \omega(x)$ such that $M(y,u) < M(x,p)$, where u is the greatest lower bound of $\omega(y)$.

First, we prove (I). By Lemma 3.4, the greatest lower bound p of $\omega(x)$ is a fixed point of T. Since $\omega(x) \cap E = \phi$, $p \,\overline{\in}\, \omega(x)$, hence $p < \omega(x)$. If (I) is not true, then $M(x,p) = n$, i.e., there is a point $y \in \omega(x)$ such that $\sigma_p(y) = n$, which implies that $p \ll y$. We claim that $[p,y]] \cap E = \{p\}$ must hold. If this equality is false, then there is a fixed point $p_o \in E$ such that $p < p_o \ll y$. Since $y \in \omega(x)$, there is an integer $k_o > 0$ such that $p_o \ll T^{k_o}x$. Kamke's theorem implies that $p_o \ll T^k x$ for $k \geq k_o$. It immediately follows that $p_o \leq \omega(x)$, i.e., p_o is a lower bound of $\omega(x)$, whence $p_o \leq p$, contradicting $p < p_o$. This proves that $[p,y]] \cap E = \{p\}$. Let q denote the least upper bound of $\omega(x)$. Then, by Lemma 3.4, $q \in E$. $\omega(x) \cap E = \phi$ implies that $p < \omega(x) < q$. Since $p \ll y$ and $y \in \omega(x)$, $p \ll q$. Obviously, $[p,q] \cap E$ is a compact set. Since $[p,y]] \cap E = \{p\}$, it is easy to prove that $B := ([p,q] \cap E) \backslash \{p\}$ is also a compact set. By Zorn's lemma, B contains a minimal element p_o. It is not difficult to see that $p < p_o$. The minimality of p_o implies that there exists no fixed point between p and p_o, that is, $[p, p_o] \cap E = \{p, p_o\}$, contradicting Lemma 3.2. (I) is proved.

Second, we prove (II). Let $M(x,p) = m \leq n-1$ and choose $y \in \omega(x)$ such that $\sigma_p(y) = m$. Without loss of generality, we may assume that $y_i \neq p_i$ (hence, $y_i > p_i$) for $i = 1, 2, \ldots, m$. The maximality of m implies that $y_j = p_j$ for $j = m+1, \ldots, n$. For any integer $k > 0$, applying Lemma 2.1, we have $p_i < T_i^k y$ for $i = 1, 2, \ldots, m$. Since $T^k y \in \omega(x)$, the maximality of m implies that $p_j = T_j^k y$ for $j = m+1, \ldots, n$.

Let $\pi^m = \{(x_1, x_2, \ldots, x_m, 0, 0, \ldots, 0) \ : \ x_i \geq 0 \text{ for } i = 1, 2, \ldots, m\}$. Then $O^+(y) \subset p + \pi^m$, which implies that $\omega(y) \subset p + \pi^m$. For an n-dimensional vector $z = (z_1, z_2, \ldots, z_n)$, we define an m-dimensional vector $\tilde{z} = (z_1, z_2, \ldots, z_m)$, and we also use the same symbols $\leq$, $<$, and $\ll$ to denote the order relation in $\mathbf{R}^m$. Suppose u and v are the greatest lower bound and the least upper bound of $\omega(y)$, respectively. Obviously, $u_j = v_j = p_j$ for $j = m+1, \ldots, n$. Therefore, $M(y,u) \leq m$.

We assert that $M(y,u) < m$. If not, $M(y,u) = m$. By the definition of $M(y,u)$, there is a point $z \in \omega(y)$ such that $\sigma_u(z) = m$, i.e., $\tilde{u} \ll \tilde{z}$. We claim that there exists a relative open set V in $[u,v]$ containing u such that $V \cap E = \{u\}$. If the claim is false, then there is a sequence $\{w_k\} \subset E$ such that $u < w_k$ for any $k > 0$ and $\lim_{k\to\infty} w_k = u$. Hence $\lim_{k\to\infty} \tilde{w}_k = \tilde{u} \ll \tilde{z}$ and there exists a $k_0 > 0$ such that $\tilde{w}_k \ll \tilde{z}$ for $k \geq k_0$. Let $w = w_{k_0}$. Then $w_j = u_j = v_j = p_j$ for $j = m+1, \dots, n$. Since $z \in \omega(y)$, there exists an integer $k_1 > 0$ such that $\tilde{w} \ll \widetilde{T^{k_1}y}$. Because $w_j = T_j^k y = p_j$ for $j = m+1, \dots, n, w < T^{k_1}y$. Lemma 2.1 implies that $w < T^k y$ for $k \geq k_1$. Hence $w \leq \omega(y)$, i.e., w is a lower bound of $\omega(y)$, whence $w \leq u$, contradicting $u < w$. This contradiction shows that our claim is true, i.e., $V \cap E = \{u\}$. Therefore, $C := ([u,v] \cap E)\backslash\{u\}$ is a compact set. Zorn's lemma implies that C contains a minimal element u_0. Obviously, $u < u_0$ and $[u, u_0] \cap E = \{u, u_0\}$, contradicting Lemma 3.2. This proves that our assertion holds, i.e., $M(y,u) < M(x,p)$.

For any $y \in \omega(x), \omega(y) \cap E = \phi$. Let $y_0 = y$ and $u_0 = u$. Then, repeatedly using (II) $m+1$ times, we have $\{y_0, y_1, \dots, y_m\} \subset \omega(x)$ such that

$$M(y_i, u_i) < M(y_{i-1}, u_{i-1}) \tag{5}$$

for $i = 1, 2, \dots, m$, where u_i is the greatest lower bound of $\omega(y_i)$. Since $M(y_0, u_0) = M(y,u) < m$, it follows from (5) that $M(y_i, u_i) < m - i$ for $i = 0, 1, 2, \dots, m$. Therefore, $M(y_m, u_m) < 0$. But by the definition of $M(y_m, u_m)$, it is a nonnegative integer—a contradiction. This shows that $\omega(x) \cap E \neq \phi$. The proof is complete.

Remark 1. For infinite-dimensional strongly monotone systems satisfying a much more general conservation law than the one used in this paper, analogous results are proved by Takáč in [7, 8]. All these systems have a common characteristic: every fixed point (equilibrium) is Liapunov stable. The author [12, 15, 16] proves the results of convergence to fixed point without the *strongly* assumption.

Remark 2. J. Mierczyński [5] proved that if (iii) and (iv) hold and $F(t,x)$ satisfies the carathéodory conditions, then the solutions of (1) uniquely exist for the initial value problem on $\mathbf{R} \times P$. Assuming further that (i) holds, we can conclude that every compact solution is asymptotically periodic of period τ under the above hypotheses.

Acknowledgments. The author is indebted to the referees for their valuable suggestions.

REFERENCES

[1] B. R. Tang, Y. Kuang, and H. Smith, *Strictly nonautonomous cooperative system with a first integral*, SIAM J. Math. Anal., 24 (1993), pp. 1331–1339.

[2] F. Nakajima, *Periodic time dependent gross-substitute systems*, SIAM J. Appl. Math., 36 (1979), pp. 421–427.

[3] G. R. Sell and F. Nakajima, *Almost periodic gross-substitute dynamical systems*, Tôhoku Math. J., 32 (1980), pp. 255–263.

[4] J. Mierczyński, *Strictly cooperative systems with a first integral*, SIAM J. Math. Anal., 18 (1987), pp. 642–646.

[5] ———, *Uniqueness for a class of cooperative systems of ordinary differential equations*, Colloq. Math., to appear.

[6] P. Takáč, *Convergence to equilibrium on invariant d-hypersurfaces for strongly increasing discrete-time semigroups*, J. Math. Anal. Appl., 148 (1990), pp. 223–244.

[7] ———, *Domains of attraction of generic ω-limit sets for strongly monotone discrete-time semigroups*, J. Reine Angew. Math., 423 (1992), pp. 101–173.

[8] ———, *Domains of attraction of generic ω-limit sets for strongly monotone semiflows*, Z. Anal. Anwendungen, 10 (1991), pp. 275–317.

[9] E. N. Dancer and P. Hess, *Stability of fixed points for order-preserving discrete-time dynamical systems*, J. Reine Angew. Math., 419 (1991), pp. 125–139.

[10] W. A. Coppel, *Stability and Asymptotic Behavior of Differential Equations*, D. C. Heath, Boston, 1969.

[11] P. Hartman, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[12] Jiang Ji-Fa, *Three- and four-dimensional cooperative systems with every equilibrium stable*, J. Math. Anal. Appl., 188 (1994), pp. 92–100.

[13] ———, *Type K monotone systems with an order-increasing invariant function*, Chinese Ann. Math. Ser. B, 1996, to appear.

[14] ———, *Periodic time dependent cooperative systems of differential equations with a first integral*, Ann. Differential Equations, 8 (1992), pp. 429–437.

[15] ———, *Sublinear discrete-time order-preserving dynamical systems*, Math. Proc. Cambridge Philos. Soc., 119 (1996), pp. 561–574.

[16] ———, *On the analytic order-preserving discrete-time dynamical systems in $\mathbf{R}^n$ with every fixed point stable*, J. London Math. Soc., 1996, to appear.

[17] F. Nikaido, *Convex Structure and Economic Theory*, Academic Press, New York, 1968.

[18] P. A. Samuelson, *Foundations of Economic Analysis*, Harvard University Press, Cambridge, MA, 1948.

[19] Q. Arino, *Monotone semi-flows which have a monotone first integral*, preprint.

SIAM J. MATH. ANAL.
Vol. 27, No. 6, pp. 1745–1769, November 1996

NONSTATIONARY SUBDIVISION SCHEMES AND MULTIRESOLUTION ANALYSIS*

ALBERT COHEN† AND NIRA DYN‡

Abstract. Nonstationary subdivision schemes consist of recursive refinements of an initial sparse sequence with the use of masks that may vary from one scale to the next finer one. This paper is concerned with both the convergence of nonstationary subdivision schemes and the properties of their limit functions. We first establish a general result on the convergence of such schemes to C^∞ compactly supported functions. We show that these limit functions allow us to define a multiresolution analysis that has the property of spectral approximation. Finally, we use these general results to construct C^∞ compactly supported cardinal interpolants and also C^∞ compactly supported orthonormal wavelet bases that constitute Riesz bases for Sobolev spaces of any order.

Key words. subdivision schemes, multiresolution analysis, spectral approximation, dyadic interpolation, wavelets.

AMS subject classifications. 41A28, 41A30, 41A10, 42C15

1. Introduction. Subdivision schemes constitute a useful tool for the fast generation of smooth curves and surfaces from a set of control points by means of iterative refinements. In the most-often-considered binary univariate case, one starts from a sequence $s_0(k)$ and obtains at step j a sequence $s_j(2^{-j}k)$ generated from the previous one by linear rules:

$$s_j(2^{-j}k) = 2 \sum_{n\in k+2\mathbb{Z}} c_{j,k}(n) s_{j-1}(2^{-j}(k-n)). \tag{1.1}$$

The masks $c_{j,k} = \{c_{j,k}(n)\}_{n\in\mathbb{Z}}$ are, in general, finite sequences, a property that is clearly useful for the practical implementation of (1.1).

A natural problem is then to study the convergence of such an algorithm to a limit function. In particular, the scheme is said to be strongly convergent if and only if there exists a continuous function $f(x)$ such that $\lim_{j\to+\infty}(\sup_k |s_j(2^{-j}k) - f(2^{-j}k)|) = 0$. One can study more general types of convergence with the use of a smooth function g that is well localized in space (for example, compactly supported) and satisfies the interpolation property $g(k) = \delta_k$. One can then define $f_j(x) = \sum_k s_j(2^{-j}k) g(2^j x - k)$ and study the convergence of f_j to f in a functional sense.

A subdivision scheme is said to be stationary and uniform when the masks $c_{j,k}(n) = c_n$ are independent of the parameters j and k. In that case, one can rewrite (1.1) as

$$s_j(2^{-j}k) = 2\sum_n c_{k-2n} s_{j-1}(2^{-j+1}n). \tag{1.2}$$

Note that (1.2) is equivalent to filling in the sequence s_{j-1} with zeros at the intermediate points $2^{-j}(2k+1)$ and applying a discrete convolution with the sequence (c_k). Detailed reviews of stationary subdivision have been done by Cavaretta, Dahmen, and Micchelli (1991) and Dyn (1992).

* Received by the editors July 5, 1994; accepted for publication (in revised form) June 5, 1995.

† Laboratoire d'Analyse Numerique, Université Pierre et Marie Curie, 4 Place Jussieu, 75005 Paris, France.

‡ Department of Mathematics, Tel Aviv University, Tel Aviv 69978, Israel.

These algorithms apply in a natural way to computer-aided geometric design. Moreover, the interest in stationary subdivision schemes has grown in the digital-image-processing and numerical-analysis communities since they have been connected to multiresolution analysis and wavelet bases.

A multiresolution analysis consists of a nested sequence of approximation subspaces

$$\{0\} \to \cdots V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \cdots \to L^2(\mathbb{R}) \tag{1.3}$$

that are generated by a "scaling function" $\varphi \in V_0$ in the sense that the set $\{\varphi(2^j x - k)\}_{k\in\mathbb{Z}}$ constitutes a Riesz basis for V_j. By $V_j \to L^2(\mathbb{R})$, we mean here that for any f in $L^2(\mathbb{R})$, $\lim_{j\to+\infty} \|P_j f - f\|_0 = 0$, where $P_j f$ is the L^2-projection of f onto V_j and $\|\cdot\|_0$ is the L^2 norm (we shall use the notation $\|\cdot\|_s$ for the Sobolev $H^s = W_2^s$ norm). Here again, many generalizations are possible (see Meyer (1990) or Daubechies (1992) for a detailed review of this concept).

Since the spaces V_j are embedded, the scaling function satisfies an equation of the type

$$\varphi(x) = 2\sum_n c_n \varphi(2x - n). \tag{1.4}$$

We shall assume here that φ is compactly supported so that the c_n's are finite in number. In that case, φ is also an L^1 function and by taking the Fourier transform of (1.4), we have

$$\hat{\varphi}(\omega) = m(\omega/2)\hat{\varphi}(\omega/2), \tag{1.5}$$

where $m(\omega) = \sum_{n\in\mathbb{Z}} c_n e^{-in\omega}$. Assuming that φ is normalized in the sense that $\int \varphi = \hat{\varphi}(0) = 1$, by iterating (1.5), we obtain

$$\hat{\varphi}(\omega) = \prod_{k=1}^{+\infty} m(2^{-k}\omega). \tag{1.6}$$

This formula indicates that φ is the limit in the weak (or distribution) sense of a stationary subdivision scheme since it represents in the Fourier domain the refinement of an initial Dirac sequence by iterative convolutions with c_n. Note also that the support of φ is contained in the convex hull of the support of the mask (c_k). Conversly, any refinable function, i.e., weak limit of such a scheme, satisfies a "refinement equation" of the type described above and is a potential candidate to generate a multiresolution analysis (see also Derfel, Dyn, and Levin (1995)).

Given a stationary subdivision scheme, we see here that two questions are relevant:

- Is the scheme convergent and in what sense?
- What are the properties of the limit functions?

By the last question, we mean in particular the approximation properties of the spaces V_j (can we approximate in norms other than L^2, in particular in Sobolev spaces H^s, with specific rates...), the exact regularity of the scaling function, and other properties of φ such as cardinal interpolation or orthonormality of its integer shifts.

Numerous contributions have been made to these two problems. The convergence of the subdivision and the approximation properties of the multiresolution spaces are strongly linked: in particular, one can prove (see Dyn and Levin (1990); Cavaretta, Dahmen, and Micchelli (1992); and Daubechies and Lagarias (1991)) that both the convergence of the subdivision scheme to a C^r function for some $r \geq 0$ and the property that $\lim_{j \to +\infty} 2^{js} \|P_j f - f\|_0 = 0$ for all $f \in H^s$ $(s \leq r)$ imply that the scaling function satisfies the Strang–Fix conditions of order N, where N is an integer such that $N \leq r < N+1$. These conditions can be expressed by three equivalent statements:

- Any polynomial of degree not exceedding N can be expressed as a combination of the integer shifts of φ.
- For all $p \leq N$ and $n \in \mathbb{Z} - \{0\}$, $(\frac{d}{d\omega})^p \hat{\varphi}(2n\pi) = 0$, $\hat{\varphi}(0) = 1$.
- For all $p \leq N$, $(\frac{d}{d\omega})^p m(\pi) = 0$ or, equivalently, $\sum_n (-1)^n n^p c_n = 0$.

Note that this last statement reveals that $m(\omega)$ can be written as

$$m(\omega) = \left(\frac{1+e^{-i\omega}}{2}\right)^{N+1} q(\omega), \tag{1.7}$$

where $q(\omega)$ is a trigonometric polynomial. (1.7) implies that there are at least $N+2$ nonzero c_n's, and thus the support length of φ is at least $N+1$. This leads to the observation that very good approximation rates for regular functions, as well as convergence of the subdivision in a smooth norm, can only be achieved if one accepts the loss of some space localization (in particular, one cannot build a refinable function that is both compactly supported and in C^∞).

More recently, attention has been given to subdivision schemes that are nonstationary in scale, i.e., for which the masks may vary from one step of the refinement process to the next. A model case is the scheme that uses at step k the same mask $c_n^k = \binom{k}{n} 2^{-k+1}$ $(0 \leq n \leq k)$ that would give rise in the stationary subdivision case to B-splines of degree $k-1$. It was proved by Derfel, Dyn, and Levin (1995) that such a scheme converges strongly to the "up-function" introduced by Rvachev and Rvachev (1971) (see also Rvachev (1990)). The limit function can thus be written in the Fourier domain as

$$\hat{\varphi}(\omega) = \prod_{k=1}^{+\infty} \left(\frac{1+e^{-i2^{-k}\omega}}{2}\right)^k. \tag{1.8}$$

The length of its support is given by $L = \sum_{k>0} k2^{-k} = 2 < +\infty$. Such a function cannot satisfy a refinement equation of the type of (1.4). However, note that the product (1.8) can also be written as

$$\begin{aligned}\prod_{k=1}^{+\infty} \left(\frac{1+e^{-i2^{-k}\omega}}{2}\right)^k &= \prod_{n=0}^{+\infty} \prod_{k=n+1}^{+\infty} \frac{1+e^{-i2^{-k}\omega}}{2} \\ &= \prod_{n=0}^{+\infty} \frac{1+e^{-i2^{-n}\omega}}{-i2^{-n}\omega} \\ &= \prod_{n=0}^{+\infty} \hat{\chi}_{[0,1]}(2^{-n}\omega).\end{aligned}$$

It follows that

$$\varphi = \chi_{[0,1]} * 2\chi_{[0,1/2]} * \cdots * 2^j \chi_{[0,2^{-j}]} * \cdots$$

is a C^∞ function that satisfies a "continuous-refinement equation" of the type

$$\varphi(x) = 2\chi_{[0,1]} * \varphi(2\cdot) = \int_0^2 \varphi(2x-y)dy. \tag{1.9}$$

The idea of using iterative convolutions to build C^∞ compactly supported test functions is older than Rvachev's work: it can be found in Mandelbrojt (1942).

Returning to subdivision schemes, we see that by letting the masks grow linearly, it is possible to obtain a C^∞ function while preserving the compact-support property. It was also shown by Dyn and Ron (1995) that a "semimultiresolution analysis" can be derived by defining, for all $j \geq 0$, $V_j = \text{Span}\{\varphi_j(2^j x - k)\}_{k\in\mathbb{Z}}$ with

$$\hat{\varphi}_j(\omega) = \prod_{k=1}^{+\infty} \left(\frac{1+e^{-i2^{-k}\omega}}{2}\right)^{k+j} \tag{1.10}$$

and that these spaces have the property of spectral approximation in L^2: for any $r \geq 0$ and for all $f \in H^r$, $\lim_{j\to+\infty} 2^{jr}\|P_j f - f\|_0 = 0$.

Our goal in this paper is to generalize these results to a large class of nonstationary subdivision schemes.

Assuming that such a scheme converges at least in the sense of tempered distributions, the general form of its limit function will be given in the Fourier domain by

$$\hat{\varphi}(\omega) = \prod_{k=1}^{+\infty} m_k(2^{-k}\omega), \tag{1.11}$$

where m_k is the sequence of trigonometric polynomials associated with the masks of the subdivision. Note that, since we do not assume any particular form for m_k, the function φ will not in general satisfy any type of refinement equation, discrete or continuous, thus making the analysis of its smoothness and approximation properties more difficult.

What is the interest of such a generalization? It is important to remark that the approximation properties of the up-function and its associated multiresolution analysis, very attractive from the theoretical point of view, suffer from a major numerical disadvantage: the computation of the L^2 projection onto V_j is difficult to manage at high scales since the Gram matrix of the basis $\{\varphi^j(2^j x - k)\}_{k\in\mathbb{Z}}$ becomes ill conditioned. More precisely, its condition number $C(j)$ grows exponentially with j:

$$(\sqrt{2})^{j+1} \leq C(j) \leq \text{const}\left(\frac{\pi}{2}\right)^j. \tag{1.12}$$

The upper bound is taken from Dyn and Ron (1995) and the lower bound is obtained here:

$$\begin{aligned} C(j)^2 &= \left(\sup_\omega \sum_k |\hat{\varphi}^j(\omega+2k\pi)|^2\right)\left(\inf_\omega \sum_k |\hat{\varphi}^j(\omega+2k\pi)|^2\right)^{-1} \\ &\geq \left(\inf_\omega \sum_k |\hat{\varphi}^j(\omega+2k\pi)|^2\right)^{-1} \end{aligned}$$

$$\geq \left(\sum_k |\hat{\varphi}^j((2k+1)\pi)|^2\right)^{-1}$$
$$= \left(\sum_k |\cos^{j+1}(\pi/4)\hat{\varphi}^{j+1}((2k+1)\pi/2)|^2\right)^{-1}$$
$$\geq [\cos(\pi/4)]^{-2j-2} = 2^{j+1}.$$

The same problem occurs when one wants to interpolate data on the grid $2^{-j}\mathbb{Z}$ by a function in V_j for j odd: one checks from a similar computation that the condition number $D(j)$ of the system grows exponentially.

In a more general setting, it is possible to keep these condition numbers bounded as j grows. One can even fix one of them to 1 by imposing constraints on the trigonometric polynomials m_k so that the limit functions have orthonormality or cardinal interpolation properties (see §4).

Finally, an important property of multiresolution analysis is the equivalence

$$\|f\|_r^2 \approx \|P_0 f\|_0^2 + \sum_{j>0} 2^{2jr}\|P_{j+1}f - P_j f\|_0^2 \tag{1.13}$$

that is the key to multilevel preconditionning techniques (see Dahmen and Kunoth (1992)) and that can also be expressed in terms of wavelet coefficients. Thus far, we have only been able to prove this equivalence in the orthonormal case for all $r > 0$ (see §4).

Our paper is organized as follows. In §2, we give a general result on the convergence of a nonstationary subdivision scheme in C^∞ under very mild conditions on the masks. In §3, we study the approximation properties of the associated multiresolution spaces and prove that spectral approximation can be achieved for all Sobolev norms. Finally, in §4, we apply these results to dyadic interpolation and to orthonormal wavelets that constitute Riesz bases for all Sobolev spaces. This particular wavelet basis has recently been introduced in a paper by Berkolaiko and Novikov (1992) which was concerned with the existence of a multiscale orthonomal basis of compactly supported C^∞ functions.

For the sake of simplicity, we limit ourselves to the one-dimensional setting and our results are stated in the case where the length of the masks grows at least linearly. We show in the appendix how this can be extended to more general growth rates of the mask length.

2. Nonstationary subdivision schemes. Let $\{m_k\}_{k>0}$ be a sequence of finite masks, i.e., $m_k(n) = 0$ if $|n| > d(k)$. We denote by $m_k(\omega) = \sum_n m_k(n)e^{-in\omega}$ their representation in the Fourier domain, i.e., a sequence of trigonometric polynomials of degree $d(k)$. Let us consider the nonstationary subdivision scheme that is associated with this sequence of masks, i.e., $s_j(2^{-j}k) = 2\sum_n m_j(k-2n)s_{j-1}(2^{-j+1}n)$. If the input is a Dirac sequence $\delta_{m,0}$, one obtains after n steps a sequence of samples on the grid $2^{-n}\mathbb{Z}$ that can be interpolated in a unique way by a function $\varphi^{[n]}$ that is band limited on $[-2^n\pi, 2^n\pi]$. This function is defined by

$$\hat{\varphi}^{[n]}(\omega) = \prod_{k=1}^{n} m_k(2^{-k}\omega)\chi_{[-\pi,\pi]}(2^{-n}\omega). \tag{2.1}$$

Note that the functions $\varphi^{[n]}$ are analytic and thus not compactly supported. We shall use these particular interpolants in order to study the convergence of the subdivision scheme to the limit function defined (if this is possible) by

$$\hat{\varphi}(\omega) = \prod_{k=1}^{+\infty} m_k(2^{-k}\omega). \tag{2.2}$$

After n steps, the result of the subdivision in the space domain is supported in $[-L(n), L(n)]$ with $L(n) = \sum_{k=1}^{n} 2^{-k}d(k)$. A natural condition for the compactly supported limit function is thus

$$L = \sum_{k=1}^{+\infty} 2^{-k}d(k) < +\infty. \tag{2.3}$$

Our first result shows that this condition is also instrumental in the derivation of the convergence in the sense of tempered distributions of the subdivision scheme.

THEOREM 2.1. *Assume that $r_k = 2^{-k}d(k)$ and $s_k = |m_k(0) - 1|$ are both summable sequences and that the functions $|m_k(\omega)|$ are uniformly bounded by some constant $M > 0$. Then $\hat{\varphi}^{[n]}$ converges uniformly on any compact set to $\hat{\varphi}$ and $\varphi^{[n]}$ converges to φ in the sense of tempered distributions. The tempered distribution φ is compactly supported in $[-L, L]$ with $L = \sum_{k>0} r_k$.*

Proof. We first study the convergence of the infinite product (2.2). For a fixed ω, we have to check the summability in k of $t_k(\omega) = |m_k(2^{-k}\omega) - 1|$. If, in addition, $\sum_{k>0} t_k(\omega)$ is uniformly bounded on every compact set, then (2.2) will also converge uniformly on every compact set.

We can write

$$\begin{aligned} t_k(\omega) &\le |m_k(2^{-k}\omega) - m_k(0)| + s_k \\ &\le 2^{-k}|\omega| \sup_{\omega} \left| \frac{d}{d\omega} m_k \right| + s_k. \end{aligned}$$

Using Bernstein's inequality, we obtain the estimate

$$\sup_{\omega} \left| \frac{d}{d\omega} m_k \right| \le Md(k), \tag{2.4}$$

and thus

$$t_k(\omega) \le M|\omega| r_k + s_k, \tag{2.5}$$

which proves the uniform convergence of (2.2) on every compact set.

The same argument shows that for any $n > p \ge 0$, the products

$$P_p^n(\omega) = \prod_{k=p+1}^{n} m_k(2^{-k}\omega), \tag{2.6}$$

are uniformly bounded on $[-2^{p+1}, 2^{p+1}]$ by the same $B > 0$. We can define these products to be equal to 1 whenever $n \le p$ so that this statement makes sense for all $n, p > 0$. This applies in particular to $\hat{\varphi}^{[n]} = P_0^n$ and $\hat{\varphi} = P_0^{\infty}$, which are thus uniformly bounded on $[-2, 2]$.

For $2^p \le |\omega| \le 2^{p+1}$ with $p \ge 0$, we can write

$$\begin{aligned} |\hat{\varphi}^{[n]}(\omega)| &= \left| P_p^n(\omega) \prod_{k=1}^{p} m_k(2^{-k}\omega) \right| \\ &\le BM^p \le BM^{\log_2 |\omega|} \le B|\omega|^b \end{aligned}$$

with $b = \log_2(M)$ (we have assumed here without loss of generality that $M \geq 1$). For all $\omega \in \mathbb{R}$, we thus have the estimate

$$|\hat{\varphi}^{[n]}(\omega)| \leq B(1+|\omega|)^b, \tag{2.7}$$

where the constant B does not depend on n. Consequently, it also holds for the pointwise limit $\hat{\varphi}$.

Now take any test function $g(\omega)$ in the Schwartz class $S(\mathbb{R})$. For any $\varepsilon > 0$, there exists $A > 0$ such that

$$B\int_{|\omega|>A} g(\omega)(1+|\omega|)^b d\omega \ < \ \varepsilon/2. \tag{2.8}$$

By the uniform convergence of $\hat{\varphi}^{[n]}$ to $\hat{\varphi}$ on every compact, there exists N such that for all $n > N$,

$$\left|\int_{|\omega|<A} g(\omega)(\hat{\varphi}(\omega) - \hat{\varphi}^{[n]}(\omega))d\omega\right| \ < \ \varepsilon/2. \tag{2.9}$$

Combining (2.7), (2.8), and (2.9), we immediately obtain the convergence of $\langle\hat{\varphi}^{[n]}|g\rangle$ to $\langle\hat{\varphi}|g\rangle$. □

We are now interested in finding additional hypotheses for stronger convergence of the subdivision scheme to a C^∞ compactly supported function φ. Note that, in contrast to its approximants $\varphi^{[n]}$, the function φ cannot be analytic. Our next result states general conditions for the uniform convergence of $\varphi^{[n]}$ and all its derivatives.

THEOREM 2.2. *Assume that the hypotheses of Theorem* 2.1 *are satisfied and that we have the estimate*

$$|m_k(\omega)| \leq (1+\alpha_k)|m(\omega)|^k \tag{2.10}$$

with $\sum_k |\alpha_k| < +\infty$ *and* $m(\omega) = \cos^\beta(\omega/2)\tilde{m}(\omega)$, *for some* $\beta \geq 0$ *(not necessarily integer), where the function* $\tilde{m}(\omega)$ *is bounded, Hölder continuous at the origin and satisfies* $\tilde{m}(0) = 1$ *and* $\sigma_i = \sup_\omega |\prod_{k=1}^{i} \tilde{m}(2^k\omega)| < 2^{\beta i}$ *for some fixed integer* $i > 0$.

Then φ *is a* C^∞ *compactly supported function and, for all* $s \in \mathbb{Z}_+$, $(\frac{d}{dx})^s\varphi^{[n]}$ *converges uniformly to* $(\frac{d}{dx})^s\varphi$.

Proof. It is sufficient to show that for all $s \in \mathbb{Z}_+$, the functions $|\omega|^s|\hat{\varphi}^{[n]}(\omega)|$ are dominated by an L^1 function $f_s(\omega)$ that does not depend on n: by dominated convergence, this implies

$$\lim_{n\to+\infty}\int |\omega|^s|\hat{\varphi}(\omega) - \hat{\varphi}^{[n]}(\omega)|d\omega \ = \ 0 \tag{2.11}$$

and thus the uniform convergence of all the derivatives of $\varphi^{[n]}$ in the space domain. We shall construct these dominating functions using the additional hypotheses that we have made on the functions $m_k(\omega)$. First, we need a technical estimate that will be useful: for any $q \geq 0$, there exists $C_q > 0$ such that for any sequence $\{a_k\}_{k>0}$ with $0 \leq a_k \leq 1$ and any $n \geq p \geq 0$,

$$\prod_{k=p}^{n} |m_{q+k}(2^{-k}\omega)|^{a_k} \leq C_q(1+|\omega|)^b \tag{2.12}$$

with $b = \log_2(M)$. (As in the previous theorem, we assume without loss of generality that $M \geq 1$.) Indeed, using the same argument (Bernstein's inequality) as in the proof of Theorem 2.1, we observe that $\prod_{k=p}^{n} |m_{q+k}(2^{-k}\omega)|^{a_k}$ is uniformly bounded in $[-1, 1]$ by a constant C_q that does not depend on a_k, p, and n since we have

$$\begin{aligned} |1 - |m_{q+k}(2^{-k}\omega)|^{a_k}| &\leq |m_{q+k}(2^{-k}\omega) - 1| \\ &\leq |m_{q+k}(2^{-k}\omega) - m_{q+k}(0)| + |m_{q+k}(0) - 1| \\ &\leq 2^q M|\omega| r_{q+k} + s_{q+k} \leq 2^q M r_{q+k} + s_{q+k}. \end{aligned}$$

For $2^l \leq |\omega| \leq 2^{l+1}$ with $p \leq l < n$, we now derive

$$\begin{aligned} \prod_{k=p}^{n} |m_{q+k}(2^{-k}\omega)|^{a_k} &= \prod_{k=p}^{l} |m_{q+k}(2^{-k}\omega)|^{a_k} \prod_{k=l+1}^{n} |m_{q+k}(2^{-k}\omega)|^{a_k} \\ &\leq M^l \prod_{k=l+1}^{n} |m_{q+k}(2^{-k}\omega)|^{a_k} \\ &\leq C_q (M)^{\log_2 |\omega|} = C_q |\omega|^b. \end{aligned}$$

In the cases where $l \geq n$, this estimate still holds since $M^n \leq M^l$, while for $l < p$, the bound is C_q. This proves (2.12) for all $\omega \in \mathbb{R}$.

We are now ready to build the dominating functions $f_s(\omega)$. For fixed $s \geq 0$, choose $p \in \mathbb{N}$ such that $p((\log_2 \sigma_i)/i - \beta) + s + b < -1$. (This is always possible since we have assumed $(\log_2 \sigma_i)/i < \beta$.) For $n \geq p$, we can estimate $\hat{\varphi}^{[n]}(\omega)$ on $[-2^n\pi, 2^n\pi]$ by

$$\begin{aligned} |\hat{\varphi}^{[n]}(\omega)| &= \prod_{k=1}^{n} |m_k(2^{-k}\omega)| \\ &\leq M^{p-1} \prod_{k=p}^{n} |m_k(2^{-k}\omega)| \\ &= M^{p-1} \prod_{k=p}^{n} |m_k(2^{-k}\omega)|^{\frac{p}{k}} \prod_{k=p}^{n} |m_k(2^{-k}\omega)|^{\frac{k-p}{k}}. \end{aligned}$$

Using estimate (2.12) and hypothesis (2.10), we thus obtain

$$\begin{aligned} |\hat{\varphi}^{[n]}(\omega)| &\leq M^{p-1} C_p (1 + |\omega|)^b \prod_{k=p}^{n} |m(2^{-k}\omega)|^p \\ &= M^{p-1} C_p (1 + |\omega|)^b \prod_{k=p}^{n} |\cos(2^{-k-1}\omega)|^{\beta p} |\tilde{m}(2^{-k}\omega)|^p \\ &= M^{p-1} C_p (1 + |\omega|)^b \frac{|\mathrm{sinc}(2^{-p}\omega)|^{\beta p}}{|\mathrm{sinc}(2^{-n-1}\omega)|^{\beta p}} \prod_{k=p}^{n} |\tilde{m}(2^{-k}\omega)|^p \\ &\leq A_p (1 + |\omega|)^{b-\beta p} \prod_{k=p}^{n} |\tilde{m}(2^{-k}\omega)|^p, \end{aligned}$$

where A_p depends only on p, since $|\mathrm{sinc}(2^{-n-1}\omega)|^{\beta p}$ is bounded below away from 0 on $[-2^n\pi, 2^n\pi]$ by a constant that does not depend on n but only on p. To estimate the remaining product, we remark that since $\tilde{m}(\omega)$ is bounded and Hölder continuous at the origin and $\tilde{m}(0) = 1$, then for all $n \geq p \geq 0$, the products $\prod_{k=p}^{n} |\tilde{m}(2^{-k}\omega)|^p$ are uniformly bounded on $[-1, 1]$ by a constant B_p that is independent of n. For $2^l \leq |\omega| \leq 2^{l+1}$ with $p \leq l < n$, using the hypothesis on $\tilde{m}$, we obtain

$$\begin{aligned}
\prod_{k=p}^{n} |\tilde{m}(2^{-k}\omega)|^p &= \prod_{k=p}^{l} |\tilde{m}(2^{-k}\omega)|^p \prod_{k=l+1}^{n} |\tilde{m}(2^{-k}\omega)|^p \\
&= \prod_{k=p}^{l} |\tilde{m}(2^{-k}\omega)|^p \prod_{k=1}^{n-l} |\tilde{m}(2^{-k-l}\omega)|^p \\
&\leq B_p \prod_{k=p}^{l} |\tilde{m}(2^{-k}\omega)|^p \\
&\leq B_p \sigma_i^{[\frac{l-p}{i}]p} \left(\sup_{\omega} |\tilde{m}(\omega)|\right)^{(i-1)p} \\
&= D_p \sigma_i^{lp/i} \\
&\leq D_p \sigma_i^{(p/i)\log_2 |\omega|} = A_p |\omega|^{(p/i)\log_2 \sigma_i},
\end{aligned}$$

where D_p depends only on p (again, in the cases where $l \geq n$ or $l < p$, this still holds by replacing the product that does not make sense by 1). Combining this with the previous estimate, we obtain

$$|\hat{\varphi}^{[n]}(\omega)| \leq K_p(1 + |\omega|)^{b+p(\frac{\log_2 \sigma_i}{i} - \beta)}, \tag{2.13}$$

where K_p depends only on p, and thus

$$|\omega|^s |\hat{\varphi}^{[n]}(\omega)| \leq K_p(1 + |\omega|)^{b+s+p(\frac{\log_2 \sigma_i}{i} - \beta)}. \tag{2.14}$$

This also holds trivially for $|\omega| > 2^n\pi$. Since we have assumed that $b+s+p((\log_2 \sigma_i)/i - \beta) < -1$, this gives us the desired uniform L^1 estimate. This concludes the proof of the theorem. □

Remarks. The hypotheses of Theorem 2.2 imply, in particular, that the degree $d(k)$ of m_k grows at least linearly (m_k has a zero of order βk at $\omega = \pi$). This is not strictly necessary: we show in the appendix that it is possible to obtain strongly converging subdivision schemes with a C^∞ limit function as soon as $d(k)$ tends to $+\infty$ without any assumption on its asymptotic behavior (but with the assumption $|m(\omega)| \leq 1$, which removes a lot of technicalities).

These hypotheses can also be weakened by assuming that estimate (2.10) is satisfied only for k sufficiently large: the limit behavior of the subdivision does not depend on the first iterations.

3. Multiresolution approximation. Let $\{m_k\}_{k>0}$ be a sequence of finite masks that satisfy the hypotheses of Theorem 2.2. We define a sequence of C^∞ compactly supported functions by

$$\hat{\varphi}_j(\omega) = \prod_{k=1}^{+\infty} m_{k+j}(2^{-k}\omega), \quad j \geq 0. \tag{3.1}$$

We see that $\varphi_0 = \varphi$ and that φ_j is obtained as the limit of the same subdivision algorithm by cancelling the first j iterations. It follows that φ_j is also in C_0^∞. Since

$\hat{\varphi}_j(\omega) = m_{j+1}(\omega/2)\hat{\varphi}_{j+1}(\omega/2)$, we see that this sequence of functions satisfies a series of recursive refinement equations:

$$\varphi_j(x) = \sum_{|n|\le d(j+1)} m_{j+1}(n)\varphi_{j+1}(2x-n). \tag{3.2}$$

It is thus natural to define a "semimultiresolution analysis" $\{V_j\}_{j\ge 0}$ by $V_j = \mathrm{Span}\{\varphi_j(2^j x - k)\}_{k\in\mathbb{Z}}$. The inclusion $V_j \subset V_{j+1}$ comes from (3.2).

We shall now study the approximation properties of these spaces in Sobolev spaces. Given a function $f \in H^r$, we can define

$$d(f, V_j)_s = \inf_{g\in V_j} \|f-g\|_s \tag{3.3}$$

for $s \le r$, where $\|\cdot\|_s$ is the H^s norm. We are concerned here with the behavior of $d(f, V_j)$ as j goes to $+\infty$. By definition, the spaces V_j have approximation order (resp. density order) r in H^s if $2^{(r-s)j}d(f,V_j)_s$ is bounded (resp. goes to 0) as $j \to +\infty$.

We shall first establish a general result using a technique introduced in a paper by de Boor, DeVore, and Ron (1992) in which the authors are concerned with approximation in the L^2 norm from shift-invariant spaces. Here we adapt their technique to the derivation of density orders in Sobolev norms. Approximation orders in Sobolev norms by shift-invariant spaces are studied in Zao (1995) and Ron (1995).

THEOREM 3.1. *Let $\{\varphi^j\}_{j\ge 0}$ be a sequence of compactly supported functions in H^s for some $s \ge 0$ and define $V_j = \mathrm{Span}\{\varphi_j(2^j x - k)\}_{k\in\mathbb{Z}}$. For $r \ge s$, assume that there exists $t \in]0,\pi]$ such that, for all $0 \le v \le s$,*

$$\sup_{|\omega|<t}\left(|\omega|^{-2r}\frac{\sum_{n\ne 0}|\omega+2n\pi|^{2v}|\hat{\varphi}_j(\omega+2n\pi)|^2}{|\hat{\varphi}_j(\omega)|^2}\right) \to 0 \quad as\ j\to+\infty. \tag{3.4}$$

Then the spaces V_j have density order r in H^s: let P_j be the L^2 projection onto V_j and let S_j be the operator defined by $\mathcal{F}S_jf(\omega) = \hat{f}(\omega)\chi_{[-t,t]}(2^{-j}\omega)$, where $\mathcal{F}$ represents the Fourier-transform operator ($\mathcal{F}f(\omega) = \hat{f}(\omega)$). Then for all $f \in H^r$, we have $d(f,V_j) \le \|P_jS_jf - f\|_s \le C2^{j(s-r)}\|f\|_r\varepsilon(f,j)$ with $0 \le \varepsilon(f,j) \le 1$ and $\lim_{j\to+\infty}\varepsilon(f,j) = 0$.

Proof. First, observe that we can always associate with φ_j a function ϕ^j defined by

$$\hat{\phi}_j(\omega) = \frac{\hat{\varphi}_j(\omega)}{\left(\sum_{n\in\mathbb{Z}}|\hat{\varphi}_j(\omega+2n\pi)|^2\right)^{1/2}} \tag{3.5}$$

such that $\{2^{j/2}\phi_j(2^jx-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis of V_j: in the case where $\sum_{n\in\mathbb{Z}}|\hat{\varphi}_j(\omega+2n\pi)|^2 = \sum_k\langle\varphi^j(\cdot)|\varphi^j(\cdot-k)\rangle e^{-ik\omega}$ vanishes at some isolated point, we can easily check that ϕ_j is still the L^2 limit when $\varepsilon \to 0$ of $\phi_{j,\varepsilon}$ defined by

$$\hat{\phi}_{j,\varepsilon}(\omega) = \frac{\hat{\varphi}_j(\omega)}{\left(\varepsilon + \sum_{n\in\mathbb{Z}}|\hat{\varphi}_j(\omega+2n\pi)|^2\right)^{1/2}} \tag{3.6}$$

and that $\phi_{j,\varepsilon}$ is an ℓ^2 combination of $\varphi_j(x-k)$.

Consequently, we can write

$$P_jf(x) = 2^j \sum_{k\in\mathbb{Z}} \langle f|\phi_j(2^j\cdot -k)\rangle \phi_j(2^jx-k) \tag{3.7}$$

for any $f \in L^2$. For all $j \geq 0$, we define $Q_j = I - P_j$ and $T_j = I - S_j$. We can thus estimate the approximation error as follows:

$$\begin{aligned}\|P_jS_jf - f\|_s &\leq \|T_jf\|_s + \|P_jS_jf - S_jf\|_s\\ &\leq \|T_jf\|_s + \|S_jP_jS_jf - S_jf\|_s + \|T_jP_jS_jf\|_s\\ &\leq \|T_jf\|_s + \|S_jQ_jS_jf\|_s + \|T_jP_jS_jf\|_s.\end{aligned}$$

Let f be in H^r, i.e., $\|f\|_r^2 = (2\pi)^{-1}\int|\hat{f}(\omega)|^2(1+|\omega|^{2r})d\omega < +\infty$. We shall examine these three quantities separately and prove that they all satisfy the estimate that we want for $d(V_j,f)_s$.

The "truncation error" $\|T_jf\|_s$ is independent of the approximating subspaces V_j. It is clear that we have

$$\begin{aligned}\|T_jf\|_s^2 &= (2\pi)^{-1}\int_{|\omega|>2^jt}|\hat{f}(\omega)|^2(1+|\omega|^{2s})d\omega\\ &\leq (2\pi)^{-1}2^{2j(s-r)}t^{2(s-r)}\int_{|\omega|>2^jt}|\hat{f}(\omega)|^2(1+|\omega|^{2r})d\omega\\ &\leq C2^{2j(s-r)}\|f\|_r^2\varepsilon(f,j),\end{aligned}$$

with $0 \leq \varepsilon(f,j) \leq 1$ and $\varepsilon(f,j) \to 0$ as $j \to +\infty$.

For the second term, we have

$$\begin{aligned}\|S_jQ_jS_jf\|_s^2 &= (2\pi)^{-1}\int_{|\omega|<2^jt}|\mathcal{F}Q_jS_jf(\omega)|^2(1+|\omega|^{2s})d\omega\\ &\leq (2\pi)^{-1}(1+2^{2js}t^{2s})\int_{|\omega|<2^jt}|\mathcal{F}Q_jS_jf(\omega)|^2d\omega\\ &\leq C2^{2js}\|S_jQ_jS_jf\|_0^2.\end{aligned}$$

To estimate $\|S_jQ_jS_jf\|_0^2$, we note that

$$\begin{aligned}\mathcal{F}P_jS_jf(\omega) &= \hat{\phi}_j(2^{-j}\omega)\sum_{k\in\mathbb{Z}}\langle S_jf|\phi_j(2^j\cdot -k)\rangle e^{-i2^{-j}k\omega}\\ &= (2^{j+1}\pi)^{-1}\hat{\phi}_j(2^{-j}\omega)\sum_{k\in\mathbb{Z}}\langle \mathcal{F}S_jf(\cdot)|\hat{\phi}_j(2^{-j}\cdot)e^{i2^{-j}k\cdot}\rangle e^{-i2^{-j}k\omega}.\end{aligned}\tag{3.8}$$

Since the above sum defines a $2^{j+1}\pi$-periodic function which coincides on $[-2^j\pi,2^j\pi]$ with $\hat{f}(\omega)\chi_{[-t,t]}(2^{-j}\omega)\overline{\hat{\phi}(2^{-j}\omega)}$, it follows that, on the interval $[-2^j\pi,2^j\pi]$,

$$\mathcal{F}P_jS_jf(\omega) = |\hat{\phi}_j(2^{-j}\omega)|^2\hat{f}(\omega)\chi_{[-t,t]}(2^{-j}\omega).$$

From this, we derive

$$\|S_jQ_jS_jf\|_0^2 = (2\pi)^{-1}\int_{|\omega|<2^jt}|\hat{f}(\omega) - \mathcal{F}P_jS_jf(\omega)|^2d\omega$$

$$
\begin{aligned}
&= (2\pi)^{-1}\int_{|\omega|<2^jt}|\hat f(\omega)|^2(1-|\hat\phi_j(2^{-j}\omega)|^2)^2d\omega\\
&\leq (2\pi)^{-1}2^{-2jr}\sup_{|\omega|<t}\left(\frac{1-|\hat\phi_j(\omega)|^2}{|\omega|^r}\right)^2\int_{|\omega|<2^jt}|\omega|^{2r}|\hat f(\omega)|^2d\omega\\
&= 2^{-2jr}\|f\|_r^2\sup_{|\omega|<t}\left(\frac{\sum_{n\neq0}|\hat\varphi_j(\omega+2n\pi)|^2}{|\omega|^r\sum_n|\hat\varphi_j(\omega+2n\pi)|^2}\right)^2\\
&\leq 2^{-2jr}\|f\|_r^2\sup_{|\omega|<t}\left(\frac{\sum_{n\neq0}|\hat\varphi_j(\omega+2n\pi)|^2}{|\omega|^r|\hat\varphi_j(\omega)|^2}\right)^2.
\end{aligned}
$$

Combining these estimates with hypothesis (3.4) in the case where $v=0$, we obtain

$$\|S_jQ_jS_jf\|_s^2\leq C2^{2j(s-r)}\|f\|_r^2\varepsilon(j) \tag{3.9}$$

with $0\leq\varepsilon(j)\leq1$ and $\varepsilon(j)\to0$ as $j\to+\infty$.

Finally, for the last term $\|T_jP_jS_jf\|_s$, we note that for ω such that $|\omega-2^{j+1}n\pi|<2^jt$, we have, by (3.8) and the observation following it,

$$\mathcal{F}P_jS_jf(\omega)=\hat\phi_j(2^{-j}\omega)\overline{\hat\phi_j(2^{-j}\omega-2n\pi)}\hat f(\omega-2^{j+1}n\pi), \tag{3.10}$$

and $\mathcal{F}P_jS_jf(\omega)=0$ for $2^jt<|\omega-2^{j+1}n\pi|<2^j\pi$. Consequently, we can estimate this last term as follows:

$$
\begin{aligned}
\|T_jP_jS_jf\|_s^2 &= (2\pi)^{-1}\int_{|\omega|>2^jt}|\mathcal{F}P_jS_jf(\omega)|^2(1+|\omega|^{2s})d\omega\\
&\leq C\sum_{n\neq0}\int_{|\omega|<2^jt}|\mathcal{F}P_jS_jf(\omega+2^{j+1}n\pi)|^2|\omega+2^{j+1}n\pi|^{2s}d\omega\\
&= C\int_{|\omega|<2^jt}|\hat\phi_j(2^{-j}\omega)\hat f(\omega)|^2\sum_{n\neq0}|\omega+2^{j+1}n\pi|^{2s}|\hat\phi_j(2^{-j}\omega+2n\pi)|^2d\omega\\
&\leq C\int_{|\omega|<2^jt}|\hat f(\omega)|^2\left(\sum_{n\neq0}|\omega+2^{j+1}n\pi|^{2s}|\hat\phi_j(2^{-j}\omega+2n\pi)|^2\right)d\omega\\
&\leq C2^{2j(s-r)}\sup_{|\omega|<t}\left(\frac{\sum_{n\neq0}|\omega+2n\pi|^{2s}|\hat\phi_j(\omega+2n\pi)|^2}{|\omega|^{2r}}\right)\int_{|\omega|<2^jt}|\hat f(\omega)|^2|\omega|^{2r}d\omega\\
&\leq C2^{2j(s-r)}\sup_{|\omega|<t}\left(|\omega|^{-2r}\frac{\sum_{n\neq0}|\omega+2n\pi|^{2s}|\hat\varphi_j(\omega+2n\pi)|^2}{|\hat\varphi_j(\omega)|^2}\right)\|f\|_r^2.
\end{aligned}
$$

Combining this estimate with hypothesis (3.4) in the case where $v=s$, we obtain

$$\|T_jP_jS_jf\|_s^2\leq C2^{2j(s-r)}\|f\|_r^2\varepsilon(j) \tag{3.11}$$

with $0\leq\varepsilon(j)\leq1$ and $\varepsilon(j)\to0$ as $j\to+\infty$. This concludes the proof of the theorem. □

We now return to the case of multiresolution spaces generated by the functions φ_j defined by (3.1) with the conditions on the masks stated in Theorem 2.2. The following result shows that, under an additional assumption on these masks, hypothesis (3.4) is satisfied for any $r,v\geq0$.

THEOREM 3.2. *Assume that the family* $\{m_k(\omega)\}_{k>0}$ *satisfies the hypotheses of Theorem* 2.2 *and*

$$|m_k(\omega)| \geq (1-\tilde{\alpha}_k)|a(\omega)|^k, \tag{3.12}$$

where $|\tilde{\alpha}_k| < 1$, $\sum_k |\tilde{\alpha}_k| < +\infty$, $a(\omega)$ *is Hölder continuous at the origin with exponent* $\gamma > 0$ *and* $a(0) = 1$. *Then there exists* $t > 0$ *such that for all* $r, v \geq 0$,

$$\sup_{|\omega|<t} \left(|\omega|^{-2r} \frac{\sum_{n\neq 0} |\omega + 2n\pi|^{2v} |\hat{\varphi}_j(\omega+2n\pi)|^2}{|\hat{\varphi}_j(\omega)|^2} \right) \to 0 \quad \textit{as } j \to +\infty. \tag{3.13}$$

Proof. We shall prove that, for fixed $v > 0$, there exist constants $A, B, C, D, T > 0$ such that if $|\omega| < T$, the following holds:

$$|\hat{\varphi}_j(\omega)|^2 \geq AB^j \tag{3.14}$$

and

$$\sum_{n\neq 0} |\omega + 2n\pi|^{2v} |\hat{\varphi}_j(\omega + 2n\pi)|^2 \leq C|D\omega|^{2\beta j}, \tag{3.15}$$

where β is the exponent that appears in the hypothesis of Theorem 2.2. It will then be sufficient to choose $t \in]0, T[$ such that $(Dt)^{2\beta}/B < 1$ in order to ensure (3.13).

First, we introduce four functions:

$$g_m(\omega) = \prod_{k=1}^{\infty} |m(2^{-k}\omega)|, \qquad h_m(\omega) = \prod_{k=1}^{\infty} |m(2^{-k}\omega)|^k \tag{3.16}$$

and

$$g_a(\omega) = \prod_{k=1}^{\infty} |a(2^{-k}\omega)|, \qquad h_a(\omega) = \prod_{k=1}^{\infty} |a(2^{-k}\omega)|^k. \tag{3.17}$$

The above infinite products are convergent since the functions $m(\omega)$ and $a(\omega)$ are Hölder continuous at the origin, achieving the value 1 there, and hence the logarithm of each of the four infinite products is a convergent series.

By (3.1), (2.10), and (3.12), we have

$$M_a h_a(\omega)[g_a(\omega)]^j \leq |\hat{\varphi}_j(\omega)| \leq M_m h_m(\omega)[g_m(\omega)]^j, \tag{3.18}$$

where

$$M_a = \prod_{k=1}^{\infty} (1 - |\tilde{\alpha}_k|), \qquad M_m = \prod_{k=1}^{\infty} (1 + |\alpha_k|) \tag{3.19}$$

To get an upper bound for the numerator of (3.13), define

$$g_m(\omega) = \operatorname{sinc}^{\beta}(\omega/2)\tilde{g}(\omega) \tag{3.20}$$

with $\tilde{g}(\omega) = \prod_{k=1}^{\infty} |\tilde{m}(2^{-k}\omega)|$. For $l \geq 0$ and $2^l \leq |\omega| \leq 2^{l+1}$, we have, by the hypotheses of Theorem 2.2,

$$\begin{aligned} \tilde{g}(\omega) &= \tilde{g}(2^{-l}\omega) \prod_{k=1}^{l} |\tilde{m}(2^{-k}\omega)| \\ &\leq \sup\nolimits_{|\omega|<2}(\tilde{g}(\omega))(\sup|\tilde{m}|)^{i-1}(\sigma_i)^{[l/i]} \\ &\leq \sup\nolimits_{|\omega|<2}(\tilde{g}(\omega))(\sup|\tilde{m}|)^{i-1}|\omega|^{\frac{\log_2 \sigma_i}{i}} \leq K_1|\omega|^{\beta}. \end{aligned} \tag{3.21}$$

Combining (3.21) with (3.19), we obtain the estimates

$$g_m(\omega) \le K_2(1+|\omega|)^{-\varepsilon} \tag{3.22}$$

with $\varepsilon = \beta - (\log_2 \sigma_i)/i > 0$ and

$$g_m(\omega + 2n\pi) \le K_1|\omega|^{\beta}, \quad n \in \mathbb{Z} - \{0\},\ |\omega| \le \pi. \tag{3.23}$$

In the last inequality, we used the bound

$$\operatorname{sinc}(n\pi + \omega/2) \le |\omega||\omega + 2n\pi|^{-1}, \quad n \in \mathbb{Z} - \{0\}. \tag{3.24}$$

Finally, by (3.22) and since $g_m(\omega)$ is Hölder continuous at the origin with the same exponent as $m(\omega)$, we can also write $h_m(\omega)$ as

$$h_m(\omega) = \prod_{k=0}^{\infty} g_m(2^{-k}\omega). \tag{3.25}$$

For any $\delta \in]0,1[$, choose $\omega_\delta > 0$ such that $g_m(\omega) < \delta$ if $|\omega| > \omega_\delta$. For $l \ge 0$ and $2^l\omega_\delta \le |\omega| < 2^{l+1}\omega_\delta$, we thus have

$$\begin{aligned} h_m(\omega) &= h_m(2^{-l-1}\omega)\prod_{k=0}^{l} g_m(2^{-k}\omega) \\ &\le [\sup_{|\omega|\le\omega_\delta} h_m(\omega)]\delta^{l+1} \\ &\le [\sup_{|\omega|\le\omega_\delta} h_m(\omega)]\delta^{\log_2(|\omega|/\omega_\delta)} \\ &= K_\delta|\omega|^{\log_2\delta}. \end{aligned}$$

Since δ is arbitrary, this shows that $h_m(\omega)$ has rapid decay at infinity. We can thus define

$$K_3 = \sum_{n\in\mathbb{Z}} |\omega + 2n\pi|^{2v}|h_m(\omega + 2n\pi)|^2 < +\infty, \tag{3.26}$$

and conclude estimate (3.15) from (3.18) and (3.22) with $C = M_m^2K_3$ and $D = K_1^{1/\beta}$.

To obtain a lower bound for the denominator of (3.13), we use the fact that $h_a(\omega)$ and $g_a(\omega)$ are Hölder continuous at the origin with the same exponent as $a(\omega)$ and both are 1 at the origin. Thus $h_a(\omega) \ge \theta_1 > 0$ and $g_a(\omega) \ge \theta_2 > 0$ for $|\omega|$ small enough, which together with (3.18) yields (3.14) with $A = M_a^2\theta_1^2$ and $B = \theta_2^2$. □

Combining Theorem 3.1 and 3.2, we immediatly obtain the following corollary.

COROLLARY 3.3. *Let $\{m_k(\omega)\}_{k>0}$ be a family of trigonometric polynomials that satisfy the hypotheses of Theorem* 3.2 *and let $\{\varphi_j\}_{j\ge0}$ be the associated scaling functions defined by* (3.1). *Then the semimultiresolution analysis $\{V_j\}_{j\ge0}$ generated by these functions achieve spectral approximation, i.e., it has density order r in H^s for all $r \ge s \ge 0$.*

Remarks. As in the previous section, the hypotheses here can also be weakened by assuming that the hypotheses on the functions $m_k(\omega)$ hold only for k sufficiently large since only the scaling functions at coarse scales will be affected.

Here we have concentrated on the rate of approximation in Sobolev spaces $H^s(\mathbb{R})$. If we are interested in the approximation of a function $f \in H^r(I)$, where I is a finite interval, we can use the space V_j^I defined by the restriction into I of the functions of V_j. This space has finite dimension, and we can easily check that there exists a constant $C > 0$ such that $\dim(V_j^I) \leq C2^j$. The property that $2^{(r-s)j}d(f, V_j^I)_s$ goes to 0 as $j \to +\infty$ reveals that $\{V_j^I\}_{j\geq 0}$ is a sequence of n-width approximation spaces for Sobolev spaces. In that sense, the V_j spaces are optimal for approximation in all Sobolev spaces.

One may also like to obtain similar optimal results for other types of Besov spaces $B_q^s(L^p)$, e.g., Hölder space $C^s = B_\infty^s(L^\infty)$. The generalizations of the results of this section depend on p in an essential way:

- When $p = 2$, for $H^s = B_2^s(L^2)$ in particular, these spaces can be fully described in the Fourier domain because of the isometry of $f \to \hat{f}$ with respect to the L^2 norm, and Theorem 3.1 can be generalized to $q \neq 1$.
- When $p \neq 2$, this task seems much more difficult since one has to deal with the space domain in which the support of the functions φ_j grows with j, at least linearly. Take, for example, the Hölder space C^s ($s = N + r$, $r \in]0, 1]$): a natural method of approximating $f \in C^s$ by a function in V_j is to approximate f locally by a polynomial of degree N with an L^∞ error of order $(\Delta x)^s$ and to generate these polynomials locally by the functions $\{\varphi_j(2^j x - k)\}_{k\in\mathbb{Z}}$ provided that the Strang–Fix conditions of order N are satisfied. In the case that we consider, due to the particular properties of $m_k(\omega)$, the V_j spaces contain polynomials of arbitrarily high degree as j grows. However, since the support of the functions φ_j grows at least linearly, one can only predict an L^∞ error of order $(j2^{-j})^s$ for arbitrary s. In the standard multiresolution-analysis framework, the error will be of order 2^{-sj} since the generating function is the same at every scale, but only for $s \leq N + 1$, N being the order of the Strang–Fix conditions, which is constant for all $j \geq 0$.

For C^∞ functions, better rates of convergences can be expected. An interesting perspective is the exploration of the approximation properties of the V_j spaces in classes of C^∞ functions defined by the growth of the sup norm of their derivatives (Gevrey classes; see Hormander (1983)).

Finally, let us remark that an important problem—besides knowing the rate of convergence of the best approximation in a fixed norm—is the construction of this best approximation numerically. In the case of Sobolev spaces, we gave an explicit approximation operator that combines truncation in the Fourier domain with L^2 projection onto V_j. An interesting case is when the functions $\{\varphi_j(x - k)\}_{k\in\mathbb{Z}}$ form orthonormal systems for all $j \geq 0$ so that the L^2 projection can be expressed directly in terms of theses compactly supported functions. The next section will focus on an example of such a system, showing that it leads to orthonormal bases of smooth compactly supported wavelets that are Riesz bases for all Sobolev spaces and that the asymptotically best H^s approximation is always achieved by the L^2 projection.

4. Smooth cardinal interpolation and orthonormal wavelet bases with compact support. By definition, a scaling function has the property of cardinal interpolation if

$$\varphi(k) = \delta_{0,k} \tag{4.1}$$

and the property of orthonormality if

$$\langle \varphi(\cdot - k) | \varphi(\cdot - l) \rangle = \delta_{k,l}. \tag{4.2}$$

In the framework of standard multiresolution-analysis and refinement equations, such functions were first studied by Deslauriers and Dubuc (1987) and Dyn, Gregory, and Levin (1987) in the interpolating case and by Daubechies (1988) and Meyer (1990) in the orthonormal case. They can be generated by imposing specific necessary constraints on the coefficients c_n in (1.4) or, equivalently, on the function $m(\omega)$. These constraints are $c_{2n} = (1/2)\delta_{n,0}$ or, equivalently,

$$m(\omega) + m(\omega + \pi) = 1 \tag{4.3}$$

for interpolation and

$$|m(\omega)|^2 + |m(\omega + \pi)|^2 = 1 \tag{4.4}$$

for orthonormality. We recall here the main results of these constructions without proving them since they are now classics and can be found in detail in Daubechies (1992) and Meyer (1990):

- The constraint (4.3) (resp. (4.4)) will be sufficient to ensure (4.1) (resp. (4.2)) if and only if the associated subdivision scheme (i.e., the sequence $\varphi^{[n]}$) converges in C^0 (resp. in L^2) since the functions $\varphi^{[n]}$ that interpolate the subdivision at step n satisfy by recursion the interpolation (resp. orthonormality) property.
- It is clear that a trigonometric polynomial solution of (4.4) leads to one of (4.3) by taking its square modulus. The converse is also true assuming that the solution $M(\omega)$ of (4.3) is a positive, even trigonometric polynomial; then the Riesz lemma ensures the existence of a trigonometric polynomial $m(\omega)$ that satisfies $|m|^2 = M$.
- An important family of solutions of (4.3) is given by

$$M_N(\omega) = \cos^{2N}(\omega/2) \sum_{j=0}^{N-1} \binom{N-1+j}{j} \sin^{2j}(\omega/2) \tag{4.5}$$

for all $N > 0$. It is associated with a family $\{m_N(\omega)\}_{N>0}$ of solutions of (4.4) by the previous argument. It was shown by Daubechies (1988) that the regularity of the associated orthonormal and interpolating scaling functions increases linearly with N in an asymptotical sense.

- Let $\{u(x-k)\}_{k\in\mathbb{Z}}$ be an orthonormal basis of a closed subspace $U \subset L^2(\mathbb{R})$, let $m(\omega)$ be a solution of (4.4), and define $v, w \in U$ by

$$\begin{aligned} \hat{v}(\omega) &= m(\omega/2)\hat{u}(\omega/2), \\ \hat{w}(\omega) &= e^{-i\omega/2}\overline{m(\omega/2+\pi)}\hat{u}(\omega/2). \end{aligned} \tag{4.6}$$

Then $\{2^{-1/2}v(x/2-k)\}_{k\in\mathbb{Z}} \cup \{2^{-1/2}w(x/2-k)\}_{k\in\mathbb{Z}}$ constitutes an orthonormal basis of U. Applying this splitting trick to the refinable orthonormal function detemined by $m(\omega)$, one defines the "mother wavelet" ψ by $\hat{\psi}(\omega) = e^{-i\omega/2}\overline{m(\omega/2+\pi)}\hat{\varphi}(\omega/2)$. The complete system $\{2^{j/2}\psi(2^jx-k)\}_{j,k\in\mathbb{Z}}$ constitutes an orthonormal wavelet basis of $L^2(\mathbb{R})$ and so does $\{\varphi(x-k)\}_{k\in\mathbb{Z}} \cup \{2^{j/2}\psi(2^jx-k)\}_{j\geq 0,k\in\mathbb{Z}}$.

- The function ψ has the same support length and global regularity as φ (in the compactly supported univariate case). The Strang–Fix conditions of order N for φ are equivalent to the property that ψ has $N+1$ vanishing moments, i.e., $\int x^k\psi(x)dx = 0$ for $k = 0, \dots, N$, and hold if φ and ψ are in $H^s(\mathbb{R})$ ($s = N+r$, $r \in]0,1]$). In addition, the wavelet basis is also a Riesz basis for all Sobolev spaces with exponent smaller than s when $\varphi, \psi \in H^s(\mathbb{R})$.

We shall now use the families $\{M_N(\omega)\}_{N>0}$ and $\{m_N(\omega)\}_{N>0}$ described by (4.5) and $|m_N|^2 = M_N$ as the masks of nonstationary subdivision schemes. We will be able to apply the results of the previous sections thanks to an estimate borrowed from Volkmer (1992).

LEMMA 4.1. *The functions $M_N(\omega)$ satisfy*

$$|\cos^2(\omega/2)|^N \leq M_N(\omega) \leq |M(\omega)|^N, \tag{4.7}$$

where $M(\omega)$ is a 2π-periodic function satisfying $M(\omega) = 1$ if $|\omega| \leq \pi/2$ and $M(\omega) = \sin^2(\omega)$ if $\pi/2 \leq |\omega| \leq \pi$. The function $M(\omega)$ can be written as $M(\omega) = \cos^2(\omega/2)\tilde{M}(\omega)$, where $\tilde{M}(\omega)$ is bounded and Hölder continuous and satisfies

$$\sigma_2 = \sup|\tilde{M}(\omega)\tilde{M}(2\omega)| < 2^4 = 16. \tag{4.8}$$

Proof. From (4.5), we clearly have $M_N(\omega) \geq \cos^{2N}(\omega/2)$. The upper estimate is also trivial for $|\omega| \leq \pi/2$ in view of (4.3). For $|\omega| \geq \pi/2$, note that the polynomial $P_N(y) = \sum_{j=0}^{N-1} \binom{N-1+j}{j} y^j$ with $1/2 \leq y \leq 1$ satisfies

$$\begin{aligned} P_N(y) &= \sum_{j=0}^{N-1} \binom{N-1+j}{j}(2y)^j 2^{-j} \leq (2y)^{N-1} \sum_{j=0}^{N-1} \binom{N-1+j}{j} 2^{-j} \\ &= (2y)^{N-1} P_N(1/2) = (4y)^{N-1} \leq (4y)^N. \end{aligned}$$

(We find $P_N(1/2) = 2^{N-1}$ by taking $y = 1/2$ in the equation $(1-y)^N P_N(y) + y^N P_N(1-y) = 1$, which expresses that M_N is a solution of (4.3).) Consequently, we have, for $\pi/2 \leq |\omega| \leq \pi$,

$$\begin{aligned} M_N(\omega) &= \cos^{2N}(\omega/2) P_N(\sin^2(\omega/2) \leq [4\cos^2(\omega/2)\sin^2(\omega/2)]^N \\ &= [\sin^2(\omega)]^N = |M(\omega)|^N. \end{aligned}$$

To prove (4.8), we remark that $\sup_\omega(\tilde{M}(\omega)) = 4$ and that this supremum is attained only at the points $\omega = (2n+1)\pi$. Consequently, $\sup_\omega(\tilde{M}(\omega)\tilde{M}(2\omega)) < \sup_\omega |\tilde{M}(\omega)|^2 = 16$. □

We are now ready to apply the results of the previous sections. For all $j \geq 0$, we define

$$\hat{\Phi}_j(\omega) = \prod_{k=1}^{+\infty} M_{j+k}(2^{-k}\omega) \tag{4.9}$$

and

$$\hat{\varphi}_j(\omega) = \prod_{k=1}^{+\infty} m_{j+k}(2^{-k}\omega). \tag{4.10}$$

THEOREM 4.2. *For all $j > 0$, the subdivision algorithms associated with Φ_j and φ_j converge to these functions in the sense of the uniform convergence of all the derivatives. The functions ϕ^j (resp. φ_j) are in C^∞, are compactly supported in $[-3-2j, 3+2j]$ (resp. $[0, 3+2j]$), and have the property of cardinal interpolation (resp. orthonormality). The families $\{\phi^j\}_{j\geq 0}$ and $\{\varphi_j\}_{j\geq 0}$ both generate multiresolution analyses that have the property of spectral approximation.*

Proof. This result is a simple application of the theorems in the previous sections: Lemma 4.1 indicates that the functions $M_N(\omega)$ and $m_N(\omega)$ satisfy the hypotheses of Theorems 2.1, 2.2, and 3.2. The convergence of the subdivision schemes and the spectral approximation property follow from these results.

Since $M_N(\omega) = \sum_{n=0}^{2N-1} C_n \cos(n\omega)$ and $m_N(\omega) = \sum_{n=0}^{2N-1} c_n e^{-in\omega}$, the supports of ϕ^j and φ_j are, respectively, the intervals $[-L_j, L_j]$ and $[0, L_j]$ with $L_j = \sum_{k>0}(2k+2j-1)2^{-k} = 3+2j$.

Finally, the properties of cardinal interpolation and orthonormality can be derived with the same arguments as in the stationary case: we consider the functions that interpolate the subdivision values after n steps, i.e.,

$$\hat{\Phi}_j^{[n]}(\omega) = \prod_{k=1}^{n} M_{j+k}(2^{-k}\omega)\chi_{[-\pi,\pi]}(2^{-n}\omega) \tag{4.11}$$

and

$$\hat{\varphi}_j^{[n]}(\omega) = \prod_{k=1}^{n} m_{j+k}(2^{-k}\omega)\chi_{[-\pi,\pi]}(2^{-n}\omega). \tag{4.12}$$

The function that is defined by $\hat{\Phi}_j^{[0]} = \hat{\varphi}_j^{[0]} = \chi_{[-\pi,\pi]}$ is clearly both orthonormal and interpolatory. The standard theory (see Daubechies (1992) and Deslauriers and Dubuc (1987)) indicates that the property of cardinal interpolation (resp. orthonormality) of a function u is preserved by the transformation $u \to v$ such that $\hat{v}(\omega) = m(\omega/2)\hat{u}(\omega/2)$ if $m(\omega)$ is a solution of (4.3) (resp. (4.4)).

Consequently, we can easily check by recursion that $\Phi_j^{[n]}$ (resp. $\varphi_j^{[n]}$) is interpolatory (resp. orthonormal) for all $n, j \geq 0$, and the same holds for Φ_j (resp. φ_j) by uniform convergence of the subdivision algorithm. □

Since in the interpolatory case, the subdivision with a finite number of iterations already has the property of cardinal interpolation, it is possible to use these masks to interpolate a sequence on a grid of points $\Gamma = \gamma\mathbb{Z}$ into a sequence on a finer grid $2^{-j}\Gamma$. As j goes to $+\infty$, the limit curve is C^∞ and the influence of a single point in Γ on this curve is limited in space since the limit function from a Dirac sequence is compactly supported. This is clearly a numerical advantage over interpolation with cardinal sines. Figure 1 represents the function Φ_0 obtained after 10 iterations of the subdivision.

For the orthonormal case, in Figure 2, we have represented the function φ_0. In this setting, we can use the "splitting trick" that we have mentionned for the standard case to build orthonormal wavelet bases: for each $j \geq 0$, we define

$$\hat{\psi}_j(\omega) = e^{-i\omega/2}\overline{m_{j+1}(\omega/2+\pi)}\hat{\varphi}_{j+1}(\omega/2) \tag{4.13}$$

so that the family

$$\psi_{j,k} = 2^{j/2}\psi_j(2^j x - k), \quad k \in \mathbb{Z}, \tag{4.14}$$

constitutes an orthonormal basis of the orthogonal complement W_j of V_j in V_{j+1}. Since the spaces V_j approximate any L^2 function with arbitrarily small L^2 error as j grows, the system $\{\varphi(x-k)\}_{k\in\mathbb{Z}} \cup \{\psi_{j,k}(x)\}_{j\geq 0, k\in\mathbb{Z}}$ constitutes an orthonormal basis of $L^2(\mathbb{R})$. Note that the function ψ_j has j vanishing moments and that, although its support grows linearly with j, this growth is dominated by the scaling 2^{-j} so that these wavelets can still be used to analyze and represent the local features of a function.

The last result of this section indicates that this particular system has very good properties with respect to Sobolev spaces.

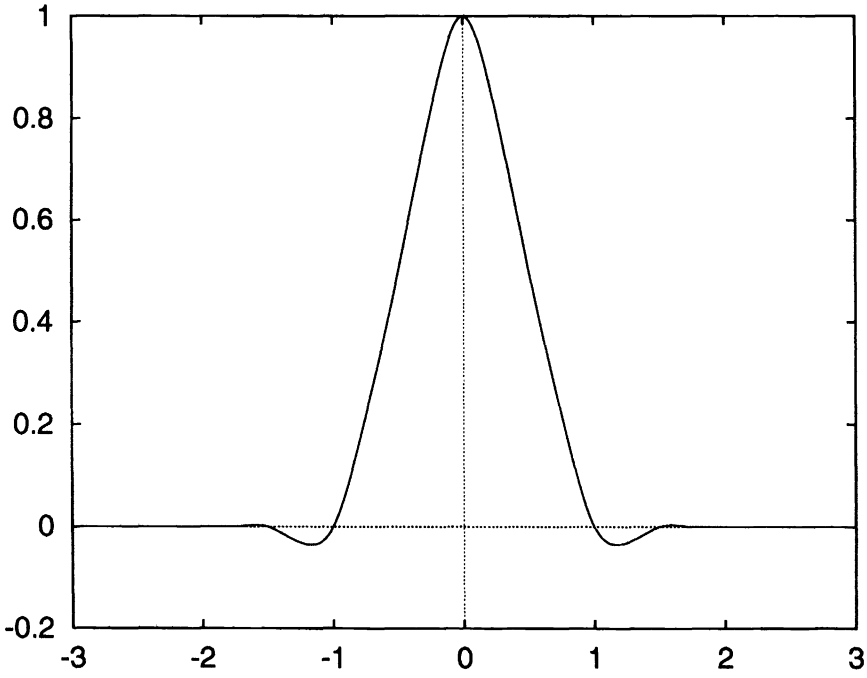


FIG. 1.

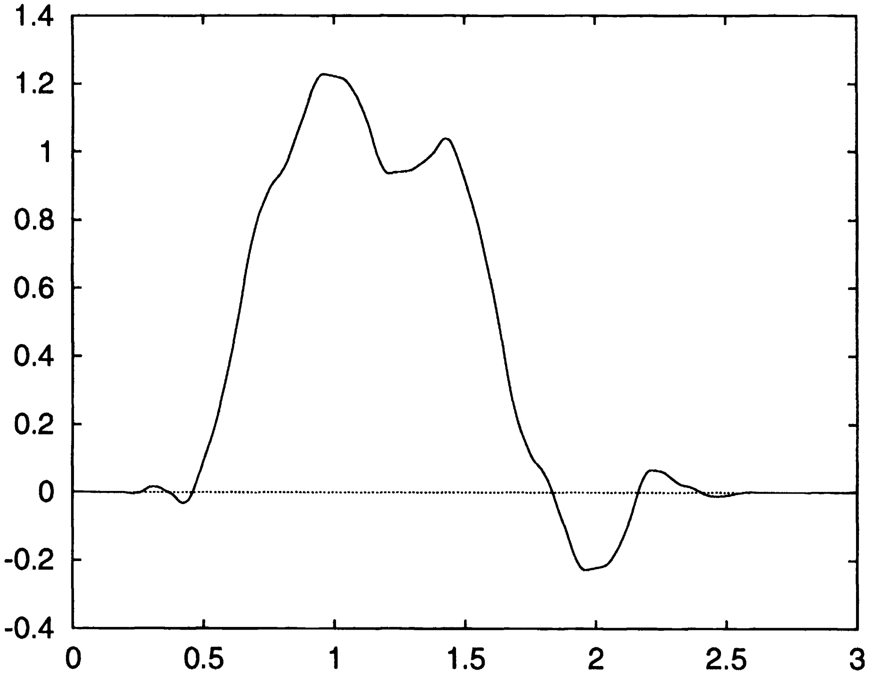


FIG. 2.

THEOREM 4.3. *The system $\{\varphi(x-k)\}_{k\in\mathbb{Z}}\cup\{\psi_{j,k}(x)\}_{j\in\mathbb{N},k\in\mathbb{Z}}$ constitutes a Riesz basis for all Sobolev spaces H^s:*

If $f(x)=\sum_{k\in\mathbb{Z}}c_k\varphi(x-k)+\sum_{j\geq 0,k\in\mathbb{Z}}d_{j,k}\psi_{j,k}(x)$ is a function in H^r, then its wavelet series converges in H^s for $0\leq s\leq r$ and we have the equivalence

$$\|f\|_r^2\approx\sum_k|c_k|^2+\sum_{j,k}2^{2rj}|d_{j,k}|^2=\|P_0(f)\|_0^2+\sum_{j\geq 0}2^{2jr}\|R_j(f)\|_0^2, \tag{4.15}$$

where $R_j=P_{j+1}-P_j$ is the L^2 projector onto W_j. Also, density order r in H^s norm is achieved by the L^2 projection in the sense that $2^{j(r-s)}\|P_jf-f\|_s\to 0$ as $j\to+\infty$ for any $f\in H^r$.

Proof. First, we define a function $u(\omega)$ and a sequence of functions $\{v_n(\omega)\}_{n>0}$ by

$$u(\omega)=\prod_{k=1}^{+\infty}|m(2^{-k}\omega)|^k \tag{4.16}$$

and

$$v_n(\omega)=|m(\omega/2+\pi)|^{n+1}\prod_{k=2}^{+\infty}|m(2^{-k}\omega)|^k, \tag{4.17}$$

where $m(\omega)=\sqrt{M(\omega)}$. The same arguments used for the function $h(\omega)$ in the proof of Theorem 3.2 indicate that $u(\omega)$ and $v_n(\omega)$ have rapid decay at infinity. Using Lemma 4.1 and the fact that $m(\omega)$ is bounded by 1, we obtain

$$|\hat{\varphi}_j(\omega)|\leq u(\omega) \tag{4.18}$$

for all $j\geq 0$ and

$$|\hat{\psi}_j(\omega)|\leq v_n(\omega) \tag{4.19}$$

for all $j\geq n\geq 0$. Now consider a function $f\in H^r$ with $n\leq r<n+1$. We clearly have

$$\|P_0f\|_0^2+\sum_{j=0}^{2n}2^{2rj}\|R_jf\|_0^2\leq C(r)\|f\|_0^2. \tag{4.20}$$

with $C(r)=1+\sum_{j=0}^{2n}2^{2rj}$. For the scales $j>2n$, by Parseval and Poisson identities, we have

$$\begin{aligned}\|R_jf\|_0^2&=\sum_{k\in\mathbb{Z}}|\langle f|\psi_{j,k}\rangle|^2\\&=(2\pi)^{-2}\sum_{k\in\mathbb{Z}}|\langle\hat{f}|\hat{\psi}_{j,k}\rangle|^2\\&=(2\pi)^{-2}2^{-j}\sum_{k\in\mathbb{Z}}\left|\int\hat{f}(\omega)\overline{\hat{\psi}_j(2^{-j}\omega)e^{i2^{-j}k\omega}}d\omega\right|^2\\&=(2\pi)^{-2}2^j\sum_{k\in\mathbb{Z}}\left|\int\hat{f}(2^j\omega)\overline{\hat{\psi}_j(\omega)e^{ik\omega}}d\omega\right|^2\\&=(2\pi)^{-1}2^j\int_{-\pi}^{\pi}\left|\sum_{l\in\mathbb{Z}}\hat{f}(2^j(\omega+2l\pi))\overline{\hat{\psi}_j(\omega+2l\pi)}\right|^2d\omega.\end{aligned}$$

Using (4.20) and the Schwarz inequality, we then obtain

$$\begin{aligned}\|R_jf\|_0^2 &\leq (2\pi)^{-1}2^j\int_{-\pi}^{\pi}\left(\sum_{l\in\mathbb{Z}}|\hat{f}(2^j(\omega+2l\pi))v_{2n+1}(\omega+2l\pi)|\right)^2 d\omega\\ &\leq (2\pi)^{-1}2^jA(r)\int_{-\pi}^{\pi}\left(\sum_{l\in\mathbb{Z}}|\hat{f}(2^j(\omega+2l\pi))|^2v_{2n+1}(\omega+2l\pi)\right)d\omega\\ &\leq (2\pi)^{-1}A(r)\int|\hat{f}(\omega)|^2v_{2n+1}(2^{-j}\omega)d\omega\end{aligned}$$

with $A(r)=\sup_\omega[\sum_{l\in\mathbb{Z}}v_{2n+1}(\omega+2l\pi)]<+\infty$. For $r<n+1$, we can also define $B(r)=\sup_\omega(\sum_{j\in\mathbb{Z}}|2^{-j}\omega|^{-2r}v_{2n+1}(2^{-j}\omega))<+\infty$ (because v_{2n+1} has a zero of order $2n+2$ at the origin). Summing on $j>2n$ our previous estimate with the appropriate weights, we thus obtain

$$\begin{aligned}\sum_{j=2n+1}^{+\infty}2^{2rj}\|R_jf\|_0^2 &= (2\pi)^{-1}A(r)\sum_{j=2n+1}^{+\infty}2^{2rj}\int|\hat{f}(\omega)|^2v_{2n+1}(2^{-j}\omega)d\omega\\ &= (2\pi)^{-1}A(r)\sum_{j=2n+1}^{+\infty}\int|\omega|^{2r}|\hat{f}(\omega)|^2|2^{-j}\omega|^{-2r}v_{2n+1}(2^{-j}\omega)d\omega\\ &\leq A(r)B(r)\|f\|_r^2.\end{aligned}$$

Combining our results for all $j\geq 0$, we finally obtain

$$\sum_k|c_k|^2+\sum_{j,k}2^{2rj}|d_{j,k}|^2\leq[A(r)B(r)+C(r)]\|f\|_r^2. \tag{4.21}$$

For the converse inequality, we first prove the following lemma.

LEMMA 4.4. *For every $r\geq 0$, there exists $K(r)$ such that for all $j\in\mathbb{N}$, the functions $f\in V_j$ satisfy the following Bernstein-type estimate:*

$$f\in V_j\ \Rightarrow\ \|f\|_0\leq\|f\|_r\leq K(r)2^{rj}\|f\|_0. \tag{4.22}$$

Proof. Let $f=\sum_{k\in\mathbb{Z}}a_k\varphi_j(2^j\cdot -k)$, $\sum|a_k|^2<\infty$. Then

$$\|f\|_r^2=\frac{2^{-j}}{2\pi}\int_{\mathbb{R}}\left|\sum_{k\in\mathbb{Z}}a_ke^{-i2^{-j}k\omega}\right|^2|\hat{\varphi}_j(2^{-j}\omega)|^2(1+|\omega|^{2r})d\omega. \tag{4.23}$$

Using (4.1), we get

$$\begin{aligned}\|f\|_r^2 &\leq (2\pi)^{-1}\int\left|\sum_{k\in\mathbb{Z}}a_ke^{-ik\omega}\right|^2|u(\omega)|^2(1+|\omega|^{2r}2^{2rj})d\omega\\ &\leq (2\pi)^{-1}2^{2rj}\int\left|\sum_{k\in\mathbb{Z}}a_ke^{-ik\omega}\right|^2|u(\omega)|^2(1+|\omega|^{2r})d\omega\\ &\leq (K(r))^22^{2rj}(2\pi)^{-1}\int_{-\pi}^{\pi}\left|\sum_{k\in\mathbb{Z}}a_ke^{-ik\omega}\right|^2d\omega\\ &\leq (K(r))^22^{2rj}\|f\|_0^2\end{aligned}$$

with $(K(r))^2 = \sup_\omega \sum_{l\in\mathbb{Z}}(1+|\omega+2\pi l|^{2r})u^2(\omega+2\pi l) < \infty$. □

Returning to the proof of the theorem, to conclude (4.15), we use the following result about Besov spaces (see, for example, Frazier and Jawerth (1985)): for $r > 0$, $1 \leq p, q \leq +\infty$, the norm of a function f in $B^r_{p,q}$ is equivalent to the quantity

$$\inf \left(\sum_{j=0}^{+\infty} \varepsilon_j^q\right)^{1/q}, \tag{4.24}$$

where the infimum is taken over all the sequences of positive numbers $\{\varepsilon_j\}_{j\geq 0}$ such that for some integer $m > r$, there exists a sequence of functions $\{f_j\}_{j\geq 0}$ in W^m_p that satisfies

$$f = \sum_j f_j, \left\|\left(\frac{d}{dx}\right)^m f_j\right\|_{L^p} \leq \varepsilon_j 2^{(m-r)j}, \quad \|f_j\|_{L^p} \leq \varepsilon_j 2^{-rj}. \tag{4.25}$$

It is clear that if $\|P_0(f)\|_0^2 + \sum_j 2^{2jr}\|R_j(f)\|_0^2$ is bounded, $f_0 = P_0 f$ and $f_j = R_{j-1}f$ for $j > 0$ will be a decomposition that satifies (4.25) with $\varepsilon_j = 2^{rj}\|f_j\|_0 K(m)$ in $\ell^2(\mathbb{N})$ since by Lemma 4.4,

$$\left\|\left(\frac{\mathrm{d}}{\mathrm{dx}}\right)^m f_j\right\|_0 \leq \|f_j\|_m \leq K(m)2^{mj}\|f_j\|_0 = \varepsilon_j 2^{(m-r)j}. \tag{4.26}$$

Consequently, we have

$$\|f\|_r^2 \leq D(r)\left(\|P_0(f)\|_0^2 + \sum_j 2^{2jr}\|R_j(f)\|_0^2\right), \tag{4.27}$$

and (4.15) is proved.

Finally, we can use (4.27) and (4.21) to evaluate $\|P_j f - f\|_s^2$ for $f \in H^r$, $r > s$. We obtain

$$\begin{aligned}\|P_j f - f\|_s^2 &\leq D(s)\sum_{l=j}^{+\infty} 2^{2ls}\|R_l(f)\|_0^2\\ &\leq 2^{2j(s-r)}D(s)\sum_{l=j}^{+\infty} 2^{2lr}\|R_l(f)\|_0^2\\ &\leq 2^{2j(s-r)}K(r,s)\|f\|_r^2\varepsilon(j,f),\end{aligned}$$

where $K(r,s) = D(s)(A(r)B(r) + C(r))$, $\varepsilon(j,f) \in [0,1]$ going to 0 as $j \to +\infty$. Spectral approximation is thus achieved by the L^2 projection onto V_j. □

Remarks. One may be interested in the characterization of other Besov spaces, such as Hölder spaces, by the properties of their wavelet coefficients. This task seems difficult for exactly the same reason that was mentioned at the end of the previous section: one would have to deal with the space domain in which the support of ψ^j grows linearly.

In the more general framework of a nonstationary multiresolution analysis $\{V_j\}_{j\geq 0}$ described in the previous section, it would be interesting to prove an equivalence of the type of (4.15),

$$\|f\|_s^2 \approx \|P_0 f\|_0^2 + \sum_{j\geq 0} 2^{2js}\|P_{j+1}f - P_j f\|_0^2, \tag{4.28}$$

since it is the key to multilevel preconditioning techniques (Dahmen and Kunoth (1992)), even when there is no explicit wavelet basis. We do not know so far under which general conditions such an equivalence holds.

5. Conclusions. We have studied the convergence of nonstationary subdivision schemes and the properties of their limit functions in terms of approximation. Some open problems are raised in a natural way:

- the approximation and characterization of functions in spaces other than H^s (more general Besov spaces and Gevrey classes of C^∞ functions);
- the possible generalizations to the multidimensional framework (other than simple tensor products) and to biorthogonal wavelets;
- the search for weaker (if possible, necessary and sufficient) conditions for convergence of the subdivision algorithm, generating C_0^∞ functions.

The orthonormal wavelet basis that we described in the last section gives an answer to an important question in approximation: find a basis that has spectral approximation properties, like the trigonometric system in Sobolev spaces with boundary conditions, and allows us to describe functions locally in space. In particular, we can adapt this system to an interval using for example the approach described in Cohen, Daubechies, and Vial (1993) so that the expansion in the adapted basis asymptotically achieves the n-width in any Sobolev space $H^s(I)$.

The numerical implementation of such an expansion has, of course, the same pyramidal structure as in the standard fast wavelet transform algorithm, using longer filters at the high scales and smaller filters at the coarse scales. This idea is very natural in terms of signal processing since it does not make sense to use a filter with a comparable size to that of the whole signal at the coarse scales. Our results on the convergence of nonstationary subdivision indicate that the regularity of the reconstruction will not be affected by the use of small filters (corresponding to nonregular scaling functions) at the coarsest scales. An interesting perspective is the application of these algorithms in image compression, where both smoothness and space localization of the limit function are required.

Appendix. We want to show here that, under mild conditions, C^∞ scaling functions can be obtained with any type of growth of the mask length in the subdivision.

Let $\{m_k(\omega)\}_{k>0}$ be a family of trigonometric polymomials that satisfies

$$|m_k(\omega)| \le |m(\omega)|^k, \tag{A.1}$$

where $|m(\omega)| = |\cos^\beta(\omega/2)\tilde{m}(\omega)| \le 1$ with $\beta \ge 0$ and $\tilde{m}(\omega)$ is bounded and Hölder continuous at the origin and satisfies $\sigma_i = \sup_\omega |\prod_{k=1}^i \tilde{m}(2^k\omega)| < 2^{\beta i}$ for some $i > 0$.

Now consider a sequence of strictly positive numbers $\{l(k)\}_{k>0}$ with the following properties:

$$\lim_{k\to+\infty} l(k) = +\infty \tag{A.2}$$

and

$$\sum_{k>0} 2^{-k} d(l(k)) < +\infty. \tag{A.3}$$

From Theorem 2.1, we know that we can define a compactly supported tempered distribution φ by

$$\hat{\varphi}(\omega) = \prod_{k=1}^{+\infty} m_{l(k)}(2^{-k}\omega), \tag{A.4}$$

where the product converges uniformly on every compact set.

To study the convergence of the associated subdivision algorithm in the strong sense, we define

$$\hat{\varphi}^{[n]}(\omega) = \prod_{k=1}^{n} m_{l(k)}(2^{-k}\omega)\chi_{[-\pi,\pi]}(2^{-n}\omega), \tag{A.5}$$

$$h_n(\omega) = \prod_{k=1}^{n} |m(2^{-k}\omega)|^{l(k)}\chi_{[-\pi,\pi]}(2^{-n}\omega), \tag{A.6}$$

and

$$g_n(\omega) = \prod_{k=1}^{n} |m(2^{-k}\omega)|\chi_{[-\pi,\pi]}(2^{-n}\omega). \tag{A.7}$$

Now

$$g_n(\omega) = (\mathrm{sinc}(\omega/2))^{\beta}(\mathrm{sinc}(2^{-n-1}\omega))^{-\beta} \prod_{k=1}^{n} |\tilde{m}(2^{-k}\omega)|\chi_{[-\pi,\pi]}(2^{-n}\omega) \tag{A.8}$$

and $\inf_{|\omega|\leq 2^n\pi} |\mathrm{sinc}(2^{-n-1}\omega) = 2/\pi$; hence the same technique used for the estimation of $g(\omega)$ in the proof of Theorem 3.2 yields

$$g_n(\omega) \leq C(1+|\omega|)^{-\varepsilon} \tag{A.9}$$

with $\varepsilon = \beta - (\log_2 \sigma_i)/i > 0$. For a fixed $s > 0$, consider an integer $p > 0$ such that $p\varepsilon \geq s$. Since $l(k)$ goes to $+\infty$, there exists an integer $m \geq 0$ such that $l(k) \geq p$ for all $k > m$. For $n > m$, we can thus estimate $\hat{\varphi}^{[n]}(\omega)$ as follows:

$$\begin{aligned}
|\hat{\varphi}^{[n]}(\omega)| &\leq h_n(\omega)\\
&= \prod_{k=1}^{n} |m(2^{-k}\omega)|^{l(k)}\chi_{[-\pi,\pi]}(2^{-n}\omega)\\
&\leq \prod_{k=m+1}^{n} |m(2^{-k}\omega)|^{l(k)}\chi_{[-\pi,\pi]}(2^{-n}\omega)\\
&\leq \prod_{k=m+1}^{n} |m(2^{-k}\omega)|^{p}\chi_{[-\pi,\pi]}(2^{-n}\omega)\\
&= [g_{n-m}(2^{-m}\omega)]^p\\
&\leq C_s(1+|\omega|)^{-s},
\end{aligned}$$

where $C_s = C^p \sup_\omega((1+|\omega|)/(1+2^{-m}|\omega|))^s$ depends only on s. Since s is arbitrary, we obtain the C^∞ convergence of the subdivision by dominated convergence, as in Theorem 2.2.

This more general result applies in particular to the trigonometric polynomials of §4 since the function $M(\omega)$ of Lemma 4.1 is bounded by 1.

REFERENCES

[1] S. Berkolaiko and I. Novikov (1992), *Sur des presque-ondelettes indéfiniment différentiables à support compact*, Dokl. Acad. Nauk., 326, pp. 935–938.

[2] A. Cavaretta, W. Dahmen, and C. A. Micchelli (1991), *Stationary subdivision*, Mem. Amer. Math. Soc., 93, pp. 1–186.

[3] A. Cohen, I. Daubechies, and P. Vial (1993), *Wavelets and fast wavelet transforms on an interval*, Appl. Comput. Harmonic Anal., 1, pp. 54–81.

[4] W. Dahmen and A. Kunoth (1992), *Multilevel preconditionning*, Numer. Math., 63, pp. 315–345.

[5] I. Daubechies (1988), *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41, pp. 909–996.

[6] ——— (1992), *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia.

[7] I. Daubechies and J. Lagarias (1991), *Two scale difference equations* I: *Existence and global regularity of solutions*, SIAM J. Math. Anal., 22, pp. 1388–1410.

[8] C. de Boor, R. DeVore, and A. Ron (1993), *Approximation from shift-invariant subspaces of* $L^2(\mathbb{R}^d)$, Report CMS TSR 92-2, University of Wisconsin, Madison, WI.

[9] G. Derfel, N. Dyn, and D. Levin (1995), *Generalized functional equations and subdivision processes*, J. Approx. Theory, 80, pp. 272–297.

[10] G. Deslauriers and S. Dubuc (1987), *Interpolation dyadique*, in Fractals, Dimensions Non Entières et Applications, G. Cherbit, ed., Masson, Paris, pp. 44–55.

[11] N. Dyn (1992), *Subdivision schemes in computer-aided geometric design*, in Advances in Numerical Analysis II: Wavelets, Subdivision Algorithms and Radial Functions, W. A. Light, ed., Oxford University Press, Oxford, UK, pp. 36–104.

[12] N. Dyn and D. Levin (1990), *Interpolating subdivision schemes for the generation of curves and surface*, in Multivariate Approximation and Interpolation, W. Haussmann and K. Jetter, eds., Birkhäuser-Verlag, Basel, Switzerland, pp. 91–106.

[13] N. Dyn, J. A. Gregory, and D. Levin (1987), *A four-point interpolatory subdivision scheme for curve design*, Comput. Aided Geom. Design, 4, pp. 257–268.

[14] N. Dyn and A. Ron (1995), *Multiresolution analysis by infinitely differentiable compactly supported functions*, Appl. Comput. Harmonic Anal., 2, pp. 15–20.

[15] M. Frazier and B. Jawerth (1985), *Decomposition of Besov spaces*, Indiana Univ. Math. J., 34, pp. 777–799.

[16] L. Hormander (1983), *The Analysis of Partial Differential Equations*, vol. I, Springer-Verlag, Berlin, New York, Heidelberg.

[17] S. Mandelbrojt (1942), *Analytic functions and classes of infinitely differentiable functions*, Rice Inst. Pamphlet, 29, pp. 1–142.

[18] Y. Meyer (1990), *Ondelettes et Opérateurs*, Hermann, Paris.

[19] A. Ron (1995), *Approximation orders of and approximation maps from local principal shift invariant spaces*, J. Approx. Theory, 81, pp. 38–65.

[20] V. L. Rvachev and V. A. Rvachev (1971), *On a function with compact support*, Dopov. Dokl. Akad. Nauk. Ukraïni, 8, pp. 705–707 (in Ukrainian).

[21] V. A. Rvachev (1990), *Compactly supported solutions of functional-differential equations and their applications*, Russian Math. Surveys, 45, pp. 87–120.

[22] H. Volkmer (1992), *On the regularity of wavelets*, IEEE Trans. Inform. Theory, 38, pp. 872–876.

[23] K. Zao (1995), *Simultaneous approximation from PSI spaces*, J. Approx. Theory, 81, pp. 166–184.

SIAM J. MATH. ANAL.
Vol. 27, No. 6, pp. 1770–1790, November 1996

A NONLINEAR OPERATOR RELATED TO SCALING FUNCTIONS AND WAVELETS*

YING HUANG†

Abstract. This paper studies a certain nonlinear operator T from $L^2(\mathbb{R})$ to itself under which every scaling function is a fixed point. The iterations $T^n f$ of T on any L^2-function f with the Riesz basis property are investigated; they turn out to be the subdivision-scheme iterates of f with weights depending on f only. The paper gives conditions for convergence of $T^n f$ to a limit in different topologies and studies the regularity of the limit functions. The results are illustrated with examples.

Key words. scaling functions, wavelets, the Riesz basis property, subdivision schemes

AMS subject classifications. 26A18, 39A10, 42A38

1. Introduction. Orthonormal bases of wavelets $\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k)$ $(j, k \in \mathbb{Z})$ for $L^2(\mathbb{R})$ have many useful properties [Ru], [Ch], [I1], [I2], [BF]. The construction of such wavelet bases is well understood; every such basis corresponds to a multiresolution analysis characterized by a *scaling function* $\phi(x)$ (see, e.g., [Dau1]). What makes a randomly chosen ϕ a scaling function? There exist many different possible choices for ϕ. We shall see that scaling functions can all be viewed as the fixed points of a nonlinear operator.

In a multiresolution analysis, for appropriately chosen c_n and d_n, the functions ϕ and ψ satisfy,

$$\phi(x) = \sum_n c_n \phi(2x - n), \qquad \psi(x) = \sum_n d_n \phi(2x - n),$$

and $\psi \perp \phi(\bullet - m)$ for all $m \in \mathbb{Z}$. It follows that $\psi \perp \psi(2\bullet - n)$ for all $n \in \mathbb{Z}$. Now define a nonlinear operator, denoted T, which projects any nonzero function in $L^2(\mathbb{R})$ onto the closed subspace spanned by its own scaled translates. More precisely,

$$T : L^2(\mathbb{R}) \to L^2(\mathbb{R}), \quad Tf := \text{orthorgonal projection of } f \text{ onto } \overline{\text{Span}\{f_{-1,n}; n \in \mathbb{Z}\}},$$

where, as usual, $f_{j,k} = 2^{-j/2}f(2^{-j}x - k)$. It is clear that $T\phi = \phi$ and $T\psi = 0$. This paper studies the properties of T.

The operator T is nonlinear: we have $T(\lambda f) = \lambda\, Tf$ for all $\lambda \in \mathbb{C}$, but there is no a priori reason to expect $T(f+g) = Tf+Tg$, and we shall see explicit counterexamples in §3. Because T is defined as an orthogonal projection, we always have

$$\|Tf\| \leq \|f\|. \tag{1.1}$$

As pointed out above, multiresolution scaling functions ϕ are fixed points of T. We can study the iterations $T^n f$ of T on an arbitrary $f \in L^2(\mathbb{R})$. If these $T^n f$ have a nontrivial limit, then this limit is a fixed point of T and is therefore a candidate of a scaling function. It leads to a multiresolution analysis "naturally" associated with f. When does such a nontrivial limit exist? How stable is the procedure under small perturbations of f? This paper contains answers to these questions. A key observation in our analysis is that T is closely related to subdivision schemes.

* Received by the editors June 9, 1994; accepted for publication (in revised form) May 30, 1995.

† Mathematics Department, Rutgers University, New Brunswick, NJ 08903.

The contents of this paper are as follows. We begin with a presentation of the basic properties of T: in §2, we give an explicit formula for Tf for a large class of functions f, and then we discuss the continuity and the fixed point set of T. In §3, we show the connection of $T^n f$ with subdivision schemes. Then §§4 and 5 study sufficient conditions for convergence of T^n (or rescaled iterates of T^n) in different topologies. The final section, §6, consists of examples of functions and their corresponding limit functions, illustrating the results of §§4 and 5.

We conclude this introduction by fixing some notations for the rest of the paper. We normalize the Fourier transform as follows:

$$\hat{f}(\xi) \equiv (f)^\wedge(\xi) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x)\, e^{-i\xi x}\, dx. \tag{1.2}$$

Here the integral converges absolutely for all ξ if f is in $L^1(\mathbb{R})$; for general $f \in L^2(\mathbb{R})$, (1.2) should be understood via the standard limiting process. Throughout this paper, the symbol $\| \bullet \|$ (without a subscript) will be reserved for the $L^2(\mathbb{R})$ norm

$$\|f\|^2 = \langle f, f\rangle = \int_{-\infty}^{+\infty} f(x)\, \overline{f(x)}\, dx.$$

With our normalization of the Fourier transform, the Plancherel identity is $\|f\| = \|\hat{f}\|$.

2. Basic properties of T. In this section, we first introduce the definition of the Riesz basis property for an L^2-function f and obtain an explicit formula for the Fourier transform of Tf. With this, we then study the continuity of T and the set of fixed points of T.

2.1. A formula for T. Since Tf is defined by the orthogonal projection of f onto $\overline{\text{Span}\{f_{-1,n}; n \in \mathbb{Z}\}}$, it is tempting to write

$$Tf(x) = \sum_{n\in\mathbb{Z}} \alpha_n \cdot f(2x - n), \tag{2.1}$$

and to try to determine the α_n's in (2.1). This is not always straightforward, however. First of all, the $f(\bullet - n)$ may not be independent, so that several sequences of coefficients α_n in (2.1) would lead to the same answer. An example is given by $f(x) = 1$ for $0 \le x < 2$, $f(x) = 0$ otherwise. For any $x \in \mathbb{R}$, we have $\sum_{n\in\mathbb{Z}}(-1)^n f(x-n) = 0$; since $Tf = f$, we thus have $Tf(x) = f(x) = f(2x) + f(2x-2) + \gamma \cdot \sum_{n\in\mathbb{Z}}(-1)^n f(2x-n)$, where $\gamma \in \mathbb{C}$ is arbitrary. This nonuniqueness of the α_n's can be circumvented by prescribing an algorithm for determining the α_n's. Since $\overline{\text{Span}\{f_{-1,n}; n \in \mathbb{Z}\}} = \overline{\bigcup_{N\in\mathbb{N}} V_{f,N}}$, where $V_{f,N} = \text{Span}\{f_{-1,n}; |n| \le N\}$, and since any finite set of $f_{-1,n}$ is always independent, we have, in general,

$$Tf(x) = \lim_{N\to+\infty} \text{Proj}_{\;V_{f,N}} f(x) = \lim_{N\to+\infty} \sum_{|n|\le N} a_n^N \cdot f(2x-n),$$

where the a_n^N's are now determined uniquely by solving the linear system

$$\left\langle \sum_{|n|\le N} a_n^N \cdot f(2\bullet - n) - f, f(2\bullet - k) \right\rangle = 0 \quad \forall |k| \le N.$$

If, for every fixed n, the a_n^N's tend to a limit as $N \to +\infty$, we could define this limit as α_n, understanding (2.1) in this limiting sense. In the example above, we have

$a_0^N = 1$, $a_2^N = 1$, and all other $a_n^N = 0$ for all $N \geq 2$. However, for general f, the a_n^N need not tend to a limit. A technical example in which the a_n^N's blow up as $N \to +\infty$ is $f(x) = e^{i(\pi/2)x}(2 - |x|)$ for $|x| \leq 2, f(x) = 0$ otherwise. For details, see [Hu].

In the rest of this paper, we shall restrict ourselves to the (very large) class of functions that do not exhibit such problems. We shall say that a function f has the *Riesz basis property* (abbreviated: f is RBP) if $f \in L^2(\mathbb{R})$ and if the set $\{f(x-k)\}_{k\in\mathbb{Z}}$ is a Riesz basis for its closed linear span, i.e., there exist two constants $0 < C_1 \leq C_2 < +\infty$ such that for all finite linear combinations of the $f(x-k)$, we have

$$2\pi C_1 \sum |c_k|^2 \leq \left\| \sum_{k\in\mathbb{Z}} c_k f(x-k) \right\|^2 \leq 2\pi C_2 \sum |c_k|^2.$$

Using Plancherel's formula, this is equivalent to

$$C_1 \leq \sum_{l\in\mathbb{Z}} |\hat{f}(\xi + 2\pi l)|^2 \leq C_2 \quad \text{a.e. in } \xi. \tag{2.2}$$

When f is RBP, any function g in $\overline{\text{Span}\{f(\bullet - n); n \in \mathbb{Z}\}}$ can be uniquely written as $g(x) = \sum_{n\in\mathbb{Z}} \beta_n \cdot f(x-n)$ with l^2-coefficients β_n. One can therefore always write an expression of type (2.1) for Tf. The coefficients α_n can then be found as follows.

PROPOSITION 2.1. *If f is RBP, then $(Tf)^\wedge(\xi) = a(\frac{\xi}{2})\hat{f}(\frac{\xi}{2})$, where*

$$a(\xi) = \frac{\sum_{l\in\mathbb{Z}} \hat{f}(2\xi + 4\pi l)\overline{\hat{f}(\xi + 2\pi l)}}{\sum_{l\in\mathbb{Z}} |\hat{f}(\xi + 2\pi l)|^2}. \tag{2.3}$$

Proof. We can rewrite (2.1) as either $Tf = \sum_{k\in\mathbb{Z}} a_k\, f_{-1,k}$ for some $a_k = \frac{1}{\sqrt{2}}\alpha_k \in \mathbb{C}$ or $(Tf)^\wedge(\xi) = a(\frac{\xi}{2})\hat{f}(\frac{\xi}{2})$, where $a(\xi) = \frac{1}{\sqrt{2}} \sum_{k\in\mathbb{Z}} a_k\, e^{-i\xi k}$. By the definition of T as an orthogonal projection operator, we have $\langle Tf - f, f_{-1,m}\rangle = 0$ for any $m \in \mathbb{Z}$. Thus

$$\langle f, f_{-1,m}\rangle = \langle Tf, f_{-1,m}\rangle = \sum_k a_k^n \langle f_{-1,k}, f_{-1,m}\rangle.$$

Multiplying both sides of this equation by $e^{-i\xi m}$ and summing over $m \in \mathbb{Z}$, we get

$$\sum_{m\in\mathbb{Z}} \langle f, f_{-1,m}\rangle\, e^{-i\xi m} = \sum_{m,k} a_k \langle f_{-1,k}, f_{-1,m}\rangle e^{-i\xi m} = \sqrt{2}\, a(\xi) \sum_{l\in\mathbb{Z}} \langle f, f_{0,l}\rangle e^{-i\xi l}.$$

Hence

$$a(\xi) = \frac{\sum_{m\in\mathbb{Z}} \langle f, f_{-1,m}\rangle\, e^{-i\xi m}}{\sqrt{2}\sum_{l\in\mathbb{Z}} \langle f, f_{0,l}\rangle e^{-i\xi l}}.$$

On the other hand,

$$\begin{aligned}\langle f, f_{-1,m}\rangle &= \sqrt{2} \int_{-\infty}^{+\infty} \hat{f}(\xi) \frac{1}{2} \overline{\hat{f}\left(\frac{\xi}{2}\right)} e^{i\frac{m}{2}\xi} d\xi \\ &= \sqrt{2} \int_0^{2\pi} e^{im\xi} \left(\sum_{l\in\mathbb{Z}} \hat{f}(2\xi + 4\pi l)\overline{\hat{f}(\xi + 2\pi l)} \right) d\xi.\end{aligned}$$

Then

$$\sum_{m\in\mathbb{Z}} \langle f, f_{-1,m}\rangle e^{-im\xi} = 2\pi\sqrt{2} \sum_{l\in\mathbb{Z}} \hat{f}(2\xi + 4\pi l)\overline{\hat{f}(\xi + 2\pi l)}.$$

Similarly, $\sum_{l\in\mathbb{Z}} \langle f, f_{0,l}\rangle e^{-i\xi l} = 2\pi \sum_{l\in\mathbb{Z}} |\hat{f}(\xi + 2\pi l)|^2$. □

2.2. Continuity of T. We already know from (1.1) that T is bounded. This does not immediately imply that T is continuous because T is nonlinear (see §3.1). Thus we need to prove the following.

THEOREM 2.2. *Assume that f is RBP. For any $\varepsilon > 0$, we can find δ, depending only on ε and f, such that if g is RBP and $\|f - g\| \leq \delta$, then $\|Tf - Tg\| \leq \varepsilon$.*

Remark. Here we require that both f and g are RBP because we want to use (2.3). Note that the Riesz basis property for f and $\|f - g\| \leq \delta$ together do *not* necessarily imply that g is RBP.

Proof of Theorem 2.2. Define $f^{\#}$ via its Fourier transform by

$$(f^{\#})^{\wedge}(\xi) = \frac{\hat{f}(\xi)}{\left(2\pi \sum_{l \in \mathbb{Z}} |\hat{f}(\xi + 2\pi l)|^2\right)^{\frac{1}{2}}}. \tag{2.4}$$

This is well defined a.e. in ξ. Then, again a.e. in ξ,

$$\sum_{l \in \mathbb{Z}} |(f^{\#})^{\wedge}(\xi + 2\pi l)|^2 = \frac{1}{2\pi}, \tag{2.5}$$

or, equivalently, the $\{f^{\#}_{-1,k}; k \in \mathbb{Z}\}$'s constitute an orthonormal basis for V_f. Thus by an argument similar to the one that proved Proposition 2.1, we have for any $h \in L^2(\mathbb{R})$ that $(\mathrm{Proj}_{V_{f^{\#}}} h)^{\wedge}(\xi) = a_{f^{\#},h}(\frac{\xi}{2}) \cdot (f^{\#})^{\wedge}(\frac{\xi}{2})$ with

$$a_{f^{\#},h}(\xi) = 2\pi \sum_{l \in \mathbb{Z}} \hat{h}(2\xi + 4\pi l)(f^{\#})^{\wedge}(\xi + 2\pi l).$$

Thus

$$\begin{aligned} &\|\mathrm{Proj}_{V_{f^{\#}}} f - \mathrm{Proj}_{V_{g^{\#}}} f\| = 2\|a_{f^{\#},f} \cdot (f^{\#})^{\wedge} - a_{g^{\#},f} \cdot (g^{\#})^{\wedge}\| \\ &\leq 2(\|(a_{f^{\#},f} - a_{g^{\#},f})(f^{\#})^{\wedge}\| + \|a_{g^{\#},f}((f^{\#})^{\wedge} - (g^{\#})^{\wedge})\|) := 2(I_1 + I_2). \end{aligned}$$

However,

$$\begin{aligned} I_1^2 &= \int_{|\xi| \leq \pi} \sum_l |(f^{\#})^{\wedge}(\xi + 2\pi l)|^2 |a_{f^{\#},f}(\xi) - a_{g^{\#},f}(\xi)|^2 \, d\xi \\ &= \frac{1}{2\pi} \int_{|\xi| \leq \pi} |a_{f^{\#},f}(\xi) - a_{g^{\#},f}(\xi)|^2 \, d\xi \quad \text{(by (2.5))} \\ &\leq 2\pi \int_{|\xi| \leq \pi} \sum_l |\hat{f}(2\xi + 4\pi l)|^2 \sum_k |\hat{f^{\#}}(\xi + 2\pi k) - \hat{g^{\#}}(\xi + 2\pi k)|^2 d\xi \\ &\leq 2\pi C_{2,f} \|f^{\#} - g^{\#}\|^2 \quad (C_{2,f} \text{ is the upper bound for } f \text{ in (2.2)}). \end{aligned}$$

On the other hand,

$$\begin{aligned} I_2^2 &:= \int_{|\xi| \leq \pi} |a_{g^{\#},f}(\xi)|^2 \sum_l |\hat{f^{\#}}(\xi + 2\pi l) - \hat{g^{\#}}(\xi + 2\pi l)|^2 d\xi \\ &\leq \max_{|\xi| \leq \pi} |a_{g^{\#},f}(\xi)|^2 \|f^{\#} - g^{\#}\|^2. \end{aligned}$$

By (2.2), (2.5), and the Cauchy–Schwarz inequality, $|a_{g^{\#},f}|^2 \leq 2\pi C_{2,f}$. Thus

$$\|\mathrm{Proj}_{V_{f^{\#}}} f - \mathrm{Proj}_{V_{g^{\#}}} f\| \leq 4\sqrt{2\pi C_{2,f}} \|f^{\#} - g^{\#}\|.$$

To control $\|f^{\#} - g^{\#}\|$, we define the sequences $b^f(\xi)$ and $b^g(\xi)$ by $b^f_l(\xi) := \hat{f}(\xi + 2\pi l)$ and $b^g_l(\xi) := \hat{g}(\xi + 2\pi l)$; then $b^f(\xi) \in l^2(\mathbb{Z})$ a.e. in ξ, and $\|b^f(\xi)\|_{l^2} \geq C_{1,f} > 0$, $\|b^g(\xi)\|_{l^2} > 0$. We have

$$\begin{aligned}\|f^{\#} - g^{\#}\|^2 &= \int_{|\xi|\leq\pi} \sum_{l\in\mathbb{Z}} \left| \frac{b^f_l(\xi)}{\|b^f(\xi)\|_{l^2}} - \frac{b^g_l(\xi)}{\|b^g(\xi)\|_{l^2}} \right|^2 d\xi \\ &\leq \int_{|\xi|\leq\pi} \left(\frac{2}{\|b^f(\xi)\|_{l^2}} \|b^f(\xi) - b^g(\xi)\|_{l^2} \right)^2 d\xi \\ &\leq \frac{4}{C_{1,f}} \|f - g\|^2 \quad (C_{1,f} \text{ is the lower bound for } f \text{ in (2.2)}).\end{aligned}$$

Therefore,

$$\begin{aligned}\|Tf - Tg\| &\leq \|\mathrm{Proj}_{V_{f^{\#}}} f - \mathrm{Proj}_{V_{g^{\#}}} f\| + \|\mathrm{Proj}_{V_{g^{\#}}} f - \mathrm{Proj}_{V_{g^{\#}}} g\| \\ &\leq 8\sqrt{2\pi C_{2,f}/C_{1,f}} \cdot \|f - g\| + \|f - g\|.\end{aligned}$$

Thus we just need to choose $\delta = \varepsilon/(8\sqrt{2\pi C_{2,f}/C_{1,f}} + 1)$ to obtain $\|Tf - Tg\| \leq \varepsilon$. □

By using a stronger topology, we can avoid the explicit requirement that g be RBP.

COROLLARY 2.3. *Assume that f is RBP and $\int |f(x)|^2(1+|x|^2)dx < \infty$. For any $\varepsilon > 0$, we can find $\delta > 0$ such that if g satisfies $\int |f(x) - g(x)|^2(1 + |x|^2)dx \leq \delta$, then $\|Tf - Tg\| \leq \varepsilon$.*

Proof. 1. Assume that $\int |f(x) - g(x)|^2(1 + |x|^2)dx \leq \delta_1^2$, where the $\delta_1 > 0$ will be determined later. Then

$$\|f - g\|_{L^1} \leq \left(\int |f(x) - g(x)|^2(1 + |x|^2)dx \right)^{\frac{1}{2}} \left(\int \frac{1}{1 + |x|^2} dx \right)^{\frac{1}{2}} \leq \delta_1\sqrt{\pi}.$$

Hence we also have $|\hat{f}(\xi) - \hat{g}(\xi)| \leq \delta_1/\sqrt{2}$.

2. Similarly, $\|f\|_{L^1} \leq M$ and $\|g\|_{L^1} \leq M + \delta_1/\sqrt{2}$ for some constant $M > 0$ depending only on f. Thus $\hat{f}$ and $\hat{g}$ are continuous and $|\hat{f}(\xi) + \hat{g}(\xi)| \leq 2M + \delta_1/\sqrt{2}$. Consequently, $||\hat{f}|^2 - |\hat{g}|^2| \leq (2M + \delta_1/\sqrt{2})\delta_1/\sqrt{2} \leq M_1\delta_1$ if we choose $\delta_1 \leq M\sqrt{2}$ and let $M_1 = 3M/\sqrt{2}$.

3. We are now ready to prove that

$$\sum_{l\in\mathbb{Z}} |\hat{g}(\xi + 2\pi l)|^2 \geq C_{1,f}/4 \quad \text{for any} \quad \xi \in [-\pi, \pi].$$

(Here $C_{1,f}$ is the lower bound for f in (2.2).) By the same argument (due to A. Cohen) as in part 1 of the proof of Theorem 6.3.1 in [Dau2], the continuity of $\hat{f}$ implies that there exists $L_0(f) \in \mathbb{N}$, such that for all $\xi \in [-\pi, \pi]$, $\sum_{|l|\leq L_0} |\hat{f}(\xi + 2\pi l)|^2 \geq C_{1,f}/2$. Thus, with the choice $\delta_1 := C_{1,f}/(4(2L_0 + 1)M_1)$,

$$\begin{aligned}\sum_{|l|\leq L_0} |\hat{g}(\xi + 2\pi l)|^2 &\geq \sum_{|l|\leq L_0} |\hat{f}(\xi + 2\pi l)|^2 - (2L_0 + 1)M_1\delta_1 \\ &\geq C_{1,f}/2 - (2L_0 + 1)M_1\delta_1 \geq C_{1,f}/4.\end{aligned}$$

4. On the other hand, since $\int |g(x)|^2(1+|x|^2)dx < \infty$, it follows that $\sum |\hat{g}(\xi + 2\pi l)|^2$ is uniformly continuous in $[-\pi, \pi]$. Thus it is bounded above and g is RBP. We can now use same argument as in the proof of Theorem 2.2. □

2.3. A characterization of fixed points of T. It is obvious from the definition of T that any function f with the Riesz basis property which is also a fixed point of T should be a function satisfying the *refinement equation*

$$f(x) = \sqrt{2} \sum_{k \in \mathbb{Z}} a_k\, f(2x-k), \tag{2.6}$$

where the sequence $\{\sqrt{2}a_k\}$ is called the *mask*. Such functions are called *refinable functions* (see [CDM] for a review). Conversely, if an L^2-function f satisfies (2.6), then f will be a fixed point of T. When—that is, for which a_k—does (2.6) have L^2-solutions?

Formula (2.6) implies, at least formally, that

$$\hat{f}(\xi) = \hat{f}(0) \prod_{j=1}^{\infty} a(2^{-j}\xi), \tag{2.7}$$

where again $a(\xi) = \frac{1}{\sqrt{2}} \sum_{k \in \mathbb{Z}} a_k e^{-ik\xi}$. The following standard result gives a sufficient condition guaranteeing pointwise convergence in ξ of the right-hand side of (2.7).

PROPOSITION 2.4. *If $a(\xi)$ is a 2π-periodic function with $a(0) = 1$ and if for some constant c and $\alpha \in (0,1]$, $|a(\xi) - a(0)| \le c|\xi|^\alpha$ for $\xi \to 0$, then*

$$\hat{\phi}(\xi) = \prod_{j=1}^{\infty} a(2^{-j}\xi) \tag{2.8}$$

is well defined for $\xi \in \mathbb{R}$ and $\hat{\phi}$ is continuous in $\mathbb{R}$.

Next, we address L^2-convergence. In [Her1], Hervé gives the following conditions guaranteeing that ϕ is in $L^2(\mathbb{R})$.

THEOREM 2.5 (Hervé). *Let $a(\xi)$ be a 2π-periodic function with $a(0) = 1$ and suppose that*

$$|a(\xi)| = \left|\cos\frac{\xi}{2}\right|^r \cdot |v(\xi)|, \qquad r > 0, \quad v(\pi) \ne 0,$$

where $|v(\xi) - v(0)| \le c|\xi|^\alpha$ for $\xi \to 0$. Let $P_{|v|^2}$ be the operator which acts on any $F \in C[0,\pi]$ via

$$P_{|v|^2} F(\xi) = \left|v\left(\frac{\xi}{2}\right)\right|^2 \cdot F\left(\frac{\xi}{2}\right) + \left|v\left(\frac{\xi}{2}+\pi\right)\right|^2 \cdot F\left(\frac{\xi}{2}+\pi\right)$$

and let β be the spectral radius of $P_{|v|^2}$ restricted to $C[0,2\pi]$. If $\beta < 2^{2r}$, then $\hat{\phi}$ defined as in (2.8) is in $L^2(\mathbb{R})$.

We give a proof for this result in §4 when we discuss the L^2-convergence of $T^n f$.

This criterion becomes practical when one also has an algorithm for computing the spectral radius of the operator $P_{|v|^2}$. In the same paper [Her1], Hervé also derived a formula for computing this spectral radius; see also [CD2].

Proposition 2.4 requires that $|a(\xi) - a(0)| \le c|\xi|^\alpha$ for ξ near 0, and this condition implies decay for the $\{a_k\}$'s. Interestingly enough, there exist refinement equations without decay (even growth) for the $\{a_k\}$'s which nevertheless have L^2-solutions.

Example 2.6. Consider $f(x) = (1 - \frac{1}{2}|x|) \cdot \chi_{|x| \leq 2}$. Then it is easy to check that $\sum_n 2nf(x-2n) = x$ and $\sum_n (2n - \frac{1}{2}) f(x - 2n + 1) = x$. Thus

$$f(x) = \frac{1}{2} f(2x+2) + f(2x) + \frac{1}{2} f(2x-2) \\ + \sum_n 2nf(2x-2n) - \sum_n \left(2n - \frac{1}{2}\right) f(2x - 2n + 1).$$

Here $|a_k| \sim |k|$, while f is in $L^2(\mathbb{R})$. Of course, this example can be generalized to higher-order B-splines, leading to examples with $|a_k| \sim |k|^l$. Note that these examples also show that a refinable function can be the solution of several different refinement equations.

3. General formula for T^n. It follows from (1.1) that $\|T^{n+1}f\| \leq \|T^n f\| \leq \|f\|$ for all $n \geq 1$. By the weak compactness of the unit ball, we can therefore expect at least weak accumulation points for the sequence $\{T^n f\}_{n \in \mathbb{N}}$. When is this accumulation point unique and nontrivial? Also, when do we have convergence in the L^2-norm as well? We look at an example in §3.1. In §3.2, we derive a general formula for $(Tf)^\wedge$, which will allow us to answer these questions in §4.

3.1. A linear combination of a wavelet and its scaling function. The different orthonormality relations make this example particularly easy. Noting that $T\phi = \phi, T\psi = 0$, we study the action of T on a linear combination of ψ and ϕ:

PROPOSITION 3.1. *Let ϕ, respectively, ψ, be the scaling function, respectively, wavelet, associated with a multiresolution analysis with $|\phi(x)| \leq C(1+|x|)^{-1}$. Define $f := \alpha\phi + \beta\psi$ with $|\alpha|^2 + |\beta|^2 = 1$. Then $\lim_{N\to\infty}(T^N f)^\wedge(\xi) = 0$ a.e. in ξ unless $\beta = 0$, in which case $f = \alpha\phi$ and $T^N f = f$ for all $N \in \mathbb{N}$. Moreover, $T^N f$ converges to 0 weakly in L^2 as well.*

Before starting the proof, let us review some "standard" notations and properties of wavelet analysis. For more details, we refer to [Dau2].

Let $m_0(\xi) = \frac{1}{\sqrt{2}} \sum_{n\in\mathbb{Z}} h_n e^{-in\xi}$ and $m_1(\xi) = \frac{1}{\sqrt{2}} \sum_{n\in\mathbb{Z}} g_n e^{-in\xi}$ with $h_n = \langle \phi, \phi_{-1,n} \rangle$ and $g_n = \langle \psi, \phi_{-1,n} \rangle$. In the following proof, we will use the fact that

$$\text{(3.1)} \qquad \sum_{n\in\mathbb{Z}} h_n \overline{h_{n+2k}} = \delta_{k,0}, \qquad \sum_{n\in\mathbb{Z}} h_n \overline{g_{n+2k}} = 0, \qquad \sum_{n\in\mathbb{Z}} g_n \overline{g_{n+2k}} = \delta_{k,0},$$

$m_0(0) = 1$, and $m_1(0) = 0$. Note that the decay of ϕ ensures that the h_n's are absolutely summable.

Proof of Proposition 3.1. If $\beta = 0$, we are done. We now assume $\beta \neq 0$, so $|\alpha| < 1$. Since the functions $f_{-1,n} = \alpha\phi_{-1,n} + \beta\psi_{-1,n}$ are orthonormal, we have

$$\begin{aligned} f^1 = Tf &= \sum_m \langle \alpha\phi + \beta\psi, \alpha\phi_{-1,m} + \beta\psi_{-1,m} \rangle (\alpha\phi_{-1,m} + \beta\psi_{-1,m}) \\ &= \sum_m (|\alpha|^2 h_m + \overline{\alpha}\beta g_m)(\alpha\phi_{-1,m} + \beta\psi_{-1,m}), \\ f^1(\bullet - m) &= \sum_k (|\alpha|^2 h_k + \overline{\alpha}\beta g_k)(\alpha\phi_{-1,2m+k} + \beta\psi_{-1,2m+k}), \\ \langle f^1, f^1(\bullet - m) \rangle &= \sum_{n,k} (|\alpha|^2 h_n + \overline{\alpha}\beta\, g_n)(|\alpha|^2\, \overline{h_k} + \alpha\overline{\beta}\ \overline{g_k})(|\alpha|^2 + |\beta|^2)\delta_{n,2m+k} \\ &= |\alpha|^2\, \delta_{m,0}. \end{aligned}$$

It follows that T^2f, the projection of $f^1 = Tf$ onto $\overline{\operatorname{Span}\{f^1(2\bullet-n); n \in \mathbb{Z}\}}$, is given by

$$f^2 = T^2 f = |\alpha|^{-2} \sum_m \langle f^1, f^1_{-1,m}\rangle f^1_{-1,m} = \sum_m (|\alpha|^2\, h_m + \overline{\alpha}\beta\, g_m) f^1_{-1,m}.$$

By induction, we then obtain, for all $N \in \mathbb{N}$,

$$\begin{aligned} f^{N+1} &= T^{N+1} f = \sum_m (|\alpha|^2\, h_m + \overline{\alpha}\beta\, g_m) f^N_{-1,m}, \\ \langle f^N, f^N_{-1,m}\rangle &= |\alpha|^{2N}(|\alpha|^2\, h_m + \overline{\alpha}\beta\, g_m), \\ \langle f^N, f^N(\bullet - m)\rangle &= |\alpha|^{2N}\, \delta_{m,0}. \end{aligned} \tag{3.2}$$

Formula (3.2) can be rewritten as

$$\begin{aligned} (f^{N+1})^\wedge(\xi) &= \sum_m (|\alpha|^2 h_m + \overline{\alpha}\beta\, g_m)\sqrt{2}\,(f^N)^\wedge\left(\frac{\xi}{2}\right)\frac{1}{2}e^{-i\frac{m}{2}\xi} \\ &= \left(|\alpha|^2\, m_0\left(\frac{\xi}{2}\right) + \overline{\alpha}\beta\, m_1\left(\frac{\xi}{2}\right)\right)(f^N)^\wedge\left(\frac{\xi}{2}\right) \\ &= \overline{\alpha}^{N+1}\prod_{j=1}^{N+1}(\alpha\, m_0(2^{-j}\xi) + \beta\, m_1(2^{-j}\xi))\hat{f}(2^{-N-1}\xi). \end{aligned} \tag{3.3}$$

Since $\hat{f}(0) = \alpha$ and the h_n's and g_n's are absolutely summable, $[\alpha\, m_0(2^{-j}\xi) + \beta\, m_1(2^{-j}\xi)]$ goes to α pointwise in ξ as $j \to \infty$ with $|\alpha| < 1$, thus the limit of (3.3) for $N \to \infty$ exists, and

$$\lim_{N\to\infty}(T^N f)^\wedge(\xi) = 0 \quad \text{a.e. in } \xi.$$

Now observing that $|\alpha\, m_0(2^{-j}\xi) + \beta\, m_1(2^{-j}\xi)| \le 1$, we have, again from (3.3), that

$$|(T^N f)^\wedge(\xi)| \le |\hat{f}(2^{-N-1}\xi)| \le \frac{1}{\sqrt{2\pi}}\|f\|_{L^1} \le \frac{1}{\sqrt{2\pi}}(|\alpha| + |\beta|)\|\phi\|_{L^1}.$$

Therefore, for the dense set in $L^2(\mathbb{R})$ of u with Fourier transform $\hat{u} \in L^2(\mathbb{R})\bigcap L^1(\mathbb{R})$, we can use Lebesgue's dominated-convergence theorem to get

$$\lim_{N\to\infty}\langle T^N f, u\rangle = \lim_{N\to\infty}\langle (T^N f)^\wedge, \hat{u}\rangle = \left\langle \lim_{N\to\infty}(T^N f)^\wedge, \hat{u}\right\rangle = 0.$$

Since the $T^N f$'s are also uniformly bounded, this implies the weak convergence of $T^N f$ to 0. □

Actually, we shall show below (in Example 4.7) that the convergence holds in the L^2-norm as well. Note that the f^N's are, in fact, linear combinations of *wavelet packets* as introduced by Coifman and Meyer (for a discussion, see, e.g., [CMW]). Note also that this proposition demonstrates the nonlinearity of T since $\alpha\phi = \alpha T\phi + \beta T\psi \neq T(\alpha\phi + \beta\psi)$.

COROLLARY 3.2. *Iterating T on f does* not *necessarily lead to the closest scaling function (in the L^2-sense) $\tilde{\phi}$ to f.*

Proof. Take f as above, $f = \alpha\,\phi + \beta\,\psi$ with $\beta \neq 0$. Then the distance of f to the family of fixed points of T satisfies $\inf\{\|f - \tilde{\phi}\|; T\tilde{\phi} = \tilde{\phi}\} \le \|f - \phi\| = [2 - 2\Re(\alpha)]^{1/2}$. However, by Proposition 3.1, $T^n f \to 0$ so that $\|f - T^n f\| \to 1 > \sqrt{2 - 2\Re(\alpha)}$ if $\Re(\alpha) > \frac{1}{2}$. □

3.2. The general case: Link with subdivision schemes. In our previous examples, $T^n f \to 0$, at least in the distributional sense, unless f is a scaling function, in which case $Tf = f$. Suitable dilations of scaling functions also give fixed points of T. For example, if $f(x) = \chi_{[0,m)}(x)$, where m is an integer ≥ 2, then $f(x) = f(2x) + f(2x - m)$ a.e., so $Tf = f$. On the other hand, if we take $f(x) = \chi_{[0,1/m)}(x)$ with $m \geq 2$, then $T^n f(x) = \chi_{[0,2^{-n}/m)}(x) \to 0$ a.e. in x as $n \to \infty$. What happens to other dilations, such as $f(x) = \chi_{[0,1+\epsilon)}(x), 0 < \epsilon < 1$? Also, more generally, what happens to arbitrary (but nice) functions, such as the Gaussian $e^{-x^2/2}$?

To answer these questions, we need to derive an explicit formula for $T^n f$ valid for general f. We will use formula (2.3) again, not only to compute Tf, but also for the iterates $T^n f$. This means that we shall need the Riesz basis property for f and all the $T^n f$'s. Fortunately, Tf inherits the Riesz basis property from f under some mild additional assumptions on f.

PROPOSITION 3.3. *If f is RBP with continuous $\hat{f}$ satisfying*

$$|\hat{f}(\xi)| \leq C\,(1 + |\xi|)^{-\frac{1}{2}-\epsilon}$$

for some $\epsilon, C > 0$, and if $a(\xi)$, defined by (2.3), *satisfies*

$$|a(\xi)|^2 + |a(\xi + \pi)|^2 > 0, \tag{3.4}$$

then Tf is RBP; *moreover, Tf inherits the other properties of f.*

Proof. 1. We begin by showing that for some $c_1, c_2 > 0$,

$$c_1 \leq |a(\xi)|^2 + |a(\xi + \pi)|^2 \leq c_2. \tag{3.5}$$

By the decay of $\hat{f}$, we can make $\sum_{|l|\geq L} |\hat{f}(\xi + 2\pi l)|^2$ arbitrarily small by choosing the positive integer L large enough. Then the finite sum $\sum_{|l|<L} |\hat{f}(\xi + 2\pi l)|^2$ is continuous because of the continuity of $\hat{f}$. It follows that $\sum_{l\in\mathbb{Z}} |\hat{f}(\xi + 2\pi l)|^2$ is continuous. Similarly, one can prove the continuity of $\sum_{l\in\mathbb{Z}} \hat{f}(2\xi + 4\pi l)\overline{\hat{f}(\xi + 2\pi l)}$. Therefore, the quotient $a(\xi)$ is also continuous. Thus (3.4) implies (3.5).

2. We now prove that Tf is RBP. Using Proposition 2.1, it is easy to check that

$$\begin{aligned}\sum_{l\in\mathbb{Z}} |(Tf)^\wedge(2\xi + 2\pi l)|^2 &= \sum_{l} |a(\xi + \pi l)\hat{f}(\xi + \pi l)|^2 \\ &= |a(\xi)|^2 \sum_{k} |\hat{f}(\xi + 2\pi k)|^2 + |a(\xi + \pi)|^2 \sum_{k} |\hat{f}(\xi + \pi + 2\pi k)|^2,\end{aligned}$$

which implies $0 < c_1 C_{1,f} \leq \sum_l |\hat{T}f(2\xi + 2\pi l)|^2 \leq c_2 C_{2,f} < +\infty$.

3. Since $(Tf)^\wedge(\xi) = a(\xi/2)\hat{f}(\xi/2)$, the continuity of $(Tf)^\wedge$ is obvious. By the Cauchy–Schwarz inequality and (2.3), $|a(\xi)| \leq C_2/C_1$. Therefore, the decay of $\hat{f}$ implies the same type of decay for $(Tf)^\wedge$.

4. Next, we compute $a^1(\xi)$. Now that Tf is RBP, we know, similarly to the case where we obtain a formula for Tf, that $f^2 := T^2 f = \sum_{k\in\mathbb{Z}} a_k^1 f_{-1,k}^1$, or

$$(f^2)^\wedge(\xi) = a^1\left(\frac{\xi}{2}\right)(f^1)^\wedge\left(\frac{\xi}{2}\right)$$

with $a^1(\xi) = \frac{1}{\sqrt{2}} \sum_{k\in\mathbb{Z}} a_k^1 \, e^{-i\xi k}$ defined by replacing $\hat{f}$ by $(Tf)^\wedge$ in (2.3). Furthermore, with

$$\alpha^n(\xi) := \sum_{k\in\mathbb{Z}} \langle f^n, f_{-1,k}^n \rangle \, e^{-i\xi k}, \qquad \beta^n(\xi) := \sum_{k\in\mathbb{Z}} \langle f^n, f_{0,k}^n \rangle \, e^{-i\xi k},$$

we have, for $n = 0, 1$ (we let $a^0 = a$ and $f^0 = f$),

$$a^n(\xi) = \frac{\alpha^n(\xi)}{\sqrt{2}\beta^n(\xi)}.$$

We now show the surprising result that $a^1 = a^0$. In fact, we have

$$\begin{aligned}\beta^1(2\xi) &= \sum_k \langle f^1, f^1_{0,k}\rangle\, e^{-2ki\xi} = \sum_k \sum_{m,l} a^0_m a^0_l \langle f^0_{-1,m}, f^0_{-1,l+2k}\rangle\, e^{-2ki\xi} \\ &= \sum_{k,m,l} a^0_m e^{-im\xi} a^0_l e^{il\xi} \beta^0_{l+2k-m}\, e^{-i(l+2k-m)\xi} \\ &= |a^0(\xi)|^2\beta^0(\xi) + |a^0(\xi+\pi)|^2\beta^0(\xi+\pi).\end{aligned}$$

Similarly, $\alpha^1(2\xi) = a^0(2\xi)\overline{a^0(\xi)}\alpha^0(\xi) + a^0(2\xi)\overline{a^0(\xi+\pi)}\alpha^0(\xi+\pi)$. Hence

$$\begin{aligned}a^1(2\xi) = \frac{\alpha^1(2\xi)}{\sqrt{2}\beta^1(2\xi)} &= \frac{a^0(2\xi)\overline{a^0(\xi)}\alpha^0(\xi) + a^0(2\xi)\overline{a^0(\xi+\pi)}\alpha^0(\xi+\pi)}{\sqrt{2}[|a^0(\xi)|^2\beta^0(\xi) + |a^0(\xi+\pi)|^2\beta^0(\xi+\pi)]} \\ &= \frac{a^0(2\xi)\left\{\overline{a^0(\xi)}\frac{\alpha^0(\xi)}{\sqrt{2}\beta^0(\xi)}\beta^0(\xi) + \overline{a^0(\xi+\pi)}\frac{\alpha^0(\xi+\pi)}{\sqrt{2}\beta^0(\xi+\pi)}\beta^0(\xi+\pi)\right\}}{|a^0(\xi)|^2\beta^0(\xi) + |a^0(\xi+\pi)|^2\beta^0(\xi+\pi)} = a^0(2\xi).\end{aligned}$$

Thus $a^1(\xi)$ also satisfies $|a^1(\xi)|^2 + |a^1(\xi+\pi)|^2 > 0$. □

This proposition motivates the following definition.

DEFINITION 3.4. *We say that f is T-*amenable *if f is RBP, if $\hat{f}$ is continuous and satisfies $|\hat{f}(\xi)| \le C(1+|\xi|)^{-(1/2)-\epsilon}$ for some $\epsilon, C > 0$, and if $a(\xi)$, defined by* (2.3)*, satisfies $|a(\xi)|^2 + |a(\xi+\pi)|^2 > 0$.*

Proposition 3.3 can now simply be rephrased as follows: if f is T-amenable, then so is Tf. By induction, all the successive iterates $T^n f$ will then also be T-amenable; in particular, they will all be RBP. It follows that for some sequence $\{a^n_k\}_{k\in\mathbb{Z}} \in l^2$, $f^{n+1} = T^{n+1}f = \sum_{k\in\mathbb{Z}} a^n_k\, f^n_{-1,k}$; an exact formula for $a^n(\xi) = \frac{1}{\sqrt{2}}\sum_{k\in\mathbb{Z}} a^n_k\, e^{-i\xi k}$ can be found as was done for $a^1(\xi)$. By the same computation that established $a^1 = a^0$, we have $a^n = a^{n-1}$; hence $a^n = a^0$ for all $n \in \mathbb{N}$. Since there will be no confusion hereafter, we will drop the superscript 0 in $a^0(\xi)$. Sometimes we will add a subscript, $a_f(\xi)$, to emphasize the dependence on the original function f.

This observation shows that for T-amenable f, the $T^n f$'s are merely the subdivision-scheme iterates of f with the weights $\{a_k\}$.

THEOREM 3.5. *If f is T-amenable, then the $T^n f$'s are given by*

$$(T^n f)^\wedge(\xi) = \left[\prod_{j=1}^{n} a(2^{-j}\xi)\right] \hat{f}(2^{-n}\xi). \tag{3.6}$$

Formula (3.6) implies, at least formally, that

$$\hat{f}^\infty(\xi) := \lim_{n\to+\infty} (T^n f)^\wedge(\xi) = \hat{f}(0)\prod_{j=1}^{\infty} a(2^{-j}\xi),$$

which can also be rewritten as $\hat{f}^\infty(\xi) = a(\xi/2)\hat{f}^\infty(\xi/2)$, or

$$f^\infty(x) = \sqrt{2}\sum_{k\in\mathbb{Z}} a_k\, f^\infty(2x-k).$$

We have therefore reduced the problem of studying the limit of the $T^n f$ to finding the solution of a refinement equation (also called a dilation equation or a two-scale difference equation) with filter $a(\xi)$. It is not surprising that this limit f^∞ should satisfy a refinement equation since f^∞ is a fixed point of T; it is surprising that the corresponding mask should be given by $a(\xi)$. There exists considerable literature on refinement equations and refinable functions. See, e.g., [CDM], [DL1], [DL2], [Her1], and the papers cited therein for a detailed study with applications to the construction of wavelets and to subdivision schemes in computer-aided geometric design. In the next section, we shall use some of these results and see how they connect with our operator.

4. Sufficient conditions for convergence of T^n. Since the functions $T^n f$ are given by a subdivision scheme, convergence questions boil down to studying the convergence of subdivision schemes, which is studied in the literature on refinable functions. Many of the articles on refinement equations concern the case where $a(\xi)$ is a trigonometric polynomial. Here $a(\xi)$ is usually not a trigonometric polynomial, and we will be particularly interested in the results obtained by Hervé [Her1]–[Her3] which also apply to nonpolynomial $a(\xi)$'s. Other recent results on convergence of the cascade algorithm are in [Du] (mostly the finite case) and in [Her4] and [CD2], which use a different approach. Related results (for another application) can also be found in [DH]. In this section, we shall state and prove several convergence theorems for our iterates $T^n f$. In particular, we study pointwise convergence of $(T^n f)^\wedge(\xi)$, L^2-convergence of $T^n f$, and pointwise convergence of $(T^n f)(x)$.

THEOREM 4.1 (pointwise convergence of $(T^n f)^\wedge$). *Assume that f is T-amenable and that $\hat{f}$ satisfies $\hat{f}(2\pi l) = \delta_{l,0}\hat{f}(0)$ with $\hat{f}(0) \neq 0$. Assume moreover that*

$$|a_f(\xi) - a_f(0)| \leq c|\xi|^\alpha \quad as \quad \xi \to 0, \tag{4.1}$$

where $a_f(\xi)$ is defined as in (2.3). Then $(T^n f)^\wedge$ converges pointwise to a nontrivial continuous limit $\hat{f}(0)\hat{\phi}$ with $\hat{\phi}$ defined as in (2.8).

Note that (4.1) is easy to satisfy. It suffices, e.g., that f has compact support or that $\sum_l |\hat{f}(\xi + 2\pi l) - \hat{f}(2\pi l)|^2 \leq C|\xi|^{2\alpha}$ for sufficiently small ξ.

Theorem 4.1 follows immediately from Proposition 2.4 and the following lemma.

LEMMA 4.2. *For a T-amenable function f, the following are equivalent:*

(1) $a_f(0) = 1$;

(2) $\hat{f}(2\pi l) = \delta_{l,0}\hat{f}(0)$ *with* $\hat{f}(0) \neq 0$, $l \in \mathbb{Z}$;

(3) $\sum_{k\in\mathbb{Z}} f(x-k) =$ *nonzero constant function.*

Proof. 1. We start by proving that (1) ⇒ (2). Assume that $a_f(0) = 1$ or, equivalently (by (2.3)), that $\sum_{l\in\mathbb{Z}} \hat{f}(4\pi l)\overline{\hat{f}(2\pi l)} = \sum_{l\in\mathbb{Z}} |\hat{f}(2\pi l)|^2$. By the Cauchy–Schwarz inequality, this implies that $(\sum_{l\in\mathbb{Z}} |\hat{f}(2\pi l)|^2)^{1/2} \leq (\sum_{l\in\mathbb{Z}} |\hat{f}(4\pi l)|^2)^{1/2}$. However, this can hold only if

$$\begin{cases} \hat{f}((2k+1)2\pi) = 0, & k \in \mathbb{Z}, \\ \hat{f}(4\pi l) = C\hat{f}(2\pi l) & \text{for some constant } C. \end{cases}$$

Since for $k \neq 0$, $k = 2^l(2m+1)$ for some $l \geq 0$ and $m \in \mathbb{Z}$, we find that

$$\hat{f}(2k\pi) = \hat{f}(2 \cdot 2^l(2m+1)\pi) = C^l\hat{f}((2m+1)2\pi) = 0.$$

2. (2) ⇒ (1) is obvious, and (2) ⇔ (3) can be checked by Poisson's summation formula. □

Note that if $a(0) = 1$, then this lemma implies that $a(\pi) = 0$. (We will use this below.) Also note that if $a(\xi)$ is continuous, then the infinite product in (2.8) makes sense only if $a(0) = 1$. Thus (1)–(3) are necessary for the right-hand side of (2.8) to be nontrivial.

Thus far, we have looked only at the weak convergence of $T^n f$ or pointwise convergence of $(T^n f)^\wedge(\xi)$. Before we can speak of convergence in L^2-norm, we need to establish conditions guaranteeing that ϕ is in $L^2(\mathbb{R})$. Note that by using Plancherel's formula and the Poisson summation formula, we can derive, for all $f \in L^2(\mathbb{R}) \bigcap L^1(\mathbb{R})$,

$$\begin{aligned} \|T^n f\|^2 &= \int_{-\infty}^{+\infty} |\hat{f}(2^{-n}\xi)|^2 \prod_{j=1}^{n} |a(2^{-j}\xi)|^2 \, d\xi \\ &= \int_{|\xi| \le 2^n \pi} F(2^{-n}\xi) \prod_{j=1}^{n} |a(2^{-j}\xi)|^2 \, d\xi, \end{aligned} \tag{4.2}$$

where $F(\xi) = \sum_{l \in \mathbb{Z}} |\hat{f}(\xi + 2\pi l)|^2$ is 2π-periodic. In all cases of interest to us, f will have exponential decay or even compact support so that (4.1) is automatically satisfied, and F is continuous. We shall assume this continuity in all that follows.

The right-hand side of formula (4.3) can be computed by using properties of a positive linear operator associated with the continuous 2π-periodic function $a(\xi)$. This operator, which we denoted $P_{|a|^2}$ in §2, is usually called the *Perron–Frobenius operator* or *transfer operator*. Recall that it acts on a continuous 2π-periodic function F as follows:

$$P_{|a|^2} F(\xi) = \left| a\left(\frac{\xi}{2}\right) \right|^2 \cdot F\left(\frac{\xi}{2}\right) + \left| a\left(\frac{\xi}{2} + \pi\right) \right|^2 \cdot F\left(\frac{\xi}{2} + \pi\right).$$

This operator often appears in the study of orthonormal wavelets and in the estimation of the regularity of the scaling functions associated with $a(\xi)$; see, e.g., [CR], [La], [CD1], [CD2], [Vi], [CDM], [Ei], [Gr], and [Her1]–[Her4]. Its connection with (4.3) is given by the following lemma.

LEMMA 4.3. *For any continuous 2π-periodic function F and all $n \in \mathbb{N}$,*

$$\int_{-\pi}^{\pi} P_{|a|^2}^n F(\xi) \, d\xi = \int_{|\xi| \le 2^n \pi} F(2^{-n}\xi) \prod_{j=1}^{n} |a(2^{-j}\xi)|^2 \, d\xi.$$

This lemma (proved by induction; see any of the references above) shows the relation, stated in Theorem 2.5, between L^2-estimates as in (4.3) and the spectral radius of $P_{|a|^2}$. We now give the proof of Theorem 2.5, borrowed from [Co]; see also [Her1]. We include it here because the same technique also proves the technical Lemma 4.4 below that we shall use extensively further on.

Proof of Theorem 2.5. It suffices to prove that

$$\sum_{n \ge 0} \int_{2^{n-1}\pi \le |\xi| \le 2^n \pi} \prod_{j=1}^{\infty} |a(2^{-j}\xi)|^2 \, d\xi < +\infty.$$

Now using the classical formulas

$$\prod_{j=1}^{n} \cos \frac{\xi}{2^{j+1}} = \frac{\sin \frac{\xi}{2}}{2^n \sin(2^{-n-1}\xi)}, \qquad \prod_{j=1}^{\infty} \cos \frac{\xi}{2^{j+1}} = \frac{2 \sin \frac{\xi}{2}}{\xi},$$

we find that

$$\begin{aligned}
I_n &:= \int_{2^{n-1}\pi\le|\xi|\le 2^n\pi} \prod_{j=1}^{\infty} |a(2^{-j}\xi)|^2\, d\xi \\
&= \int_{2^{n-1}\pi\le|\xi|\le 2^n\pi} \left|\frac{2\sin\frac{\xi}{2}}{\xi}\right|^{2r} \prod_{j=1}^{\infty} |v(2^{-j}\xi)|^2\, d\xi \\
&\le C(2^{n-1}\pi)^{-2r} \int_{2^{n-1}\pi\le|\xi|\le 2^n\pi} \prod_{j=1}^{\infty} |v(2^{-j}\xi)|^2\, d\xi.
\end{aligned}$$

Note that if we let $h(\xi) := \prod_{j=1}^{\infty} |v(2^{-j}\xi)|^2 = \prod_{j=1}^{n} |v(2^{-j}\xi)|^2\, h(2^{-n}\xi)$, then because v—and hence $|v|^2$—has Hölder exponent α in 0, we know by Proposition 2.4 that h is continuous. It follows that for $2^{n-1}\pi \le |\xi| \le 2^n\pi$, $|h(2^{-n}\xi)| \le \sup_{\pi/2\le|\zeta|\le\pi} |h(\zeta)| = C$, implying $|h(\xi)| \le C\prod_{j=1}^{n} |v(2^{-j}\xi)|^2$. Hence

$$\begin{aligned}
I_n &\le C(r)\, 2^{-2rn} \int_{2^{n-1}\pi\le|\xi|\le 2^n\pi} \prod_{j=1}^{n} |v(2^{-j}\xi)|^2 d\xi \\
&\le C(r) 2^{-2rn} \int_{|\xi|\le 2^n\pi} \prod_{j=1}^{n} |v(2^{-j}\xi)|^2 d\xi \\
&= C(r)\, 2^{-2rn} \int_{-\pi}^{\pi} P_{|v|^2}^n 1(\xi) d\xi \le C(r)\, 2^{-2rn}(\beta+\epsilon)^n = C(r)\left(\frac{\beta+\epsilon}{2^{2r}}\right)^n.
\end{aligned}$$

Since $\beta < 2^{2r}$, we can choose $\epsilon > 0$ such that $(\beta+\epsilon)/2^{2r} < 1$, so $\sum_{n\ge 0} I_n < +\infty$. □

LEMMA 4.4. *Under the same assumptions as in Theorem* 2.5, *we have, for any* $0 < \alpha \le \pi$,

$$\lim_{n\to\infty} \int_{2^n\alpha\le|\xi|\le 2^n\pi} \prod_{j=1}^{n} |a(2^{-j}\xi)|^2\, d\xi = 0. \tag{4.3}$$

Proof. The only difference between the integral in (4.3) and I_n is that the product now has only n factors instead of infinitely many. We use the estimate, for $2^n\alpha \le |\xi| \le 2^n\pi$,

$$\prod_{j=1}^{n} \left|\cos\frac{\xi}{2^{j+1}}\right|^{2r} = \left|\frac{2\sin\frac{\xi}{2}}{\xi}\right|^{2r} \left|\frac{2^{-n-1}\xi}{\sin(\frac{\xi}{2^{n+1}})}\right|^{2r} \le C(\alpha, r)\, 2^{-2rn}.$$

The rest of the argument is unchanged. □

We are now ready to state the main result of this section.

THEOREM 4.5 (L^2-convergence of $T^n f$). *Assume that f is T-amenable and that f satisfies $\hat{f}(2\pi l) = \delta_{l,0}\hat{f}(0)$. Assume moreover that $|a_f(\xi)| = |\cos\frac{\xi}{2}|^r \cdot |v(\xi)|$ and $r > 0$ with $v(\pi) \ne 0$, where $v(\xi)$ has Hölder exponent α at 0. Let β be the spectral radius of $P_{|v|^2}$ restricted to $C[0, 2\pi]$. If $\beta < 2^{2r}$, then $T^n f$ converges in $L^2(\mathbb{R})$ to $\hat{f}(0)\phi$, where ϕ is defined as in* (2.8).

Proof. Our proof mimics that of Theorem 3.3 in [CD]; our extra condition of the decay of $\hat{f}(\xi)$ is needed here since we do not just consider truncated versions of $T^n f$ as in [CD].

1. By Lemma 4.2, Proposition 2.4, and Theorem 2.5, we know that the assumptions imply that $\hat{\phi}$ is continuous and in $L^2(\mathbb{R})$.

2. Since $\hat{f}$ and $\hat{\phi}$ are continuous in 0 and since $\hat{\phi}(0) = 1$, there exists an $\alpha \in (0, \pi]$ such that

$$|\xi| \le \alpha \Rightarrow |\hat{f}(\xi)| \le C|\hat{\phi}(\xi)| \quad \text{and} \quad |\hat{\phi}(\xi)| \ge C' > 0. \tag{4.4}$$

We now divide $T^n f$ into three parts: $T^n f = \phi_n^1 + \phi_n^2 + \phi_n^3$ with

$$\begin{aligned}
\hat{\phi}_n^1(\xi) &:= (T^n f)^\wedge(\xi)\chi_{|\xi|\le 2^n\alpha}(\xi),\\
\hat{\phi}_n^2(\xi) &:= (T^n f)^\wedge(\xi)[\chi_{|\xi|\le 2^n\pi}(\xi) - \chi_{|\xi|\le 2^n\alpha}(\xi)],\\
\hat{\phi}_n^3(\xi) &:= (T^n f)^\wedge(\xi)\chi_{|\xi|> 2^n\pi}(\xi).
\end{aligned}$$

3. Clearly, $\hat{\phi}_n^1$ converges pointwise to $\hat{f}(0)\hat{\phi}$; by (4.4), $|\hat{\phi}_n^1|$ is dominated by $|\phi|$, $|\hat{\phi}_n^1(\xi)| \le C|\hat{\phi}(2^{-n}\xi)| \cdot \prod_{j=1}^n |a(2^{-j}\xi)| \le C|\hat{\phi}(\xi)|$, which implies the L^2-convergence of ϕ_n^1 to $\hat{f}(0)\phi$.

4. For $\alpha \le |\xi| \le \pi$, $\hat{f}(\xi)$ is bounded above so that

$$\begin{aligned}
\int_{-\infty}^{+\infty} |\hat{\phi}_n^2|^2 d\xi &= \int_{2^n\alpha\le|\xi|\le 2^n\pi} \prod_{j=1}^n |a(2^{-j}\xi)|^2 |\hat{f}(2^{-n}\xi)|^2 \, d\xi\\
&\le C \int_{2^n\alpha\le|\xi|\le 2^n\pi} \prod_{j=1}^n |a(2^{-j}\xi)|^2 \, d\xi.
\end{aligned} \tag{4.5}$$

By Lemma 4.4, (4.5) tends to 0 as $n \to \infty$, i.e., $\phi_n^2 \to 0$ in $L^2(\mathbb{R})$.

5. By formula (4.3),

$$\begin{aligned}
\int_{-\infty}^{+\infty} |\hat{\phi}_n^3|^2 d\xi &= \int_{|\xi|>2^n\pi} \prod_{j=1}^n |a(2^{-j}\xi)|^2 |\hat{f}(2^{-n}\xi)|^2 \, d\xi\\
&= \int_{-\infty}^{+\infty} \prod_{j=1}^n |a(2^{-j}\xi)|^2 |\hat{f}(2^{-n}\xi)|^2 d\xi - \int_{|\xi|\le 2^n\pi} \prod_{j=1}^n |a(2^{-j}\xi)|^2 |\hat{f}(2^{-n}\xi)|^2 \, d\xi\\
&= \int_{|\xi|\le 2^n\pi} \prod_{j=1}^n |a(2^{-j}\xi)|^2 \sum_{l\ne 0} |\hat{f}(2^{-n}\xi + 2l\pi)|^2 \, d\xi = J_1 + J_2,
\end{aligned}$$

where, since $|\hat{f}(\xi)| \le C(1+|\xi|)^{-1/2-\epsilon}$,

$$\begin{aligned}
J_1 &:= \int_{2^n\alpha\le|\xi|\le 2^n\pi} \prod_{j=1}^n |a(2^{-j}\xi)|^2 \sum_{l\ne 0} |\hat{f}(2^{-n}\xi + 2l\pi)|^2 \, d\xi\\
&\le C_2 \int_{2^n\alpha\le|\xi|\le 2^n\pi} \prod_{j=1}^n |a(2^{-j}\xi)|^2 \, d\xi \ \to 0
\end{aligned}$$

by Lemma 4.4 again. It remains to prove that when $n \to \infty$,

$$J_2 := \int_{|\xi|\le 2^n\alpha} G_n(\xi) d\xi = \int_{-\infty}^{+\infty} G_n(\xi)\chi_{|\xi|\le 2^n\alpha}(\xi) d\xi \to 0,$$

where $G_n(\xi) := \prod_{j=1}^n |a(2^{-j}\xi)|^2 \sum_{l\neq 0} |\hat{f}(2^{-n}\xi + 2l\pi)|^2$. By (4.4), we have

$$|G_n(\xi)\chi_{|\xi|\leq 2^n\alpha}(\xi)| \leq \frac{1}{C'}|\hat{\phi}(\xi)|^2 \cdot \sum_{l\neq 0} |\hat{f}(2^{-n}\xi + 2l\pi)|^2 \tag{4.6}$$

$$\leq \frac{C_2}{C'}|\hat{\phi}(\xi)|^2 \quad \text{(by the decay of } f). \tag{4.7}$$

For any fixed ξ, since $\hat{f}(2\pi l) = \hat{f}(0)\delta_{l,0}$ and $|\hat{f}(\xi)| \leq C/(1+|\xi|)^{(1/2)+\epsilon}$, we have

$$\lim_{n\to+\infty} \sum_{l\neq 0} |\hat{f}(2^{-n}\xi + 2l\pi)|^2 = 0.$$

Therefore, by (4.6), $G_n(\xi)\,\chi_{|\xi|\leq 2^n\alpha}(\xi) \to 0$ pointwise. By (4.7) and Lebesgue's theorem, J_2 tends to 0, which implies that $\phi_n^3 \to 0$ in $L^2(\mathbb{R})$. This finishes our proof of L^2 convergence of $T^n f$ to $\hat{f}(0)\phi$. □

Remark. If we consider $P_{|v|} = P_{(\sqrt{|v|})^2}$ and let β_1 be the spectral radius for $P_{|v|}$ restricted to $C[0,2\pi]$, then with the same assumptions except that $\beta_1 < 2^r$ and with $|\hat{f}(\xi)| \leq C(1+|\xi|)^{-1-\epsilon}$ for some $C, \epsilon > 0$, we find that the proofs of the previous lemmas and theorems concerning L^2-convergence can be copied (with small changes) to prove a similar theorem for L^1-convergence of $\hat{Tf}(\xi)$ to $\hat{f}(0)\hat{\phi}(\xi)$, which implies the pointwise convergence of $T^n f(x)$ to $\hat{f}(0)\phi(x)$.

THEOREM 4.6 (pointwise convergence of $T^n f(x)$). *Let $f(x)$ and $a(\xi)$ be as in Theorem 4.5 with the stronger regularity condition $|\hat{f}(\xi)| \leq C\,(1+|\xi|)^{-1-\epsilon}$ for some $\epsilon, C > 0$. Let β_1 be the spectral radius of $P_{|v|}$ restricted to $C[0,2\pi]$. If $\beta_1 < 2^r$, then $(T^n f)^\wedge(\xi)$ converges in $L^1(\mathbb{R})$ to $\hat{f}(0)\hat{\phi}(\xi)$; therefore, $T^n f(x)$ converges pointwise to $\hat{f}(0)\phi(x)$.*

Remark. The proofs of L^2- or L^1-convergence both depend on the property that $\hat{f}(2\pi l) = \hat{f}(0)\delta_{l,0}$. Actually, this is essential for the L^2- (or L^1-) convergence of functions like $\prod_{j=1}^n a(2^{-j}\xi)\hat{f}(2^{-n}\xi)$. We will study this in detail in the next section.

Example 4.7. In §3.1, we looked at the example where $f = \alpha\phi + \beta\psi$ with $|\alpha|^2 + |\beta|^2 = 1$. Let us revisit this example in light of the results of this section. We have

$$\hat{f}(2\pi l) = \alpha\,\hat{\phi}(2\pi l) + \beta\,\hat{\psi}(2\pi l), \quad \hat{\phi}(2\pi l) = \delta_{l,0}\hat{\phi}(0), \quad \hat{\psi}(2\pi l) = e^{i\pi l}\overline{m_0(\pi l + \pi)}\hat{\phi}(\pi l).$$

However, $\hat{\psi}(2\pi l) = e^{i\pi l}\overline{m_0(\pi)}\hat{\phi}(\pi l) = 0$ if $l = 2m$; $\hat{\psi}(2\pi l) = -\hat{\phi}(\pi(2m+1))$ if $l = 2m+1$. Since $\sum_k |\hat{\phi}(2k\pi + \xi)|^2 = \frac{1}{2\pi}$ for all ξ, it follows that $\sum_m |\hat{\phi}(2m\pi + \pi)|^2 \neq 0$ so that $\hat{\psi}(2\pi l)$ cannot be 0 for all $l = 2m+1, m \in \mathbb{Z}$. Thus f can satisfy properties (1), (2), and (3) of Lemma 4.2 only if $\beta = 0$, i.e., $f = \alpha\phi$.

Note also that we can use Lemma 4.3 for a simple proof that $T^N f \to 0$ in $L^2(\mathbb{R})$ for $\beta \neq 0$ (as promised at the end of §3.1). Indeed, we have

$$\|T^N f\|^2 = \int_{-\infty}^{+\infty} |\alpha|^{2N} \prod_{j=1}^N |\alpha\; m_0(2^{-j}\xi) + \beta\; m_1(2^{-j}\xi)|^2 |\hat{f}(2^{-N}\xi)|^2 d\xi$$

$$= |\alpha|^{2N} \int_{|\xi|\leq 2^n\pi} F(2^{-n}\xi) \prod_{j=1}^N |w(2^{-j}\xi)|^2\, d\xi = |\alpha|^{2N} \int_{-\pi}^{\pi} (P_{|w|^2})^N F(\xi) d\xi,$$

where $w(\xi) = \alpha m_0(\xi) + \beta m_1(\xi)$ and $F(\xi) = \sum_l |\hat{f}(\xi + 2\pi l)|^2$. Since the $\{f(\bullet - l)\}_{l\in\mathbb{Z}}$'s are orthonormal, $F(\xi) = \frac{1}{2\pi}$. On the other hand, as a consequence of (3.1), we have $|m_0(\xi)|^2 + |m_1(\xi)|^2 = 1$, which implies that $|w(\xi)|^2 + |w(\xi+\pi)|^2 = 1$. Hence $(P_{|w|^2})^N F(\xi) = F(\xi)$. Therefore, $\lim_{N\to\infty} \|T^N f\|^2 = \lim_{N\to\infty} |\alpha|^{2N} \cdot 2\pi \cdot \frac{1}{2\pi} = 0$, i.e., $T^N f \overset{L^2}{\to} 0$.

We can now answer the questions we raised at the beginning of §3.2.

Example 4.8. First, for $f = \chi_{[0,1+\epsilon)}$, $0 < \epsilon < 1$, we check that $\hat{f}(2k\pi) = (1 - e^{i\epsilon(2k\pi)})/i(2k\pi)$. Therefore, f does not satisfy property (2) of Lemma 4.2; hence $a_f(0) \neq 1$. By the same argument as used in the proof for (1) ⇒ (2) in Lemma 4.2, we even find $|a_f(0)| < 1$. It then follows that $(T^n f)^\wedge \to 0$.

Example 4.9. If $f(x) = e^{-x^2/2}$, then $\hat{f}(\xi) = e^{-\xi^2/2}$. A similar argument shows again that $(T^n f)^\wedge \to 0$.

5. Scaled iteration of T.

Most of our examples go to 0 under the iterations of T, usually because of a systematic "shrinking" of $T^n f$ (see Example 4.7 in the previous section). Can this can be corrected by adjusting the normalization of $T^n f$? Note that although we can control $\|T^n f\|$ (as in Proposition 5.1 below), it is difficult to give a simple formula for $\|T^n f\|$. Therefore, instead of studying the sequence $\{T^n f/\|T^n f\|\}$, we will concentrate on the convergence of the sequence $\Gamma_n = \mu_n \cdot T^n f$ for appropriately chosen μ_n. Our analysis of §3 shows that then $\hat{\Gamma}_n(\xi) = \mu_n \prod_{j=1}^n a(2^{-j}\xi)\hat{f}(2^{-n}\xi)$. Without μ_n, we have several examples where $(T^n f)^\wedge \to 0$ pointwise simply because $|a(0)| < 1$. To remedy this, we define $\mu_n = [a(0)]^{-n}$, i.e.,

$$\Gamma_n = [a(0)]^{-n} \cdot T^n f. \tag{5.1}$$

This makes sense only if $a(0) \neq 0$, which we shall assume whenever we study Γ_n.

We now concentrate on $C^\alpha[0, 2\pi] = \{f \in C[0, 2\pi] \mid \sup_{x\neq y} |f(x) - f(y)|/|x-y|^\alpha < +\infty\}$, $\alpha \in (0, 1]$. The following proposition shows that under some conditions, (5.1) is almost equivalent to simply normalizing the $T^n f$'s to 1.

PROPOSITION 5.1. *Suppose that* $F(\xi) = \sum_{l\in\mathbb{Z}} |\hat{f}(\xi + 2\pi l)|^2$ *and* $|a(\xi)|^2$ *are in* $C^\alpha[0, 2\pi]$ *with* $a(\pi) = 0$ *and* $a(0) \neq 0$. *If, furthermore, f is T-amenable, then there exists* $c_1 > 0$, *and for all* $\epsilon > 0$, *there exists* $c_2 > 0$ *so that* $c_1|a(0)|^n \leq \|T^n f\| \leq c_2|a(0)|^n(1+\epsilon)^n$.

To prove this proposition, we shall need the following result, proved in more generality by Hennion [Hen] and discussed in detail in [Her1].

THEOREM 5.2 (Hennion). *Assume* $|a(\xi)|^2 \in C^\alpha[0, 2\pi]$ *and let* ρ *be the spectral radius for* $P_{|a|^2}$ *restricted to* $C^\alpha[0, 2\pi]$. *Then there exists* $\gamma \geq 0$, $\|\gamma\| \neq 0$, *such that* $P_{|a|^2}\gamma = \rho\gamma$.

Proof of Proposition 5.1. Using results in [Her1], we can check that $a(\pi) = 0 \Rightarrow \rho = |a(0)|^2$. Using the Cauchy–Schwarz inequality and $\lim_{n\to\infty} \|P^n\|^{1/n} = \rho$, we now have

$$\begin{aligned}\|T^n f\|^2 &= \int_{|\xi|\leq 2^n\pi} F(2^{-n}\xi) \prod_{j=1}^n |a(2^{-j}\xi)|^2 \, d\xi = \int_{-\pi}^{\pi} P^n_{|a|^2} F(\xi)\, d\xi. \\ &\leq \sqrt{2\pi}\, \|P^n_{|a|^2} F(\xi)\| \leq C(\rho + \varepsilon)^n, \quad \varepsilon > 0.\end{aligned}$$

Thus we can choose $c_2 > 0$ such that $\|T^n f\| \leq c_2|a(0)|^n(1+\epsilon)^n$.

To prove the other inequality, we use the fact that f is RBP and Theorem 5.2. Take γ in $C^\alpha[0, 2\pi]$. Then we can assume $\gamma \leq M$ for a positive constant M. Hence $0 \leq P^n\gamma \leq MP^n 1$ so that

$$\rho^n\|\gamma\|^2 = \langle\gamma, P^n\gamma\rangle \leq M^2\langle 1, P^n 1\rangle \leq \frac{M^2}{C_1}\int_{-\pi}^{\pi} P^n F(\xi)\, d\xi \quad \text{(by (2.2))},$$

where we have used $\gamma \geq 0$. Thus with $c_1 = M/\sqrt{C_1}$, $\|T^n f\| \geq c_1 |a(0)|^n$. □

Remark. The condition $a(\pi) = 0$ automatically holds if $a(0) = 1$, which reduces to the unscaled case, because $\Gamma_n = T^n f$. However, $a(\pi) = 0$ is also possible when $a(0) \neq 1$. For example, if f is symmetric with respect to the line $x = k + \frac{1}{2}$ with $k \in \mathbb{Z}$, then Tf is also symmetric with respect to the same line, and $a(\pi) = 0$ follows.

Let us now investigate the convergence of the Γ_n's. We begin by considering functions supported in [0,1]. This gives explicit examples where $T^n f$ converages to 0 weakly and for which the Γ_n's converge to a nonzero limit nevertheless.

PROPOSITION 5.3. *Take $f \in L^2(\mathbb{R})$ with support $f \subset [0,1]$ and such that*

$$\int_0^{\frac{1}{2}} f(y)\overline{f(2y)}dy + \int_{\frac{1}{2}}^1 f(y)\overline{f(2y-1)}dy \neq 0. \tag{5.2}$$

Assume that $f|_{[0,1]}$ is not constant a.e. Then $(T^n F)^\wedge(\xi) \to 0$ pointwise for all ξ, but $\hat{\Gamma}_n(\xi)$ converges pointwise to a limit $\hat{\Gamma}_\infty(\xi)$ which may not be 0. This limit is in $L^2(\mathbb{R})$ if and only if

$$\int_0^{\frac{1}{2}} f(y)\overline{f(2y)}dy = \int_{\frac{1}{2}}^1 f(y)\overline{f(2y-1)}dy, \tag{5.3}$$

and in that case Γ_∞ equals $\hat{f}(0)\chi_{[0,1]}$.

Proof. Define $c_0 := \|f\|^{-2} \int_0^{1/2} f(y)\overline{f(2y)}dy$ and $d_0 := \|f\|^{-2} \int_{1/2}^1 f(y)\overline{f(2y-1)}dy$. Because the $f(\bullet - n)$'s are orthonormal, we easily find $a(\xi) = c_0 + d_0 e^{-i\xi}$, so $a(0) = c_0 + d_0$, which explains the assumption in (5.2). We also check that $|c_0|^2 + |d_0|^2 \leq \frac{1}{2}$, so $|a(0)| \leq 1$. Moreover, $|a(0)| = 1$ is only possible if $c_0 = d_0 = \frac{1}{2}$, which would imply (for details, see [Hu]) that $f|_{[0,1]}$ is constant a.e. Since this was excluded by our assumption, it follows that $|a(0)| < 1$, and $(T^n F)^\wedge \to 0$ pointwise. If we let $A(\xi) = a(\xi)/a(0) = (c_0 + d_0 e^{-i\xi})/(c_0 + d_0)$, then by Theorem 4.1, $\hat{\Gamma}_\infty(\xi) = \hat{f}(0)\prod_{j=1}^\infty A(2^{-j}\xi)$, or $\hat{\Gamma}_\infty(\xi) = A(\frac{\xi}{2})\hat{\Gamma}_\infty(\frac{\xi}{2})$. If $\Gamma_\infty \in L^2(\mathbb{R})$, then Γ_∞ is also in $L^1(\mathbb{R})$ and we can apply Theorems 3.1 and 5.1 in [DL1] to conclude $A(\xi) = \frac{1}{2}(1 + e^{-i\xi})$. This implies (5.3) and $\Gamma_\infty = \hat{f}(0)\chi_{[0,1]}$. □

There are many functions supported in [0,1] that satisfy (5.3). For instance, any f symmetric with respect to the line $x = \frac{1}{2}$ will do, as will any $\frac{1}{2}$-periodic function. An example is $f(x) = 6(x - x^2)$; for this function, the graph of Γ_n consists of 2^n arches with amplitude 1.5; this converges to $\chi_{[0,1]}$ in the distributional sense but not in $L^2(\mathbb{R})$ or pointwise. (We will see below that this absence of L^2-convergence holds in more generality.)

If $\hat{f}(2\pi l) = \hat{f}(0)\delta_{l,0}$, then $a(0) = 1$ and Γ_n is just $T^n f$. The next theorem states that the condition $\hat{f}(2\pi l) = \hat{f}(0)\delta_{l,0}$ is *necessary* for L^2- (or L^1-) convergence for $\hat{\Gamma}_n$. This means that, while we have more cases of pointwise convergence for $\hat{\Gamma}_n(\xi)$ than for $(T^n f)^\wedge(\xi)$, nothing is changed for L^2- (or L^1-) convergence: if $(T^n f)^\wedge$ does not converge in L^2 (or L^1), then neither does $\hat{\Gamma}_n$.

THEOREM 5.4. *Assume that $b(\xi)$ is 2π-periodic with Hölder exponent $\beta > 0$ and with $b(0) = 1$. For $g \in L^2(\mathbb{R}) \bigcap L^1(\mathbb{R})$ with $\hat{g}(0) \neq 0$, consider the functions G_n and G defined by $\hat{G_n}(\xi) := \prod_{j=1}^n b(2^{-j}\xi)\,\hat{g}(2^{-n}\xi)$ and $\hat{G}(\xi) := \prod_{j=1}^\infty b(2^{-j}\xi)\,\hat{g}(0)$. If $\hat{G_n}(\xi)$ converges to $\hat{G}(\xi)$ in $L^2(\mathbb{R})$ (or $L^1(\mathbb{R})$) as well as pointwise with $\hat{G} \neq 0$, then $\hat{g}(2\pi l) = \hat{g}(0)\,\delta_{l,0}$.*

Proof. We prove only the case of L^2-convergence; the L^1 case is entirely analogous.

1. Since $\hat{G}$ is continuous and $\hat{G}(0) \neq 0$, there exists an $\alpha \in (0, \pi]$ such that $|\xi| \leq \alpha \Rightarrow |\hat{G}(\xi)| \geq C > 0$. Thus we have $\chi_{|\xi| \leq \alpha}(\xi) \leq \chi_{|\xi| \leq 2^n \alpha}(\xi) \leq \hat{G}(2^{-n}\xi)/C$ for all $n \in \mathbb{N}$.

2. We now consider $\prod_{j=1}^{n} b(2^{-j}\xi)\, \chi_{|\xi| \leq \alpha}(\xi)$, which obviously converges pointwise to $\prod_{j=1}^{\infty} b(2^{-j}\xi) \chi_{|\xi| \leq \alpha}(\xi)$. As in step 3 of the proof of Theorem 4.5,

$$\left| \prod_{j=1}^{n} b(2^{-j}\xi)\, \chi_{|\xi| \leq \alpha}(\xi) \right| \leq \frac{|\hat{G}(\xi)|}{|\hat{g}(0)| C},$$

and so $\lim_{n \to \infty} \int_{|\xi| \leq \alpha} \prod_{j=1}^{n} |b(2^{-j}\xi)|^2 \, d\xi = \int_{|\xi| \leq \alpha} \prod_{j=1}^{\infty} |b(2^{-j}\xi)|^2 \, d\xi > 0$. Thus there exists an n_0 such that for all $n \geq n_0$ (with some constant $C' > 0$),

$$\int_{|\xi| \leq \alpha} \prod_{j=1}^{n} |b(2^{-j}\xi)|^2 \, d\xi \geq \frac{1}{2} \int_{|\xi| \leq \alpha} \prod_{j=1}^{\infty} |b(2^{-j}\xi)|^2 \, d\xi \equiv 2C' \int_{|\xi| \leq \alpha} |\hat{G}(\xi)|^2 d\xi. \tag{5.4}$$

3. Assume $\hat{g}(2\pi l) \neq 0$ for some $l \neq 0$. Then we can choose n_1 such that for all $n \geq n_1$ and all $|\xi| \leq \alpha$, $|\hat{g}(2\pi l + 2^{-n}\xi)|^2 \geq \frac{1}{2} |\hat{g}(2\pi l)|^2$. Consequently,

$$\begin{aligned} \int_{|\xi| \leq \alpha} |\hat{G}_n(2^n 2\pi l + \xi)|^2 d\xi &= \int_{|\xi| \leq \alpha} \prod_{j=1}^{n} |b(2^{-j}\xi)|^2 \, |\hat{g}(2\pi l + 2^{-n}\xi)|^2 d\xi \\ &\geq \frac{1}{2} |\hat{g}(2\pi l)|^2 \int_{|\xi| \leq \alpha} \prod_{j=1}^{n} |b(2^{-j}\xi)|^2 \, d\xi \\ &\geq C' |\hat{g}(2\pi l)|^2 \int_{|\xi| \leq \alpha} |\hat{G}(\xi)|^2 d\xi \quad \text{(by (5.4))}. \end{aligned}$$

4. On the other hand, we have

$$\left(\int_{|\xi| \leq \alpha} |\hat{G}_n(2^n 2\pi l + \xi)|^2 d\xi \right)^{\frac{1}{2}} \leq \|\hat{G}_n - \hat{G}\| + \left(\int_{|\xi| \leq \alpha} |\hat{G}(2^n 2\pi l + \xi)|^2 d\xi \right)^{\frac{1}{2}}.$$

For any $\epsilon > 0$, since we assume $\hat{G}_n \to \hat{G}$ in L^2, we can choose n_3 such that the first term $< \frac{\epsilon}{2}$ for all $n \geq n_3$. Because by a change of variable,

$$\int_{|\xi| \leq \alpha} |\hat{G}(2^n 2\pi l + \xi)|^2 d\xi \leq \int_{|\xi| \geq 2^{n+1}\pi |l| - \alpha} |\hat{G}(\xi)|^2 \, d\xi,$$

we can also choose n_4 so that the second term is bounded by $\frac{\epsilon}{2}$ for all $n \geq n_4$.

5. For $n \geq \max\{n_i | i = 1, 2, 3, 4\}$, we then can combine the above inequalities to get

$$\epsilon^2 \geq C' |\hat{g}(2\pi l)|^2 \int_{|\xi| \leq \alpha} |\hat{G}(\xi)|^2 d\xi > 0.$$

Since ϵ can be made arbitrarily small, this contradicts our assumption that $\hat{g}(2\pi l) \neq 0$. Thus $\hat{g}(2\pi l) = 0$ for all $l \neq 0$. □

Note that the condition $\hat{G}(\xi) \neq 0$ plays a crucial role: if it is not satisfied, then Example 4.7, where $T^n f$ converges in $L^2(\mathbb{R})$ to 0 while $\hat{f}(2\pi l) \neq 0$ for $a \neq 0$, provides a counterexample. Together with Theorem 6.2 from [Her1], Theorem 5.4 can be used to formulate conditions on f that are necessary and sufficient for the L^1- or L^2-convergence of the $(T^n f)^\wedge$'s as well as of the $\hat{\Gamma}_n$'s. (The conditions in Theorems 4.11 and 4.12 are sufficient but not necessary.)

6. Examples illustrating regularity. In this section, we present examples showing that the smoothness of f and its limit function are not related. In the process, these examples also illustrate the results in §§4 and 5. To estimate the regularity of a refinable function f with filter $a(\xi)$, we use theorems in [Her1]–[Her3] that allow us to compute Sobolev exponents s_2 $(:= \sup\{s : \int |\hat{f}(\xi)|^2(1+|\xi|^{2s})d\xi < +\infty\})$ from $|a(\xi)|$. Typically, we need $a(\xi)$ to be factorizable like $|a(\xi)| = |\cos\frac{\xi}{2}|^r|v(\xi)|$.

Example 6.1. Take f to be continuous and supported in [-1,1] with $f(x)+f(x-1) =$ nonzero constant for $x \in [0,1]$. Then f satisfies the conditions in Lemma 4.2 even though f is not necessarily a scaling function. We compare the following two examples:

1. $f_1(x) = \cos^2\frac{\pi x}{2}$, $|x| \le 1$, and $f_1(x) = 0$ otherwise. Then

$$a(\xi) = \cos^2\frac{\xi}{2}\cdot\left[\frac{1-\left(1-\frac{8}{3\pi}\right)\sin^2\frac{\xi}{2}}{1-\frac{1}{2}\sin^2\frac{\xi}{2}}\right]. \tag{6.1}$$

Using the formulas in [Her1], we find $\beta \approx 3.351 < 16$ and $\beta_1 \approx 2.564 < 4$, so $T^n f_1 \overset{L^2}{\rightarrow} f_1^\infty$ and $T^n f_1(x)$ converges pointwise to $f_1^\infty(x)$ (see Theorems 4.5 and 4.6).

2. $f_2(x)|_{[0,1]} = t(1-x)$, $t(y) = y^4(35-84y+70y^2-20y^3)$. Then

$$a(\xi) = \frac{\frac{1549}{12672}+\frac{1}{2}\cos\xi+\frac{4787}{12672}\cos 2\xi}{\frac{1042}{1287}+\frac{245}{1287}\cos\xi} = \cos^2\frac{\xi}{2}\cdot\left(\frac{1-\frac{4787}{1584}\sin^2\frac{\xi}{2}}{1-\frac{245}{2574}\sin^2\frac{\xi}{2}}\right). \tag{6.2}$$

Similar analysis shows that $T^n f_2(x)$ converges pointwise to $f_2^\infty(x)$.

Note that while $f_2(\in C^{4-\epsilon})$ is more regular than $f_1(\in C^{2-\epsilon})$, explicit computation shows that the Sobolev exponents s_2 for the limit functions do not differ much: 1.128 for f_1^∞ and 1.508 for f_2^∞; this is because the regularity of the limit is determined in large part by the number of factors $|\cos\frac{\xi}{2}|$ in $|a(\xi)|$ rather than by the regularity of f itself. For graphs of the limit functions, see [Hu].

Example 6.2. To see a similar situation for the regularity of the limit of Γ_n, we consider $f(x) = e^{-\alpha^2x^2}$, which is C^∞, but the limit of Γ_n can be very irregular depending on the values of α. We have

$$a(\xi) = \left[\sum_{l\in\mathbb{Z}} e^{-\frac{5}{4\alpha^2}(\xi-2\pi l)^2}\right] \Big/ \left[\sum_{l\in\mathbb{Z}} e^{-\frac{1}{2\alpha^2}(\xi-2\pi l)^2}\right].$$

Two extremes are given by $\alpha^2 = \frac{1}{2}$ and $\alpha^2 = 32$. In the first case, $a(\pi) \approx 0.372\times10^{-6}$ so that $a(\xi)$ has "almost" a zero at $\xi = \pi$ and the subdivision scheme produces smooth functions Γ_n that are very similar to f itself, although the true Γ_∞ must be less smooth. For $\alpha^2 = 32$, we find $a(0) \approx 0.63455711$ and $a(\pi) \approx 0.63035395$. Since these are the two extreme values of $a(\xi)$, this shows that $a(\xi)$ is almost "flat." This case is very different from the case $\alpha^2 = \frac{1}{2}$: the limit of Γ_n now is very peaked. Details and graphs can be found in [Hu].

Example 6.3. We study the convergence of the scaled iterates Γ_n for nonsmooth f. We consider $f(x) = 1$ for $-\epsilon < x \le 1+\epsilon$ and $f(x) = 0$ otherwise, which is a dilated and shifted version of $\chi_{(0,1]}$. We shall restrict ourselves to $0 < \epsilon < \frac{2}{3}$. In this case,

$$a(\xi) = \frac{\frac{3\epsilon}{2}(e^{i\xi}+e^{-2i\xi})+(\epsilon+\frac{1}{2})(1+e^{-i\xi})}{4\epsilon\cos\xi+1+2\epsilon} \tag{6.3}$$

and $a(0) = \frac{5\epsilon+1}{6\epsilon+1} < 1$. We therefore have $(T^n f)^\wedge \to 0$. On the other hand, since it is obvious that $a(\xi) \in C^\infty$, the renormalized Γ_n's do have a nontrivial limit. Figure 1 shows the limit function Γ_∞ for a few different ϵ's. Note that the smoothness of Γ_∞ changes with ϵ: an easy computation using (6.3) shows that while $a(\pi) = 0$ and $a'(\pi) \neq 0$ in general, which means that for $0 < \epsilon < \frac{2}{3}$ and $\epsilon \neq \frac{1}{7}$, $|a(\xi)|$ can be divisible only by $|\cos\frac{\xi}{2}|$. The value of $\epsilon = \frac{1}{7}$ is an exception; in this case, $a(\pi) = a'(\pi) = a''(\pi) = 0$. This explains why the graph for Γ_∞ with $\epsilon = \frac{1}{7}$ is the smoothest! This is a typical example of a nonsmooth function leading to a much more regular limit under the scaled iteration.

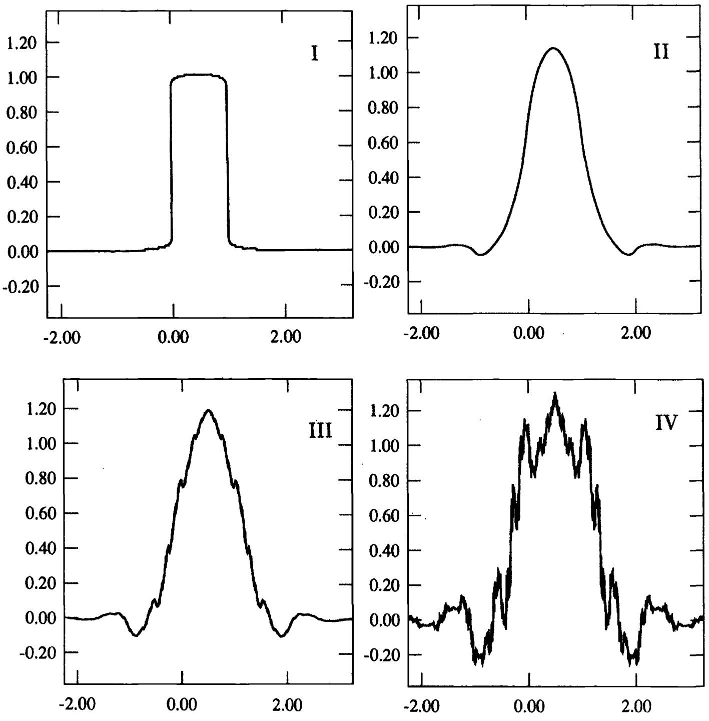


FIG. 1. *The limit function of* Γ_n *for* $f(x) = \chi_{(-\epsilon,1+\epsilon]}(x)$ *for different values of* ϵ: (I) $\epsilon = 0.01$, (II) $\epsilon = \frac{1}{7} \approx 0.14286$, (III) $\epsilon = 0.2$, *and* (IV) $\epsilon = 0.3$. *Their Sobolev exponents* s_2 *are* 1.014, 3.115, 1.138, *and* 1.163, *respectively.*

Acknowledgments. I am pleased to express my gratitude to my advisor, Professor Ingrid Daubechies, for suggesting this topic and being generous with her time and ideas. I also thank Dr. Loic Hervé for providing me with his preprints and for his comments on this paper. Finally, I thank the referees for their valuable suggestions which improved the quality of this paper.

REFERENCES

[BF] J. BENEDETTO AND M. FRAZIER, EDS., *Wavelets: Mathematics and Applications*, CRC Press, Boca Raton, FL, 1993.

[CDM] A. CAVARETTA, W. DAHMEN, AND C. MICCHELLI, *Stationary subdivision*, Mem. Amer. Math. Soc., 93 (1991), pp. 1–186.

[Ch] C. K. Chui, ed., *Wavelets: A Tutorial in Theory and Applications*, Academic Press, Boston, 1992.

[CD1] A. Cohen and I. Daubechies, *A stability criterion for biorthogonal wavelet bases and their related subband coding scheme,* Duke Math. J., 68 (1992), pp. 313–335.

[CD2] ———, *A new technique to estimate the regularity of a refinable function*, Rev. Mat. Iberoamericana, to appear.

[CMW] R. R. Coifman, Y. Meyer, and V. Wickerhauser, *Wavelet analysis and signal processing*, in Wavelets and their Applications, M. B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael, eds., Jones and Bartlett, Boston, 1992.

[Co] J. P. Conze, *Sur la régularité des solutions d'une équation fonctionnelle*, Report, Laboratoire de Probabilités, Université de Rennes I, Rennes, France, 1989.

[CR] J. P. Conze and A. Raugi, *Fonction harmonique pour un opérateur de transtion et application*, Bull. Soc. Math. France, 118 (1990), pp. 273–310.

[Dau1] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[Dau2] ———, *Ten Lectures on Wavelets*, CBMS–NSF Series in Applied Mathematics 61, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[DH] I. Daubechies and Y. Huang, *How does truncation of the mask affect a refinable function?*, Constr. Approx., 11 (1995), pp. 365–380.

[DL1] I. Daubechies and J. C. Lagarias, *Two-scale difference equations* I: *Existence and global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388–1410.

[DL2] ———, *Two-scale difference equations* II: *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[Du] S. Durand, *Convergence of cascade algorithms introduced by I. Daubechies*, Numer. Algorithms, 4 (1993), pp. 307–322.

[Ei] T. Eirola, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[Gr] G. Gripenberg, *Unconditional bases of wavelets for Sobolev spaces*, SIAM J. Math. Anal., 24 (1994), pp. 1030–1042.

[Hen] H. Hennion, *Sur un théorème spectral et son application aux noyaux lipchitziens*, Proc. Amer. Math. Soc., 118 (1993), pp. 627–634.

[Her1] L. Hervé, *Construction et régularité des fonctions d'échelle*, SIAM J. Math. Anal., 26 (1995), pp. 1361–1385.

[Her2] ———, *Méthodes d'opérateurs quasi-compacts en analyse multirésolution, applications à la construction de bases d'ondelettes et à l'interpolation*, Ph.D. thesis, Laboratoire de Probabilités, Université de Rennes I, Rennes, France, 1992.

[Her3] ———, *Régularité et conditions de bases de Riesz pour les fonctions d'échelle*, C. R. Acad. Sci. Paris Ser. I. Math., 315 (1992), pp. 1029–1032.

[Her4] ———, *Comportement asymptotique dans l'algorithme de transformée en ondelettes: Lien avec la régularité de l'ondelette*, preprint, Laboratoire de Probabilités, Université de Rennes I, Rennes, France, 1993.

[Hu] Y. Huang, *A nonlinear operator related to refinable functions*, Ph.D. thesis, Department of Mathematics, Rutgers University, New Brunswick, NJ, 1995.

[I1] *Special Issue in Wavelets*, IEEE Trans. Inform. Theory, 38 (1992).

[I2] *Special Issue in Wavelets*, IEEE Trans. Signal. Process., 41 (1993).

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

[Me] Y. Meyer, *Wavelets and Operators*, English ed., Cambridge University Press, Cambridge, UK, 1992.

[Ru] M. B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael, *Wavelets and Their Applications*, Jones and Bartlett, Boston, 1992.

[Vi] L. Villemoes, *Regularity of two-scale difference equations and wavelets*, Ph.D. thesis, Mathematical Institute, Technical University of Denmark, Lyngby, Denmark, 1992.

SIAM J. MATH. ANAL.
Vol. 27, No. 6, pp. 1791–1815, November 1996

014

INTERTWINING MULTIRESOLUTION ANALYSES AND THE CONSTRUCTION OF PIECEWISE-POLYNOMIAL WAVELETS*

GEORGE C. DONOVAN[†], JEFFREY S. GERONIMO[†], AND DOUGLAS P. HARDIN[‡]

Abstract. Let (V_p) be a local multiresolution analysis of $L^2(\mathbf{R})$ of multiplicity $r \geq 1$, i.e., V_0 is generated by r compactly supported scaling functions. If the scaling functions generate an orthogonal basis of V_0, then (V_p) is called an *orthogonal multiresolution analysis.* We prove that there exists an orthogonal local multiresolution analysis (V_p') of multiplicity r' such that

$$V_q \subset V_0' \subset V_{q+n}$$

for some integers $q \geq 0$, $n \geq 1$, and $r' > 1$.

In particular, this shows that compactly supported orthogonal polynomial spline wavelets and scaling functions (of multiplicity $r' > 1$) of arbitrary regularity exist, and we give several such examples.

Key words. multiwavelet, intertwining multiresolution analyses, orthogonal wavelet, splines

AMS subject classification. 41A15

1. Introduction.

The starting point for most wavelet constructions is a single function $\phi \in L^2(\mathbf{R})$, called a *scaling function*, whose integer translates form a Riesz basis for a closed linear subspace $V_0 \subset L^2(\mathbf{R})$. If the scaling function is compactly supported and generates an orthogonal basis of V_0, then the associated wavelet will also be compactly supported and generate an orthogonal basis. Daubechies (cf. [5]) constructed scaling functions and associated wavelets that were compactly supported, generated orthogonal bases, and were continuous (or smoother). However, these wavelets do not have closed-form representations; they are defined via a limiting process. Also, it is known that these wavelets cannot have certain other desirable properties (e.g., symmetry). By giving up compact support (cf. [1]) or orthogonality (cf. [4]), symmetric wavelet bases have been constructed using piecewise-polynomial splines. For some applications, the symmetry and simple representation are more important than having both compact support and orthogonality.

Recently, wavelet constructions generated by a finite collection of scaling functions $\Phi = \{\phi^1, \ldots, \phi^r\}$ have been studied (cf. [10], [11], [8], [12], [13], [18], [22]). In [8] and [7], symmetric, compactly supported, continuous, and orthogonal scaling functions (and associated wavelets) were constructed using $r = 2$ scaling functions. As in the case of Daubechies wavelets, these functions do not have closed-form representations. In this paper, using $r > 2$ scaling functions, we construct wavelets that not only have the three properties of compact support, arbitrary regularity, and orthogonality but are also symmetric and piecewise polynomial. In fact, we show that for any multiresolution analysis generated by compactly supported scaling functions, there is an associated intertwining orthogonal multiresolution analysis also generated by compactly supported scaling functions.

More precisely, a *multiresolution analysis of multiplicity* r is a nested sequence of closed linear subspaces (V_p) in $L^2(\mathbf{R})$ satisfying the following:

1. $f \in V_p$ iff $f(2^{-p}\cdot) \in V_0$ for $p \in \mathbf{Z}$.

*Received by the editors October 26, 1994; accepted for publication (in revised form) August 3, 1995. The research of the first author was partially supported by an NSF graduate fellowship. The research of the second and third authors was partially supported by NSF grants.

[†]School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332.

[‡]Department of Mathematics, Vanderbilt University, Nashville, TN 37240.

2. $V_0 \subset V_1$.
3. $\bigcap_{p\in\mathbf{Z}} V_p = \{0\}$.
4. $\bigcup_{p\in\mathbf{Z}} V_p$ is dense in $L^2(\mathbf{R})$.
5. There are r functions $\phi^1, \ldots, \phi^r$ such that the collection of integer translates $\{\phi^s(\cdot - n) \mid s = 1, \ldots, r \text{ and } n \in \mathbf{Z}\}$ is a Riesz basis of V_0.

The functions $\phi^1, \ldots, \phi^r$ are called scaling functions and are said to *generate* the multiresolution analysis (V_p). If there is a set of compactly supported scaling functions whose integer translates form an orthogonal basis of V_0, then we call (V_p) an *orthogonal* multiresolution analysis. The main theorem of this paper is as follows.

THEOREM 1. *If (V_p) is a multiresolution analysis generated by compactly supported scaling functions, then there is some pair of integers (q, n) and some orthogonal multiresolution analysis $(\tilde{V}_p)$ such that*

$$V_q \subset \tilde{V}_0 \subset V_{q+n}.$$

We say that (V_p) and $(\tilde{V}_p)$ are *intertwining multiresolution analyses.*

For the sake of completeness, we note that if the scaling functions are compactly supported, then the intersection property 3 follows from 1 and the fact that V_0 is finitely generated shift-invariant (FSI) space (see §2) [15, Thm. 2.2]. Furthermore, the density property 4 follows from conditions 1 and 2 (see [3] for the single-scaling-function case, which directly generalizes to the multiple-scaling-function case). The Riesz-basis condition 5 may be relaxed: if V_0 is an FSI space with compactly supported generators, then there exists a set of compactly supported generators whose collection of integer translates form a Riesz basis (cf. [2, Thm. 3.38] and [20, Thm. 3.7]). In practice, we will take V_0 to be a classical spline space, in which case it will be elementary to verify conditions 1–5 directly.

If (V_p) is an orthogonal multiresolution analysis, then there is a general procedure (cf. [7], [22], [16], [23]) for calculating compactly supported wavelets $\psi^1, \ldots, \psi^r$ that generate an orthogonal basis of $W_0 = V_1 \ominus V_0$. As in the single-scaling-function case, the work is in finding orthogonal scaling functions.

The structure of this paper is as follows. In §2, we show that, without loss of generality, we can assume that the scaling functions are supported in $[-1, 1]$. The spaces V_p for $p \geq 0$ are examples of FSI spaces, and we develop a necessary and sufficient condition for an FSI space with generators supported in $[-1, 1]$ to be orthogonal. In §3, we give a construction that, under certain conditions, gives an orthogonal intertwining multiresolution analysis. We also provide an example based on continuous piecewise-linear splines.

In §4, we first prove that if certain orthogonal projections are uniformly bounded below, then the construction of §3 works and Theorem 1 holds with $n = 1$. We then use this special case to prove Theorem 1. In §5, we apply the general theory to the piecewise-polynomial spline spaces $\mathcal{S}_{d,r}$ and give an example of orthogonal C^1 cubic spline scaling functions and wavelets.

2. Orthogonal FSI spaces.

If Φ is a subset of $L^2(\mathbf{R})$, let $\tau(\Phi) = \{\phi(\cdot - n) \mid n \in \mathbf{Z}, \phi \in \Phi\}$ denote the set of integer translates of elements in Φ and let $\sigma(\Phi)$ denote the L^2-closure of the linear span of $\tau(\Phi)$. Following [2], we call a space $V \subset L^2(\mathbf{R})$ an FSI space if $V = \sigma(\Phi)$ for some finite set Φ.

Suppose that (V_p) is a multiresolution analysis generated by r scaling functions. Then V_p for $p \geq 0$ is an FSI space with $2^p r$ generators. In particular, if the scaling functions $\{\phi^1, \ldots, \phi^r\}$ are all supported in $[-L, 1]$, then V_1 is generated by the set of

$2r$ scaling functions $\{\phi^s(2\cdot -n), \mid s=1,\ldots,r;\ n=0,1\}$ all supported in $[-L/2,1]$. Hence there is some q such that the multiresolution analysis (V_p') defined by $V_p' = V_{p+q}$ is generated by $2^q r$ scaling functions all supported in $[-1,1]$.

Thus we will be concerned with FSI spaces whose generators $\Phi = \{\phi^1,\ldots,\phi^r\}$ are supported in $[-1,1]$.

Remark 2.1. Certain linear dependencies can be removed as follows: if the restrictions $\phi^s|_{[-1,0]}$ and $\phi^{s'}|_{[-1,0]}$ are linearly dependent on $[-1,0]$, then by taking linear combinations, we can replace one of the generators with a function supported on $[0,1]$. By making such replacements, we can assume that there are k generators $\phi^1,\ldots,\phi^k$ supported in $[-1,1]$ such that

1. they are linearly independent on $[-1,0]$,
2. they are linearly independent on $[0,1]$,
3. the rest of the generators $\phi^{k+1},\ldots,\phi^r$ are supported in $[0,1]$.

Let $H_h(\Phi) = \operatorname{span}\{\phi^s(\cdot-h)\chi_{[0,1]} \mid s=1,\ldots,k\}$ for $h=0,1$. If, in addition to satisfying conditions 1–3 above, we have $H_0(\Phi)\cap H_1(\Phi)=\{0\}$, then

$$\phi^1,\ldots,\phi^k,\phi^1(\cdot-1),\ldots,\phi^k(\cdot-1),\phi^{k+1},\ldots,\phi^r$$

are linearly independent on $[0,1]$, and we say that $\phi^1,\ldots,\phi^r$ are *minimally supported on* $[-1,1]$.

Remark 2.2. In the following, $k=k(\Phi)$ will always represent the number of scaling functions supported on $[-1,1]$ but not supported on $[-1,0]$ or $[0,1]$.

The following lemma, whose proof is delayed until §4, states that any multiresolution analysis generated by compactly supported scaling functions can be converted (as in Remark 2.2) so that it is generated by scaling functions that are minimally supported on $[-1,1]$.

LEMMA 2.1. *Suppose that (V_p) is a multiresolution analysis generated by compactly supported scaling functions. Then there are some n and some set of scaling functions minimally supported on $[-1,1]$ that generate the multiresolution analysis (V_p') given by*

$$V_p' = V_{p+n}.$$

Let V be an FSI space generated by a collection Φ of r scaling functions minimally supported on $[-1,1]$. We say that V is *orthogonal with respect to* $[-1,1]$ if there is some Φ' supported on $[-1,1]$ such that $\tau(\Phi')$ is an orthogonal basis for V' (which is some dilate of V). We will give a necessary and sufficient condition that V be orthogonal with respect to $[-1,1]$. Towards this end, let $\mathcal{A}(V),\mathcal{B}_0(V),\mathcal{B}_1(V)\subset L^2(\mathbf{R})$ be the subspaces defined by

$$\mathcal{A}(V) = \operatorname{span}\{\phi^s \mid s=k+1,\ldots,r\}$$

and

$$\mathcal{B}_h(V) = \operatorname{span}\left(\{\phi^s(\cdot-h)\chi_{[0,1]} \mid s=1,\ldots,k\}\cup\mathcal{A}(V)\right) \text{ for } h=0,1,$$

where $\chi_{[0,1]}$ denotes the characteristic function of $[0,1]$. Finally, define $\mathcal{C}_h(V)$ by $\mathcal{C}_h(V) = \mathcal{B}_h(V)\ominus\mathcal{A}(V)$, the orthogonal complement of $\mathcal{A}(V)$ in $\mathcal{B}_h(V)$.

Remark 2.3.

1. It follows from Remark 2.1 that the spaces $\mathcal{A}(V)$, $\mathcal{B}_0(V)$, and $\mathcal{B}_1(V)$ are independent of the choice of minimally supported generators. For instance, we have

$$\mathcal{A}(V) = \{f \in V \mid \operatorname{supp} f \subset [0,1]\},$$

$$\mathcal{B}_0(V) = \{f\chi_{[0,1]} \mid f\chi_{(-\infty,1]} \in V\}$$

and similarly for $\mathcal{B}_1(V)$.

2. $\mathcal{C}_0(V)$ and $\mathcal{C}_1(V)$ are k dimensional.

LEMMA 2.2. *Suppose that V is an FSI space generated by r functions $\phi^1, \ldots, \phi^r$ minimally supported on $[-1,1]$. Then V is orthogonal with respect to $[-1,1]$ iff $\mathcal{C}_0(V) \perp \mathcal{C}_1(V)$.*

Proof. If V is orthogonal, then it is easy to see directly that $\mathcal{C}_0(V) \perp \mathcal{C}_1(V)$.

Suppose $\mathcal{C}_0(V) \perp \mathcal{C}_1(V)$. By using the Gram–Schmidt procedure, if necessary, we can assume that

- $\phi^s(\cdot - h)$ is orthogonal to $\mathcal{A}(V)$ for $s = 1, \ldots, k$, and $h = 0, 1$;
- $\phi^s \perp \phi^{s'}$ for $s \neq s'$ and $s, s' = 1, \ldots, r$.

Since $\mathcal{C}_0(V) \perp \mathcal{C}_1(V)$, it follows that $\phi^s \perp \phi^{s'}(\cdot - 1)$ for $s, s' = 1, \ldots, k$, and hence $\{\phi^s(\cdot - n) \mid n \in \mathbf{Z}, s = 1, \ldots, r\}$ is an orthogonal basis for V. □

3. Constructing orthogonal multiresolution analyses. The following lemma provides the basic idea we will use to construct orthogonal multiresolution analyses.

LEMMA 3.1. *Let (V_p) be a multiresolution analysis generated by r functions $\phi^1, \ldots, \phi^r$ minimally supported on $[-1,1]$. Suppose there is a subspace $\mathcal{W}$ of $\mathcal{A}(V_1) \ominus \mathcal{A}(V_0)$ such that*

$$(I - P_{\mathcal{W}})\mathcal{C}_0(V_0) \perp (I - P_{\mathcal{W}})\mathcal{C}_1(V_0), \tag{1}$$

where $P_{\mathcal{W}}$ is the orthogonal projection onto $\mathcal{W}$. Let $w_1, \ldots, w_\kappa$ be a basis for $\mathcal{W}$ and let $(\tilde{V}_p)$ be the multiresolution analysis generated by $\phi^1, \ldots, \phi^r, w_1, \ldots, w_\kappa$. Then $(\tilde{V}_p)$ is an orthogonal multiresolution analysis and

$$V_0 \subset \tilde{V}_0 \subset V_1.$$

Proof. Since $\mathcal{W} \subset V_1$, we clearly have $V_0 \subset \tilde{V}_0 \subset V_1$. This implies $V_p \subset \tilde{V}_p \subset V_{p+1}$ for all integers p. It then follows that $(\tilde{V}_p)$ satisfies the first four conditions of a multiresolution analysis.

Observe that $\mathcal{C}_h(\tilde{V}_0) = (I - P_{\mathcal{W}})\mathcal{C}_h(V_0)$ for $h = 0, 1$, and so the result follows from Lemma 2.2. □

Remark 3.1. Let P_h denote the orthogonal projection onto $\mathcal{C}_h(V_0)$ for $h = 0, 1$. Condition (1) can be rewritten as

$$P_0 P_1 = P_0 P_{\mathcal{W}} P_1. \tag{2}$$

Therefore, the rank of $P_{\mathcal{W}}$ must be at least the rank of $P_0 P_1$, which is less than or equal to k. This gives a lower bound on κ, the dimension of $\mathcal{W}$ (which equals the rank of $P_{\mathcal{W}}$). In all of our examples, $P_0 P_1$ has rank k and, it turns out, we can choose κ to be the lower bound k.

We will use the following notation in the rest of this paper:

$$\phi_{l,k}(x) = 2^{l/2}\phi(2^l x - k) \quad \text{for any } \phi \in L^2(\mathbf{R}). \tag{3}$$

3.1. Example: Piecewise-linear orthogonal scaling functions. Let $H : \mathbf{R} \to \mathbf{R}$ be the "hat" function defined by

$$H(x) = \begin{cases} 1 - |x| & \text{if } |x| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\phi^1(x) = \sqrt{3}H(2x)$ and $\phi^2(x) = \sqrt{3}H(2x-1)$ (note that ϕ^1 and ϕ^2 have norm 1) and let (V_p) be the multiresolution analysis generated by $\Phi = \{\phi^1, \phi^2\}$. In this case, V_{-1} is the space of continuous piecewise-linear splines with integer knots. Note that Φ is minimally supported on $[-1,1]$ with $k = k(\Phi) = 1$. Also, $\mathcal{A}(V_0)$, $\mathcal{B}_0(V_0)$, and $\mathcal{B}_1(V_0)$ are spanned by $\{\phi^2_{0,0}\}$, $\{\phi^1_{0,0}\chi_{[0,1]}, \phi^2_{0,0}\}$, and $\{\phi^1_{0,1}\chi_{[0,1]}, \phi^2_{0,0}\}$, respectively. Using

$$c = \langle \phi^1_{0,0}, \phi^2_{0,0} \rangle = 1/4$$

gives the bases $\{\phi^1_{0,0}\chi_{[0,1]} - c\phi^2_{0,0}\}$ and $\{\phi^1_{0,1}\chi_{[0,1]} - c\phi^2_{0,0}\}$ for $\mathcal{C}_0(V_0)$ and $\mathcal{C}_1(V_0)$, respectively.

Note that $\{\phi^1_{1,1}, \phi^2_{1,0}, \phi^2_{1,1}\}$ is a basis for $\mathcal{A}(V_1)$. Direct calculation yields the basis $\{3\phi^1_{1,1} - 5\phi^2_{1,0}, 3\phi^1_{1,1} - 5\phi^2_{1,1}\}$ for $\mathcal{A}(V_1) \ominus \mathcal{A}(V_0)$. Let w be an arbitrary vector in $\mathcal{A}(V_1) \ominus \mathcal{A}(V_0)$:

$$w = a(3\phi^1_{1,1} - 5\phi^2_{1,0}) + b(3\phi^1_{1,1} - 5\phi^2_{1,1}),$$

where we will choose w so that equation (1) is satisfied when $\mathcal{W}$ is the one-dimensional subspace spanned by w. Then

$$\|w\|^2 = (53a^2 + 6ab + 53b^2)/2$$

and equation (1) holds iff

$$\begin{aligned} (4) \qquad & \langle \phi^1_{0,0}\chi_{[0,1]} - c\phi^2_{0,0}, \phi^1_{0,1}\chi_{[0,1]} - c\phi^2_{0,0} \rangle \\ & = \langle \phi^1_{0,0}\chi_{[0,1]} - c\phi^2_{0,0}, w \rangle \langle w, \phi^1_{0,1}\chi_{[0,1]} - c\phi^2_{0,0} \rangle / \|w\|^2. \end{aligned}$$

The left-hand side is easily evaluated to get

$$\langle \phi^1_{0,0}\chi_{[0,1]} - c\phi^2_{0,0}, \phi^1_{0,1}\chi_{[0,1]} - c\phi^2_{0,0} \rangle = -c^2 = -1/16.$$

Using $H(x) = 1/2H(2x+1) + H(2x) + 1/2H(2x-1)$, we find

$$\phi^1_{0,0} = (1/2\sqrt{2})\phi^2_{1,-1} + (1/\sqrt{2})\phi^1_{1,0} + (1/2\sqrt{2})\phi^2_{1,0},$$

$$\phi^2_{0,0} = (1/2\sqrt{2})\phi^2_{1,0} + (1/\sqrt{2})\phi^1_{1,1} + (1/2\sqrt{2})\phi^2_{1,1},$$

which can be used to calculate

$$\langle \phi^1_{0,0}\chi_{[0,1]} - c\phi^2_{0,0}, w \rangle = (3/8\sqrt{2})(b - 9a)$$

and

$$\langle w, \phi^1_{0,1}\chi_{[0,1]} - c\phi^2_{0,0} \rangle = (3/8\sqrt{2})(a - 9b).$$

Thus equation (4) becomes

$$-4(53a^2 + 6ab + 53b^2) = 9(a - 9b)(-9a + b),$$

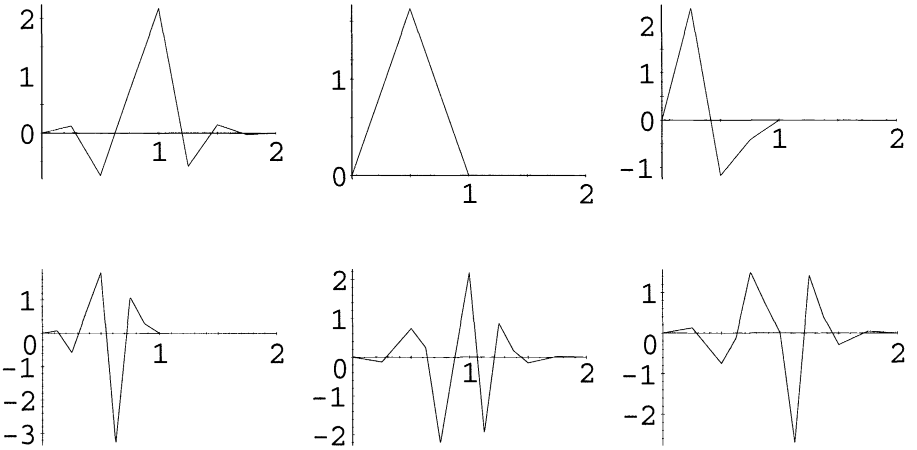


FIG. 1. *Piecewise-linear orthogonal scaling functions (top, right to left: $\tilde\phi^1(\cdot - 1)$, $\tilde\phi^2$, and $\tilde\phi^3$) and wavelets (bottom, right to left: $\tilde\psi^1$, $\tilde\psi^2$, and $\tilde\psi^3(\cdot - 1)$).*

which has the following solutions:

$$b/a = (-762 \pm 320\sqrt{5})/262.$$

We choose $b/a = (-762 + 320\sqrt{5})/262$. Set $\tilde\phi^2 = \phi^2$, $\tilde\phi^3 = w/\|w\|$,

$$y = \phi^1 - \langle \phi^1, \tilde\phi^2\rangle\tilde\phi^2 - \langle \phi^1, \tilde\phi^3\rangle\tilde\phi^3 - \langle \phi^1, \tilde\phi^2(\cdot + 1)\rangle\tilde\phi^2(\cdot + 1) - \langle \phi^1, \tilde\phi^3(\cdot + 1)\rangle\tilde\phi^3(\cdot + 1),$$

and $\tilde\phi^1 = y/\|y\|$; then from the above calculation, we see that $\tilde\Phi = (\tilde\phi^1, \tilde\phi^2, \tilde\phi^3)$ generates the multiresolution analysis $(\tilde V_p)$ with $V_0 \subset \tilde V_0 \subset V_1$. These orthonormal scaling functions and associated wavelets are shown in Figure 1. The wavelets have been constructed so that the techniques used in Theorem 4.4 of [7] lead to a wavelet basis for $L_2[0, 1]$. The interpolation points for these functions are given in Table 1 in Appendix A and the dilation coefficients are given in Table 4 in Appendix B.

3.2. Example: Symmetric piecewise-linear orthogonal scaling functions. We now choose $\tilde V_0$ between V_1 and V_2. Then $\mathcal{A}(V_2) \ominus \mathcal{A}(V_1)$ is $7 - 3 = 4$ dimensional. By applying the intertwining technique twice, we can choose symmetric w_1 and antisymmetric w_2 to obtain scaling functions and wavelets that are either symmetric or antisymmetric, as shown in Figure 2. In this case, $\tilde\Phi = (\tilde\phi^1, \tilde\phi^2, \tilde\phi^3, \tilde\phi^4)$, where $\tilde\phi^2 = \phi^2$, $\tilde\phi_3 = w_1/\|w_1\|$, $\tilde\phi_4 = w_2/\|w_2\|$, and $\tilde\phi^1 = y/\|y\|$. Here

$$y = \phi^1 - \sum_{i=2}^{4} \langle \phi^1, \tilde\phi^i\rangle\tilde\phi^i - \sum_{i=2}^{4} \langle \phi^1, \tilde\phi^i(\cdot + 1)\rangle\tilde\phi^i(\cdot + 1).$$

Because of the symmetry, it is not difficult to generate a wavelet basis for $L_2[0, 1]$ from the wavelets given below [7, Thm. 4.4].

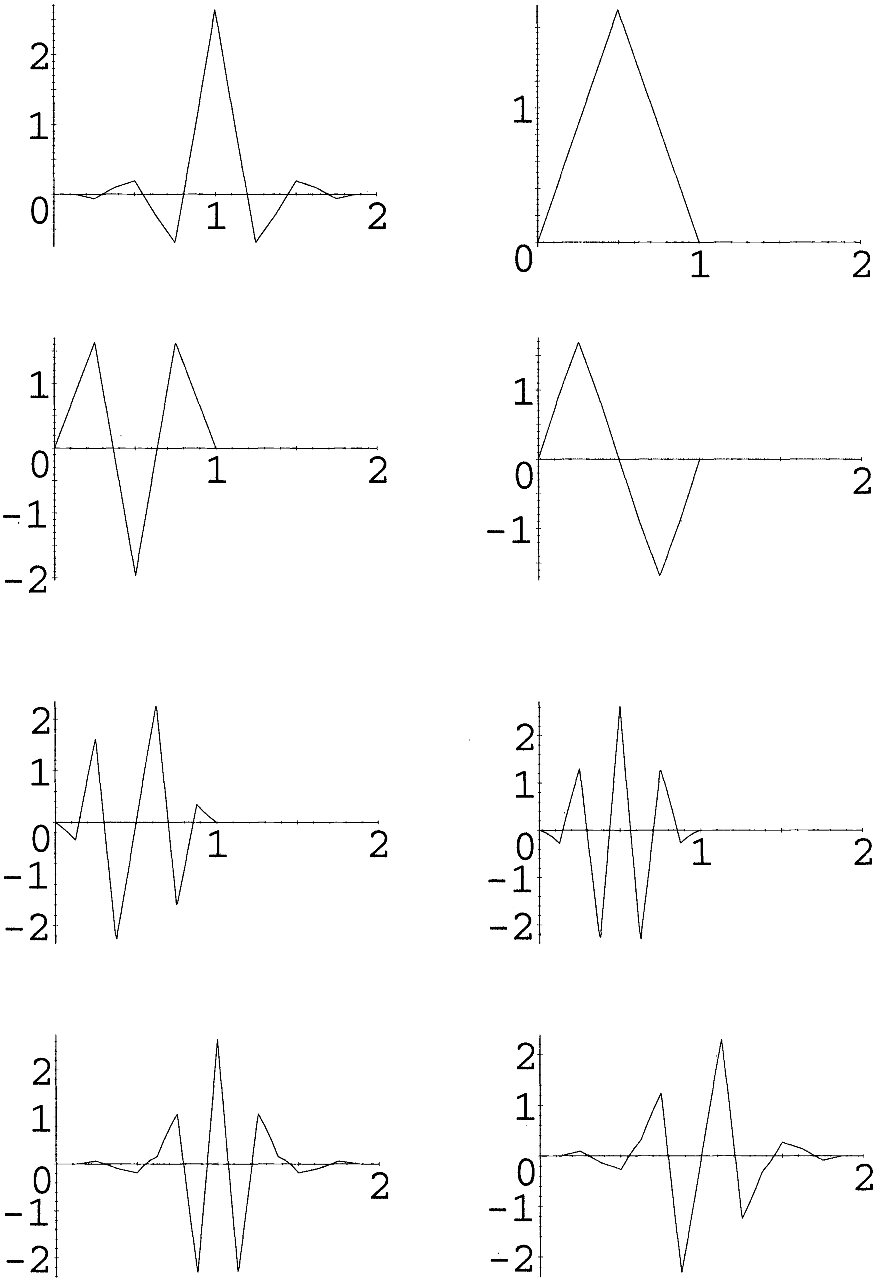


FIG. 2. *Symmetric* (*or antisymmetric*) *piecewise-linear orthogonal scaling functions and wavelets; from top to bottom and left to right:* $\tilde{\phi}^1(\cdot - 1)$, $\tilde{\phi}^2$, $\tilde{\phi}^3$, $\tilde{\phi}^4$, $\tilde{\psi}^1$, $\tilde{\psi}^2$, $\tilde{\psi}^3(\cdot - 1)$, *and* $\tilde{\psi}^4(\cdot - 1)$.

4. Main theorem. We first prove that, with an additional hypothesis, Theorem 1 holds with $n = 1$.

THEOREM 2. *Let (V_p) be a multiresolution analysis generated by r functions $\phi^1, \dots, \phi^r$ minimally supported on $[-1, 1]$. Let Q^l denote the orthogonal projection onto $\mathcal{A}(V_{l+1}) \ominus \mathcal{A}(V_l)$. Suppose there are positive numbers ϵ and l_0 such that $\|Q^l u\| \geq \epsilon\|u\|$ for all $l \geq l_0$ and u in $\mathcal{C}_0(V_l) \cup \mathcal{C}_1(V_l)$. Then there is some integer q and some orthogonal multiresolution analysis $(\tilde{V}_p)$ such that*

$$V_q \subset \tilde{V}_0 \subset V_{q+1}.$$

Proof. As noted in the proof of Lemma 2.2 we can assume that

- $\phi^s(\cdot - h)\chi_{[0,1]}$ is orthogonal to $\mathcal{A}(V_0)$ for $s = 1, \dots, k$ and $h = 0, 1$, and
- $\phi^s \perp \phi^{s'}$ for $s \neq s'$ and $s, s' = 1, \dots, r$.

For $1 \leq s \leq r$ and $l, n \in \mathbf{Z}$, let $\phi^s_{l,n}(x) = 2^{l/2}\phi^s(2^l x - n)$. Then

$$\Lambda = \{\phi^s_{l,n} \mid 1 \leq s \leq k, 1 \leq n \leq 2^l - 1\} \cup \{\phi^s_{l,n} \mid k+1 \leq s \leq r, 0 \leq n \leq 2^l - 1\}$$

is a basis for $\mathcal{A}(V_l)$,

$$\{\phi^s_{l,0}\chi_{[0,1]} \mid 1 \leq s \leq k\} \cup \Lambda$$

is a basis for $\mathcal{B}_0(V_l)$, and

$$\{\phi^s_{l,2^l}\chi_{[0,1]} \mid 1 \leq s \leq k\} \cup \Lambda$$

is a basis for $\mathcal{B}_1(V_l)$. Observe that $\mathcal{A}(V_l)$, $\mathcal{B}_h(V_l)$, and $\mathcal{C}_h(V_l)$ have linear dimensions $2^l r - k$, $2^l r$, and k, respectively.

With $\Phi_{l,n} = (\phi^1_{l,n}, \dots, \phi^k_{l,n})^\top$, we define the overlap matrix, $C = \langle \Phi_{0,0}, \Phi_{0,1}\rangle := (\langle \phi^s_{0,0}, \phi^{s'}_{0,1}\rangle)_{1 \leq s, s' \leq k}$. Let $l_2(\mathbf{R}^k)$ be the Hilbert space of sequences $y = \{y_n\}^{\infty}_{-\infty}$, $y_n \in \mathbf{R}^k$ with $\|y\|_{l_2} = (\sum(y_n, y_n))^{1/2} < \infty$ where (y_n, y_n) is the Euclidean inner product. For $y, x \in l_2(\mathbf{R}^k)$ set $\langle x, y\rangle_{l_2} = \sum(x_n, y_n)$. We first prove the following three lemmas before proceeding with the remainder of the proof of Theorem 2.

LEMMA 4.1. *Let $x = \{x_n\}^{\infty}_{-\infty} \in l_2(\mathbf{R}^k)$ and consider the infinite block-tridiagonal matrix $\mathcal{J} : l_2(\mathbf{R}^k) \to l_2(\mathbf{R}^k)$ given by $\mathcal{J}x = y$, $y = \{y_n\}^{\infty}_{-\infty}$, with $y_i = C^\top x_{i-1} + x_i + Cx_{i+1}$ for all $i \in \mathbf{Z}$. Then $\mathcal{J}$ is a bounded positive operator with* spectrum$\mathcal{J} \subset [a, b]$, *$0 < a$. Consequently, $\mathcal{J}_l = \Pi_l \mathcal{J} \Pi_l$, where $\Pi_l : l_2(\mathbf{R}^k) \to l_2(\mathbf{R}^k)$ is the projection such that $\Pi_l x = \{x_n\}^{2^l-1}_1$ is positive definite with* spectrum$\mathcal{J}_l \subset [a, b]$ *and $|(\mathcal{J}_l)^{-1}_{i,j}| \leq K\lambda^{|i-j|}$. Here K is independent of l, i, and j, and $\lambda = (\sqrt{b/a} - 1)/(\sqrt{b/a} + 1)$.*

Proof. Since $\langle \phi^i, \phi^j\rangle = 0$, $i = 1, \dots, k$, $j = k+1, \dots, r$, and $\langle \phi^i, \phi^j\rangle = 0$, $i = k+1, \dots, r$, $j = k+1, \dots, r$, $i \neq j$, it follows from Theorem 3.2 in [8] (also see [10]) that $\{\phi^i\}^r_{i=1}$ forms a Riesz basis for V_0 iff $(I + C^\top e^{i\theta} + Ce^{-i\theta}) \geq a > 0$, $0 \leq \theta < 2\pi$. Now the symbol of y, $\hat{y} = \widehat{(\mathcal{J}x)} = (I + C^\top e^{i\theta} + Ce^{-i\theta})\hat{x}$, and we find $\langle x, \mathcal{J}x\rangle_{l_2} = (1)/(2\pi)\int_0^{2\pi} \hat{x}^*(I + C^\top e^{i\theta} + Ce^{-i\theta})\hat{x}d\theta \geq a\int_0^{2\pi} \hat{x}^*\hat{x}d\theta = a\|x\|^2_{l_2}$. The upper bound follows in an analogous manner. The min–max principle [19] now implies that spectrum $\mathcal{J}_l \subset [a, b]$ and the decay given in the lemma follows from the exponential decay of elements of inverses of banded matrices (cf. [6], [9]). □

Since Q^l is orthogonal, we can choose $\{u^l_{h,i} \mid 1 \leq i \leq k\}$ to be an orthonormal basis for $\mathcal{C}_h(V_l)$, $h = 0, 1$, such that $\{Q^l u^l_{h,i} \mid 1 \leq i \leq k\}$ is an orthogonal basis for $Q^l\mathcal{C}_h(V_l)$

for $h = 0, 1$. We also define the normalized basis elements $x^l_{h,i} = Q^l u^l_{h,i}/\|Q^l u^l_{h,i}\|$ for $1 \le i \le k$ and $h = 0, 1$.

LEMMA 4.2. *Suppose there is some $\epsilon > 0$ such that $\|Q^l u\| \ge \epsilon\|u\|$ for all l sufficiently large and for u in either $\mathcal{C}_0(V_l)$ or $\mathcal{C}_1(V_l)$. Let $x^l_{h,i}$ and $u^l_{h,i}$ be as above. Then for $1 \le i, j \le k$,*

$$\lim_{l\to\infty} \langle x^l_{0,i}, x^l_{1,j}\rangle = 0 \tag{5}$$

and

$$\lim_{l\to\infty} \frac{\langle u^l_{0,i}, u^l_{1,j}\rangle}{\|Q^l u^l_{0,i}\|\|Q^l u^l_{1,j}\|} = 0. \tag{6}$$

Proof. Observe that

$$\langle \Phi_{l,n}, \Phi_{l,m}\rangle = \begin{cases} C & \text{if } m-n=1, \\ I & \text{if } m=n, \\ C^{\mathrm{T}} & \text{if } m-n=-1. \end{cases}$$

Let $f \in \mathcal{C}_0(V_l)$. Using the bases given above for $\mathcal{A}(V_l)$ and $\mathcal{C}_0(V_l)$, f may be expressed in the form

$$f = \gamma_0^{\mathrm{T}} \Phi_{l,0}\chi_{[0,1]} + \sum_{n=1}^{2^l-1} \gamma_n^{\mathrm{T}} \Phi_{l,n}, \tag{7}$$

where each γ_n is in $\mathbf{R}^k$. Since f is perpendicular to $\mathcal{A}(V_l)$, it follows that $\gamma = (\gamma_1^{\mathrm{T}}, \dots, \gamma_{2^l-1}^{\mathrm{T}})^{\mathrm{T}}$ satisfies the block-tridiagonal system of equations

$$\begin{gathered} \gamma_1 + C\gamma_2 = -C^{\mathrm{T}}\gamma_0, \\ C^{\mathrm{T}}\gamma_{n-1} + \gamma_n + C\gamma_{n+1} = 0, \quad n = 2, 3, \dots, 2^l-2, \\ C^{\mathrm{T}}\gamma_{2^l-2} + \gamma_{2^l-1} = 0. \end{gathered} \tag{8}$$

If we define $\mathcal{J}_l$ to be the matrix representation of the left-hand side of equation (8), we find that $\gamma = \mathcal{J}_l^{-1}(-\gamma_0^{\mathrm{T}} C, 0, \dots, 0)^{\mathrm{T}}$. Thus f is determined by γ_0 (and l) and we denote the dependence of f on γ_0 and l by $f_{\gamma_0,l}$. Let $\lambda = (\sqrt{b/a}-1)/(\sqrt{b/a}+1)$. It follows from Lemma 4.1 that there are positive numbers K and l_0 such that $\|\gamma_n\| \le K\lambda^n$ for $n \ge 0$, $l \ge l_0$, and $\|f_{\gamma_0,l}\| = 1$. Hence

$$\lim_{l\to\infty} \max\{\|f\chi_{[1/2,1]}\| \mid f \in \mathcal{C}_0(V_l), \|f\| = 1\} = 0. \tag{9}$$

Similarly,

$$\lim_{l\to\infty} \max\{\|g\chi_{[0,1/2]}\| \mid g \in \mathcal{C}_1(V_l), \|g\| = 1\} = 0. \tag{10}$$

Equations (9) and (10) then give

$$\lim_{l\to\infty} |\langle u^l_{0,i}, u^l_{1,j}\rangle| \le \lim_{l\to\infty} \|u^l_{0,i}\chi_{[1/2,1]}\|\|u^l_{1,j}\| + \|u^l_{0,i}\|\|u^l_{1,j}\chi_{[0,1/2]}\| = 0.$$

Since Q^l is uniformly bounded below on $\mathcal{C}_h(V_l)$ for $h = 0, 1$ and l large enough, this proves the limit (6).

If $f \in \mathcal{C}_0(V_l)$, then $Q^l f = f-(I-P_{l+1})f$, where P_{l+1} is the orthogonal projection onto $\mathcal{A}(V_{l+1})$, and so $(I-P_{l+1})f$ is in $\mathcal{C}_0(V_{l+1})$. Therefore, both f and $(I-P_{l+1})f$ decay exponentially away from 0. Thus, as above, the limit in (5) follows from equations (9) and (10) and the uniform lower bound on Q^l. □

Let $\mathcal{D}_l$ denote $\operatorname{span}(Q^l\mathcal{C}_0(V_l) \cup Q^l\mathcal{C}_1(V_l))$.

LEMMA 4.3. *There is some q such that for $l \geq q$ there exists a set $\Psi = \{z^{\pm}_{i,j} \mid 1 \leq i,j \leq k\}$ of $2k^2$ functions in $\mathcal{E}_l := (\mathcal{A}(V_{l+1}) \ominus \mathcal{A}(V_l)) \ominus \mathcal{D}_l$ such that if*

$$w^{\pm}_{i,j} = z^{\pm}_{i,j} + (x^l_{0,i} \pm x^l_{1,j})/(4k),$$

then $\{w^{\pm}_{i,j} \mid 1 \leq i,j \leq k\}$ is an orthonormal set.

Proof. Note that $\dim(\mathcal{A}(V_{l+1}) \ominus \mathcal{A}(V_l)) = r2^l$. Let $\{e_i \mid 1 \leq i \leq r2^l\}$ be an orthonormal basis of $\mathcal{A}(V_{l+1}) \ominus \mathcal{A}(V_l)$. Represent $w^{\pm}_{i,j}, z^{\pm}_{i,j}, (x^l_{0,i} \pm x^l_{1,i})/(4k)$ in terms of this basis as column vectors of size $r2^l$. Let W^+_l be the $r2^l \times k^2$ matrix whose $(i-1)k+j$ column is $w^+_{i,j}$. Similarly, define W^-_l and $Z^{\pm}_l$. Let $X^{\pm}_l$ be the $r2^l \times k^2$ matrix whose $(i-1)k+j$ column is $(x^l_{0,i} \pm x^l_{1,j})/(4k)$. In block form, we have

$$W_l := \begin{pmatrix} W^+_l & W^-_l \end{pmatrix} = \begin{pmatrix} X^+_l + Z^+_l & X^-_l + Z^-_l \end{pmatrix} =: X_l + Z_l. \tag{11}$$

Note that $\{w^{\pm}_{i,j} \mid 1 \leq i,j \leq k\}$ is an orthonormal set iff $W_l^{\top} W_l = I$, where I is the $2k^2 \times 2k^2$ identity. Since $z^{\pm}_{i,j} \in \mathcal{D}^{\perp}_l$, we have $X_l^{\top} Z_l = \mathbf{0}$. Hence $W_l^{\top} W_l = I$ iff $Z_l^{\top} Z_l = I - X_l^{\top} X_l$. We can solve this for Z_l (using Cholesky factorization, for instance) if the right-hand side $I - X_l^{\top} X_l$ is positive definite (since it is clearly symmetric).

Let q be large enough so that $|\langle x^l_{0,i}, x^l_{1,j}\rangle| < 1/4$ for $l \geq q$. Then for $l \geq q$, the absolute values of the elements of $X_l^{\top} X_l$ are of the form $|\langle x^l_{0,i} \pm x^l_{1,j}, x^l_{0,m} \pm x^l_{1,n}\rangle|/(16k^2) < 3/(16k^2)$. Thus any row or column sum of the absolute values of the elements of $X_l^{\top} X_l$ is less than $3/8 < 1$, implying that $I - X_l^{\top} X_l$ is diagonally dominant with positive terms on the diagonal and hence is positive definite. □

Let q be as in Lemma 4.3 and $l \geq q$. For $1 \leq i,j \leq k$, let $w^l_{i,j} = \alpha_{i,j} w^+_{i,j} + \beta_{i,j} w^-_{i,j}$, where $\alpha^2_{i,j} + \beta^2_{i,j} = 1$ and $\alpha_{i,j}, \beta_{i,j} \in [0,1]$. Let $\mathcal{W} = \operatorname{span}\{w^l_{i,j} \mid 1 \leq i,j \leq k\}$. Then

$$\begin{aligned}
\langle u^l_{0,i}, P_{\mathcal{W}} u^l_{1,j}\rangle &= \sum_{m,n=1}^{k} \langle u^l_{0,i}, w^l_{m,n}\rangle\langle w^l_{m,n}, u^l_{1,j}\rangle \\
&= \|Q^l u^l_{0,i}\|\|Q^l u^l_{1,j}\| \sum_{m,n=1}^{k} \langle x^l_{0,i}, w^l_{m,n}\rangle\langle w^l_{m,n}, x^l_{1,j}\rangle \\
&= \frac{\|Q^l u^l_{0,i}\|\|Q^l u^l_{1,j}\|}{16k^2} \sum_{m,n=1}^{k} \langle x^l_{0,i}, (\alpha_{m,n}+\beta_{m,n})x^l_{0,m} + (\alpha_{m,n}-\beta_{m,n})x^l_{1,n}\rangle \\
&\quad \cdot\langle(\alpha_{m,n}+\beta_{m,n})x^l_{0,m} + (\alpha_{m,n}-\beta_{m,n})x^l_{1,n}, x^l_{1,j}\rangle \\
&= \frac{\|Q^l u^l_{0,i}\|\|Q^l u^l_{1,j}\|}{16k^2} \Bigg\{ \alpha^2_{i,j} - \beta^2_{i,j} \\
&\quad + \langle x^l_{0,i}, x^l_{1,j}\rangle \left(\sum_{n=1}^{k} (\alpha_{i,n}+\beta_{i,n})^2 + \sum_{m=1}^{k} (\alpha_{m,j}-\beta_{m,j})^2 \right) \\
&\quad + \sum_{m,n=1}^{k} (\alpha^2_{m,n} - \beta^2_{m,n}) \langle x^l_{0,i}, x^l_{1,n}\rangle\langle x^l_{o,m}, x^l_{1,j}\rangle \Bigg\}.
\end{aligned}$$

Let $\gamma_{i,j} = \alpha_{i,j}^2$ and $\underline{\gamma} = (\gamma_{i,j}) \in [0,1]^{k\times k}$. Then $\alpha_{i,j} = \sqrt{\gamma_{i,j}}$ and $\beta_{i,j} = \sqrt{1-\gamma_{i,j}}$. We can rewrite the above calculation as

$$\frac{16k^2}{\|Q^l u_{0,i}^l\| \, \|Q^l u_{1,j}^l\|} \langle u_{0,i}^l, P_{\mathcal{W}} u_{1,j}^l \rangle = 2\gamma_{i,j} - 1 + 2f_{i,j}^l(\underline{\gamma}), \tag{12}$$

where $f_{i,j}^l(\underline{\gamma})$ is continuous for $\underline{\gamma} \in [0,1]^{k\times k}$. Furthermore, if $|\langle x_{0,i}^l, x_{1,n}^l \rangle| < \epsilon$, then $|f_{i,j}^l(\underline{\gamma})| < 4k\epsilon + k^2\epsilon^2$. Let $\theta_{i,j}^l = (1+((16k^2)/(\|Q^l u_{0,i}^l\|\|Q^l u_{1,j}^l\|))\langle u_{0,i}^q, u_{1,j}^q \rangle)/2$. Then equation (2) is equivalent to

$$\theta_{i,j}^l - f_{i,j}^l(\underline{\gamma}) = \gamma_{i,j} \tag{13}$$

or

$$\underline{\theta}^l - \underline{f}^l(\underline{\gamma}) = \underline{\gamma}, \tag{14}$$

where $\underline{\theta}^l = (\theta_{i,j}^l) \in \mathbf{R}^{k\times k}$ and $\underline{f}^l = (f_{i,j}^l) : [0,1]^{k\times k} \to \mathbf{R}^{k\times k}$. Using Lemma 3.1, the proof will be finished once it is shown that equation (14) has a solution.

By Lemma 4.2, we can find l large enough so that $\underline{\theta}^l \in [1/3, 2/3]^{k\times k}$ and so that $\underline{f}^l(\underline{\gamma}) \in [-1/3, 1/3]^{k\times k}$ for $\underline{\gamma} \in [0,1]^{k\times k}$. Then for l large enough,

$$\underline{g} := \underline{\theta}^l - \underline{f}^l$$

maps the compact, convex set $[0,1]^{k\times k}$ into itself continuously. By the Schauder fixed-point theorem, equation (14) has a solution $\underline{\gamma} \in [0,1]^{k\times k}$ and thus Theorem 2 is proved. □

By a suitable modification of the original multiresolution analysis, a new multiresolution analysis $(\hat{V}_p)$ may be constructed with

$$V_0 \subset \hat{V}_0 \subset V_n$$

such that the hypotheses of Lemma 4.2 are satisfied. First note that by Lemma 2.1, any multiresolution analysis generated by compactly supported functions may be assumed to have generators that are minimally supported in $[-1,1]$. Next, let $V_l[a,b] = \{\phi \in V_l : \text{supp}\phi \subset [a,b]\}, T^\infty = \{\phi \in V_0 : \phi \perp V_0(-\infty,0] \oplus V_0[0,\infty)\}, T_0^\infty = \{\phi\chi_{[0,\infty)} : \phi \in T^\infty\}$, and $T_1^\infty = \{\phi\chi_{(-\infty,0]} : \phi \in T^\infty\}$. If all the functions in T^∞ have bounded support, then for l chosen large enough so that $[-2^l, 2^l]$ supports T^∞, $\tilde{V}_0 = V_l$ gives rise to a compactly supported orthogonal multiresolution analysis. Consequently, it is only those functions in T^∞ that have unbounded support that need to be altered. To this end, define $T_0^F = \{\phi \in T_0^\infty : \text{supp}\phi \text{ is bounded}\}, T_1^F = \{\phi \in T_1^\infty : \text{supp}\phi \text{ is bounded}\}, T_0^\omega = T_0^\infty \ominus T_0^F$, and $T_1^\omega = T_1^\infty \ominus T_1^F$ with $k_1 = \dim T_0^F = \dim T_1^F$ and $k_2 = \dim T_0^\omega = \dim T_1^\omega$; also set $\mathcal{C}_0^l = \mathcal{C}_0(V_l)$ and $\mathcal{C}_1^l = \mathcal{C}_1(V_l)$.

The uniform-lower-bound condition of Theorem 2 can be replaced by a simpler one by passing to limits. Using the notation in the proof of Lemmas 4.2 and 4.1, it follows that $2^{-l/2} f_{\gamma_0,l}(2^{-l}\cdot)$ converges in $L^2(\mathbf{R})$ to some $f_{\gamma_0,\infty} \in T_0^\infty$. Furthermore, $2^{-l/2}(Q_l f_{\gamma_0,l})(2^{-l}\cdot)$ converges to $Q^\infty f_{\gamma_0,\infty}$, where Q^∞ is the orthogonal projection onto $V_1[0,\infty)$. We then have the following result.

Lemma 4.4. *If Q^∞ is nonsingular on T_0^∞, then Q^l is uniformly bounded below as in Theorem 2.*

We now complete the proof of Theorem 1 with the following lemma.

LEMMA 4.5. *If (V_p) is a multiresolution analysis generated by compactly supported scaling functions, then there are some integer n and some multiresolution analysis $(\hat{V}_p)$ with $V_0 \subset \hat{V}_0 \subset V_n$, such that for ϕ in the unit sphere of $\mathcal{C}_0(\hat{V}_0)$ or $\mathcal{C}_1(\hat{V}_0)$, $\|2^{-l/2}(\hat{P}^{l+1} - \hat{P}^l)\phi(2^l\cdot)\|$ is bounded away from zero for sufficiently large l. Here $\hat{P}^l$ is the othogonal projection onto $\mathcal{A}(\hat{V}_l)$.*

Proof. To construct $\hat{V}_0$ we will "borrow" functions from $V_1, V_2, \ldots$, taking care not to disturb T_0^F or T_1^F. This will be accomplished first by working with functions in the spaces T_0^∞ and T_1^∞ then approximating these functions by members of $T_0^l = \{\phi(2^{-l}\cdot) : \phi \in \mathcal{C}_0^l\}$ and $T_1^l = \{\phi(2^l\cdot) : \phi \in \mathcal{C}_1^l(\cdot + 1)\}$ for sufficiently large l.

Given $\phi \in T_0^\omega$ of unit length, the fact that (V_p) forms a multiresolution analysis implies that $\lim_{l\to\infty} P_\infty^l \phi = \phi$ and hence that $\lim_{l\to\infty} \|P_\infty^l \phi\| = \|\phi\| = 1$, where P_∞^l is the orthogonal projection onto $V_l(-\infty, 0] \oplus V_l[0, \infty)$. Also, since these projections are onto successively larger spaces, this convergence is monotoncially nondecreasing in l. We have continuous functions converging monotonically to a continuous function (i.e., 1) on a compact set, so the convergence must be uniform. Let n_0 be sufficiently large so that $\|P_\infty^l \phi\| > 1/4$ for all ϕ in the unit sphere of T_0^ω and all $l \geq n_0$. We may choose k_2 orthogonal functions $w_1, \ldots, w_{k_2} \in V_{n_0}[0, \infty)$ with span W_0 such that $\|P_{W_0}\phi\| > 1/4$ for all unit functions $\phi \in T_0^\infty$. Furthermore, we may assume that these functions are compactly supported since if they were not, we could approximate them with arbitrary accuracy by functions that are. Observe that W_0 may be shifted as far to the right as desired without destroying its effectiveness. Indeed, given any $\phi \in T_0^\omega$, $\phi(\cdot + j)\chi_{[0,\infty)}$ is again in T_0^ω for any positive integer, j. Thus for $\phi \in T_0^\omega$, $\|P_{W_0(\cdot - j)}\phi\| = \|P_{W_0}\phi(\cdot + j)\| = \|P_{W_0}(\phi(\cdot + j)\chi_{[0,\infty)})\|$. This last expression is a continuous function of ϕ, which, by the construction of W_0, may not assume the value zero. Therefore, it is bounded away from zero on the unit sphere, say by ϵ_0, where the value of ϵ_0 may depend on j, but will always be positive.

In light of the above observation, shift W_0 so that $\mathrm{supp}W_0 \cap \mathrm{supp}T_0^F = \emptyset$ and $\mathrm{supp}W_0 \cap \mathrm{supp}W_0(1/2\cdot) = \emptyset$. The first condition guarantees that $W_0 \perp T_0^F$, and the second guarantees that "borrowing" the dilates of W_0 will not decrease the magnitudes of projections onto W_0. Similarly choose compactly supported $w_{-1}, \ldots, w_{-k_2} \in V_{n_1}$ with span W_1 so that $\mathrm{supp}W_1 \cap \mathrm{supp}T_1^F = \mathrm{supp}W_1 \cap \mathrm{supp}W_1(1/2\cdot) = \emptyset$ and $\|P_{W_1}\phi\| > \epsilon_1$ for unit $\phi \in T_1^\omega$. Let $n = \max\{n_0, n_1\}$ and $\epsilon = \min\{\epsilon_0, \epsilon_1\}$. Form a new multiresolution analysis $(\hat{V}_p)$ by taking

$$\begin{aligned}\hat{V}_0 = V_l &\oplus \left(\textstyle\bigoplus_{j\in Z} W_0(2^{l-1}\cdot - 2^{l-1}j)\right) \oplus \cdots \oplus \left(\textstyle\bigoplus_{j\in Z} W_0(2^{l-n}\cdot - 2^{l-n}j)\right) \\ &\oplus \left(\textstyle\bigoplus_{j\in Z} W_1(2^{l-1}\cdot - 2^{l-1}j)\right) \oplus \cdots \oplus \left(\textstyle\bigoplus_{j\in Z} W_1(2^{l-n}\cdot - 2^{l-n}j)\right),\end{aligned}$$

where l is chosen sufficiently large so that

1. $\mathrm{supp}(W_0(2^{-1}\cdot) \oplus \cdots \oplus W_0(2^{-n}\cdot) \oplus W_1(2^{-1}\cdot) \oplus \cdots \oplus W_1(2^{-n}\cdot)) \subset [-2^{l-1}, 2^{l-1}]$ and, consequently, $\mathrm{supp}(T_0^F \oplus T_1^F) \subset [-2^{l-1}, 2^{l-1}]$ (we use 2^{l-1} here to assure that W_0 and W_1 do not interfere with each other);
2. T_0^l and T_1^l approximate T_0^∞ and T_1^∞ closely enough so that $\|P_{W_0}\phi\| > (3/4)\epsilon$ for unit $\phi \in T_0^l \ominus T_0^F$ and $\|P_{W_1}\phi\| > (3/4)\epsilon$ for unit $\phi \in T_1^l \ominus T_1^F$, which is possible because the scaling functions form a Riesz basis; and
3. $\|P_{W_0}(I - P_{W_1})\phi\| < \epsilon/2$ for unit $\phi \in \hat{\mathcal{C}}_1^0(\cdot - 1)$ and $\|P_{W_1}(I - P_{W_0})\phi\| < \epsilon/2$ for unit $\phi \in \hat{\mathcal{C}}_0^0$, which is possible because the functions in T_1^ω (respectively, T_0^ω) undergo exponential decay to the left (right), as shown in Lemma 4.1.

Thus we find

$$V_l \subset \hat{V}_0 \subset V_{l+n}.$$

In addition, $\|(\hat{P}^{i+1} - \hat{P}^i)(2^{i/2}\phi(2^i\cdot))\| > \epsilon/4$ for $i \geq 0$ and unit $\phi \in \mathcal{C}_0(\hat{V}_0)$, with an analogous condition holding for $\phi \in \mathcal{C}_1(\hat{V}_0)$. This is true because in the multiresolution analysis $(\hat{V}_p)$, $W_0(2^l\cdot) \oplus W_1(2^l\cdot) \subset \hat{V}_1$, so $(W_0(2^l\cdot))_m \oplus (W_1(2^l\cdot))_m \subset \hat{V}_{m+1}$ for $m \geq 1$. If $u \in \mathcal{C}(\hat{V}_0)$ is a unit vector, then so is $u_{m,0}$, and hence

$$\begin{aligned} \|\hat{P}^{m+1}(I - \hat{P}^m)u_{m,0}\| &\geq \|P_{(W_0(2^l\cdot))_m}(I - \hat{P}^m)u_{m,0}\| \\ &= \|P_{W_0(2^l\cdot)}(I - P_{\hat{V}_0[0,2^m]})u\| \\ &\geq \|P_{W_0(2^l\cdot)}u\| - \|P_{W_0(2^l\cdot)}P_{\hat{V}_0[0,2^m]}u\| \\ &\geq (3/4)\epsilon - (1/2)\epsilon = \epsilon/4, \end{aligned}$$

for all $m \geq 1$. □

Remark 4.1.

1. In the proof of Theorem 2, we took $\mathcal{W}$ to be a k^2-dimensional subspace. In Remark 3.1, it was noted that the rank of P_0P_1 (which is generically k) is a lower bound for dim $\mathcal{W}$. In all of our examples, we achieve the lower bound k.

2. In the case $k = 1$, it is possible to show that the uniform-lower-bound condition of Theorem 2 always holds. In the next section, we show that this is also true for the spline spaces $\mathcal{S}_{d,r}$. Thus Theorem 2 holds in at least the cases of most interest.

We conclude this section with the proof of Lemma 2.1. We restate the lemma for the reader's convenience.

LEMMA 2.1 *Suppose that (V_p) is a multiresolution analysis generated by compactly supported scaling functions. Then there is some n and some set of scaling functions minimally supported on $[-1, 1]$ that generate the multiresolution analysis (V_p') given by*

$$V_p' = V_{p+n}.$$

Proof. Without loss of generality, let V_0 be generated by scaling functions $\Phi = (\phi^1, \ldots, \phi^r)$ as in Remark 2.1. If $f \in H_0(\Phi)$, then f can be expressed uniquely as

$$f = \sum_{s=1}^{k} a_s\phi^s\chi_{[0,1]}. \tag{15}$$

Let $p : H_0(\Phi) \to H_1(\Phi)$ be defined by

$$p(f) = \sum_{s=1}^{k} a_s\phi^s(\cdot - 1)\chi_{[0,1]}. \tag{16}$$

Let $E = H_0(\Phi) \bigcap H_1(\Phi)$ and E' be the maximal p-invariant subspace of E, i.e., $E' = \bigcap_{n=1}^{\infty} p^{-n}(E)$ (where for the purpose of computing E', the range of p is extended to the Minkowski sum of $H_0(\Phi)$ and $H_1(\Phi)$). Since p is nonsingular, $p(E') = E'$. If we consider p as an operator on E', let $\lambda_1, \ldots, \lambda_j$ be the (possibly complex) eigenvalues of p with magnitude less than one and $\lambda_{j+1}, \ldots, \lambda_l$ be its (possibly complex) eigenvalues with magnitude greater than one. The eigenvalues are listed above including multiplicities. We claim that p has no eigenvalues that lie on the unit circle; hence $l = \mathrm{Dim}(E') \leq k$. To see this, suppose that λ is an eigenvalue of p with magnitude one and x an eigenvector associated with this eigenvalue. Let $y = \mathrm{Re}(x)$; then for λ real, set $\kappa = \|y\| = \|p^n(y)\|$; otherwise set $0 < \kappa = \min_{|u|=1}\|x + u\bar{x}\| \leq \|p^n(y)\| \leq \|x\|$. We may define a sequence of functions $\{f_\nu\}_{\nu=0}^{\infty}$ as

$$f_\nu = \sum_{i=0}^{\nu}(-1)^i g_i, \tag{17}$$

where $g_i = (p^i y)(\cdot + i) + (p^{i+1}y)(\cdot + i + 1)$. Note that

1. $g_i = \sum_{s=1}^k a_s^i \phi^s(\cdot + i)$,
2. $A\sum_{s=1}^k (a_s^i)^2 \leq \|g_i\|^2 \leq B\sum_{s=1}^k (a_s^i)^2$,
3. $\|g_i\|^2 = \|p^i y\|^2 + \|p^{i+1}y\|^2$, and
4. $f_\nu = y + (-1)^\nu (p^{\nu+1}y)(\cdot + \nu + 1)$.

Therefore, $\|f_\nu\|^2 = \|y\|^2 + \|p^{\nu+1}y\|^2 \leq 2\|x\|^2$ and $\sum_{i=1}^\nu \sum_{s=1}^k (a_s^i)^2 \geq \sum_{i=1}^\nu \frac{2\kappa^2}{B} = 2\kappa^2\nu/B$, which shows that $\tau(\Phi)$ is not a Riesz basis and thus yields a contradiction.

Let $x_1, \ldots, x_l$ be the generalized eigenvectors associated with $\lambda_1, \ldots, \lambda_l$. In the event that an eigenvalue is complex, its conjugate is also an eigenvalue and we will replace the appropriate generalized eigenvectors by their real and imaginary parts. Note that $x_1, \ldots, x_l$ form a basis for E'. For $y = x_m$, $m = 1, \ldots, j$, we define a sequence of functions $\{f_\nu\}$ as above. In this case, because of the choice of y, $\|p^n y\| \to 0$ since $\lambda_1, \ldots, \lambda_j$ are in magnitude less than one. Therefore, $f_\nu \overset{L^2}{\to} y$. For $y = x_m$, $m = j+1, \ldots, l$, choose the sequence $\{f_\nu\}$ as $f_\nu = \sum_{i=0}^\nu (-1)^i g_i$, where $g_i = (p^{-i}y)(\cdot - i) + (p^{-(i+1)}y)(\cdot - i - 1)$. We again have $f_\nu \overset{L^2}{\to} y$ and so $x_1, \ldots, x_l$ are in V_0. By replacing $\phi^1, \ldots, \phi^k$ by appropriate linear combinations of $\phi^1, \ldots, \phi^k$, we may assume that the right halves (and hence also the shifted left halves) of $\phi^1, \ldots, \phi^l$ are in E'. Since $x_1, \ldots, x_l$ span E', we may replace the generators $\phi^1, \ldots, \phi^l$ by the equivalent set of generators $x_1, \ldots, x_l$. Since $x_1, \ldots, x_l$ have support in $[0,1]$, these may be removed to A. By the linear independence of $\tau(\Phi)$, none of the $\phi^{l+1}, \ldots, \phi^k$ can have their right halves or shifted left halves in E'. If $E \ominus E' = 0$, then $\Phi = \{x_1, \ldots, x_l, \phi^{l+1}, \ldots, \phi^r\}$ is minimally supported and spans V_0.

If there are nonzero functions in $E \ominus E'$, let $E'' = H_0(\Phi) \ominus E'$. For nonzero $x \in E''$, let n_x be such that $p^{n_x}(x) \notin E$ but $p^i(x) \in E$, $1 \leq i < n_x$, and define $n_0 = \infty$. Note that n_x is finite for nonzero x since $x \notin E'$ and that $n_{\alpha x + \beta y} \geq \min\{n_x, n_y\}$ for scalars α, β, and $x, y \in E''$. Now for $i = 1, 2, \ldots$, set $E_i = \{x \in E'' : n_x \geq i\}$. Then E_i is a subspace of E'' and $E_1 = E''$. Also, there is some $\tilde{n}$ such that $E_{\tilde{n}+1} = 0$ and $E_{\tilde{n}} \neq 0$. Let $X_{\tilde{n}}$ be a set of vectors that form a basis for $E_{\tilde{n}}$. By the definition of $\tilde{n}$, if $x \in X_{\tilde{n}}$, then $x \notin E$. Furthermore, $p(X_{\tilde{n}})$ gives a linearly independent set of vectors in $E_{\tilde{n}-1} \bigcap E$. Beginning with the vectors $p(X_{\tilde{n}})$ and $X_{\tilde{n}}$, complete this set to a basis, $X_{\tilde{n}-1}$, for $E_{\tilde{n}-1}$. Now $p(X_{\tilde{n}-1})$ is a linearly independent set of vectors in $E_{\tilde{n}-2}$ and we add a sufficient number of independent vectors to form a basis, $X_{\tilde{n}-2}$, for $E_{\tilde{n}-2}$. Continue this processes until a basis, X_1, for E_1 has been formed. Since X_1 forms a basis for E'', $\{x + p(x)(\cdot + 1) : x \in X_1\}$ may be used to replace $\phi^{l+1}, \ldots, \phi^k \in \Phi$. For $x \in X_1$, set $\phi_x = x + p(x)(\cdot + 1)$. We now consider $\tilde{\phi}_x = \phi_x(\cdot + i - 1)$ for $x \in X_i - X_{i+1}$ (where we set $X_{\tilde{n}+1} = 0$). Since these are just translates of ϕ_x, $x \in X_1$, we may replace ϕ_x by $\tilde{\phi}_x$. For each $x \in X_1$, let

$$\hat{\phi}_x = \sum_{i=0}^{n_x - 1} (-1)^i \tilde{\phi}_{p^i(x)},$$

which is a linear combination of elements from $\{\tilde{\phi}_x : x \in X_1\}$. Then $\{\hat{\phi}_x : x \in X_1\}$ is linearly independent and may be used to replace $\{\tilde{\phi}_x : x \in X_1\}$. Note that the support of $\tilde{\phi}_x$ is $[-n_x, -n_x + 1] \cup [0,1]$ and that $\operatorname{span}\{\hat{\phi}_x(\cdot - a_x)\chi_{[0,1]} x \in X_1\}$ and $\operatorname{span}\{\hat{\phi}(\cdot - b_x)\chi_{[0,1]} : x \in X_1\}$ are linearly independent spaces. The set $\{\phi_x, x \in X_1\} \cup \{x_i\}_{i=1}^l \cup \{\phi^i\}_{i=k+1}^r$ spans V_0. Consequently, there is an integer h, $\log_2 \tilde{n} \leq h < \log_2 \tilde{n} + 1$, such that the above functions, scaled by 2^h, are all supported in $[-1,1]$, and there is a $\tilde{\Phi}$ formed from appropriate translates of these functions such that $\tau(\tilde{\Phi})$ forms a basis for V_h. From the above construction and by Remark 2.1 in §2, appropriate linear combinations of the generators of $\tilde{\Phi}$ may be taken so as to produce a set of generators that is minimally supported. It follows from Theorems 4.3 and 5.3 of [14] that $\tau(\tilde{\Phi})$ is a Riesz basis of V_h. □

5. Orthogonal spline scaling functions. For $h > 0$ and $1 \le r \le d+1$, let $\mathcal{S}_{d,r}(h)$ denote the space of piecewise-polynomial functions on $\mathbf{R}$ of degree d with knots of multiplicity r on $h\mathbf{Z}$. Let $\lceil a \rceil$ denote the least integer greater than or equal to a.

We let (t_i) denote the sequence of knots with multiplicity r for $\mathcal{S}_{d,r}(h)$ given by

$$t_i = h\lceil i/r \rceil, \quad i \in \mathbf{Z}.$$

Also, let N_i^d denote the B-spline in $\mathcal{S}_{d,r}(h)$ with knots $t_i, \ldots, t_{i+d+1}$. Thus N_i^d is supported on $[t_i, t_{i+d+1}] \subset [jh, (j+L)h]$, where $j = \lceil i/r \rceil$ and $L = \lceil (d+1)/r \rceil$. Observe that the support of N_0^d is exactly $[0, Lh]$.

Let q be such that $2^q + 1 \ge L$ and choose $h = 2^{-q}$. Let $R = 2^q r$ and define $\phi^s = N_{s-d-1}^d$ for $s = 1, \ldots, R$. Then $\Phi = (\phi^1, \ldots, \phi^R)^\top$ generates a multiresolution analysis (V_p) with $V_0 = \mathcal{S}_{d,r}(2^{-q}) \cap L^2(\mathbf{R})$ (see Goodman and Lee [11]). It follows from the choice of q that Φ is minimally supported in $[-1, 1]$. Observe that there are exactly $d + 1 - r$ values of i such that 0 is an interior point of $[t_i, t_{i+d+1}]$. Hence $k = k(\Phi) = d + 1 - r$.

In this section, we will show that (V_p) satisfies the hypotheses of Theorem 2 and so there is an orthogonal $\tilde{V}_0$ between $V_{q'}$ and $V_{q'+1}$ for some integer q'. Towards this end, we first develop some results concerning the sign changes of spline functions. We say that $(x_1, \ldots, x_\alpha)$ is a *sign-change sequence* on (a, b) for $f : \mathbf{R} \to \mathbf{R}$ if f is not identically zero on any subinterval (x_j, x_{j+1}), $j = 0, \ldots, \alpha$, where $a = x_0 < x_1 < \cdots < x_\alpha < x_{\alpha+1} = b$, and, for either $k = 0$ or $k = 1$, we have $(-1)^{j+k} f(x) \ge 0$ for all $x \in (x_j, x_{j+1})$ and $j = 0, \ldots, \alpha$. Let i_0, i_1 be given integers and define $\mathcal{P}_{i_0,i_1} = \operatorname{span}\{N_i^d \mid i_0 \le i \le i_1\}$ and $\sigma_i := \{x \mid N_i^d(x) \ne 0\} = (t_i, t_{i+d+1})$ for $i = i_0, \ldots, i_1$. We will need the following result concerning sign changes of B-spline expansions.

LEMMA 5.1. *Let $a, b \in \mathbf{R}$ and i_0, i_1 be integers such that $t_{i_0} < t_{i_1}$ and suppose that $r \le d$ (and thus the B-splines are continuous).*

1. *If $f \in \mathcal{P}_{i_0,i_1}$ and $(x_1, \ldots, x_\alpha)$ is a sign-change sequence for f on (a, b), then for any open interval $G \subset (a, b)$, we have*

$$\operatorname{card}\{x_j \in G\} \le \operatorname{card}\{i \mid \sigma_i \cap G \ne \emptyset\} - 1. \tag{18}$$

2. *If $a < x_1 < x_2 < \cdots < x_\alpha < b$ satisfies*

$$x_i \in \sigma_{i_0+i-1} \cap \sigma_{i_0+i} \quad \textit{for } i = 1, \ldots, \alpha,$$

then there is some $f \in \mathcal{P}_{i_0,i_0+\alpha}$ for which $(x_1, \ldots, x_\alpha)$ is a sign-change sequence on (a, b). Furthermore, f does not vanish on any subinterval of $\bigcup_{i=i_0}^{i_0+\alpha} \sigma_i$.

Proof. Part 1 follows from the variation-diminishing property of B-spline expansions (cf. [21, Thm. 4.76]). Let $x_0 \in \sigma_{i_0}, x_0 \ne x_i$. By [21, Thm. 4.61], there is a unique $f \in \mathcal{P}_{i_0,i_1}$ such that

$$f(x_0) = 1, \qquad f(x_i) = 0, \quad i = 1, \ldots, \alpha.$$

Suppose $\alpha = 1$. Then $f = c_0 N_{i_0}^d + c_1 N_{i_0+1}^d$. Since $f(x_0) = 1$, at least one of c_0 and c_1 are nonzero. Since $x_1 \ne t_k$ for all k, it follows that both c_0 and c_1 are nonzero and so f does not vanish on any subinterval of $\sigma_{i_0} \cup \sigma_{i_1}$. Hence by counting the zeros of f including their multiplicities, it follows from [21, Thm. 4.56] that f changes sign at x_1. It then follows by induction on α that f does not vanish on any subinterval of $(t_{i_0}, t_{i_0+\alpha+d+1}) = \bigcup_{i=i_0}^{i_0+\alpha} \sigma_i$. Hence [21, Thm. 4.56] shows that $(x_1, \ldots, x_\alpha)$ must be all of the zeros of f in $(t_{i_0}, t_{i_0+\alpha+d+1})$ and that all of these zeros must be isolated, simple zeros. Thus $(x_1, \ldots, x_\alpha)$ is a sign-change sequence of f on (a, b). □

LEMMA 5.2. *Let $f \in V_0$ be such that $f\chi_{[0,\infty)} \neq 0$. Then there is some $g \in V_1$ such that supp $g \subset [0,\infty)$ and $\langle f, g\rangle \neq 0$.*

Proof. If $r = d + 1$, then $f\chi_{[0,\infty)} \in V_1$ and we can take $g = f\chi_{[0,\infty)}$.

Suppose $r \leq d$. Let $s_i = t_i/2$ for all i. Note that $n_i^d := N_i^d(2\cdot)$ is the B-spline associated with the knots $s_i, \ldots, s_{i+d+1}$. Let $\sigma'_i = (s_i, s_{i+d+1}) = \{x \mid n_i^d(x) > 0\}$ and $\mathcal{P}'_{i_0,i_1}$ be the span of the B-splines associated with the knots $s_{i_0}, \ldots, s_{i_1+d+1}$.

Let $\hat{f} = f\chi_{[0,\infty)}$. If there exists a σ'_i such that $\sigma'_i \subset \operatorname{supp}\hat{f}$ and f does not change sign on σ'_i, then we can choose $g = n_i^d$. If $\operatorname{supp}\hat{f}$ is not of the form $[0, b]$ or $[0, \infty)$, then some component I of $\operatorname{supp}\hat{f}$ does not contain zero and $\hat{f}\chi_I \in V_1$. In this case, choose $g = \hat{f}\chi_I$. If $\operatorname{supp}\hat{f} = [0, b]$, then by the first part of Lemma 5.1, on $[b - h, b]$, $\hat{f}$ is a polynomial with at most $r - 1$ zeros. Thus on $[b - h, b]$, $\hat{f} = (x - b)^{d-r+1}p(x)$, where $p(x)$ is a polynomial of degree $r - 1$. Consequently, we can choose a function $g \in V_1$ such that on $[b - h, b]$, we have $g = (x - (b - h))^{d-r+1}p(x)$.

Finally, suppose $\operatorname{supp}\hat{f} = [0, \infty)$ and each σ'_i contains a sign change. Suppose there are integers i_0, j_0 and $\alpha \geq 0$ such that

$$\begin{gathered} x_{j_0} \notin \sigma'_{i_0}, \qquad x_{j_0+\alpha+1} \notin \sigma'_{i_0+\alpha}, \quad \text{and} \\ x_{j_0+i} \in \sigma'_{i_0+i-1} \cap \sigma'_{i_0+i} \quad \text{for } i = 1, \ldots, \alpha. \end{gathered} \tag{19}$$

Then part 2 of Lemma 5.1 implies that there is some $g \in V_1$ with support $(a, b) := \bigcup_{i=i_0}^{i_0+\alpha} \sigma'_i$ such that $(x_{j_0+1}, \ldots, x_{j_0+\alpha})$ is a sign-change sequence for g on (a, b). Furthermore, since g does not vanish on any subinterval of (a, b), we have $\langle f, g\rangle \neq 0$.

We now show that there must exist an i_0 and j_0 so that (19) holds. We first show that there do not exist integers i_0 and j_0 such that

$$x_{i_0+l} \in \sigma'_{j_0+l-1} \cap \sigma'_{j_0+l} \tag{20}$$

for all $l \geq 1$. Suppose not; then

$$\operatorname{card}\left\{ x_j \in \bigcup_{l=1}^{2m} \sigma'_{j_0+l} \right\} \geq 2m. \tag{21}$$

Since f is in V_0, it follows from part 1 of Lemma 5.1 that there exists a constant c such that

$$m + c > \operatorname{card}\left\{ x_j \in \bigcup_{l=1}^{2m} \sigma'_{j_0+l} \right\} \tag{22}$$

for all m. This contradicts (21) for m sufficiently large. Thus the above shows that there does not exist an infinite sign-change sequence that satisfies (20).

We now rule out the case where (19) fails for each finite α. Let $0 = x_0 < x_1 < \cdots$ be an infinite sign-change sequence of f on $(0, \infty)$, $n_0 = 1 - r$, and for each $i \geq n_0$, set $k(i) = \min\{j : x_j \in \sigma'_i\} - 1$. Assume that (19) does not hold. Since we have already ruled out (20), it must be that for each i there exists an $n \geq 1$ which depends upon i such that $x_{k(i)+l} \in \sigma'_{i+l-1} \cap \sigma'_{i+l}$, $l = 1, \ldots, n-1$, but $x_{k(i)+n} \notin \sigma'_{i+n-1} \cap \sigma'_{i+n}$ and $x_{k(i)+n} \in \sigma'_{i+n-1}$. Thus $x_{k(i)+n} < \inf \sigma'_{i+n}$, and so from the definition of $k(i+n)$, we find that $k(i+n) > (k(i)+n) - 1$ or $k(i+n) \geq k(i) + n$. Thus for each infinite sign-change sequence not satisfying (19), there exists an infinite sequence $n_0 < n_1 < \cdots$ such that $k(n_{i+1}) \geq k(n_i) + n_{i+1} - n_i$. Since $k(n_0) = 0$, we find that $k(n_l) \geq n_l - n_0$ for all l. From (21) and (22) above, we find that for l large enough, this leads to a contradiction. □

Thus Q_∞, the orthogonal projection onto $V_1[0,\infty)$, is nonsingular on T_0^∞. Then Lemma 4.4 implies that (V_p) satisfies the hypotheses of Theorem 2.

THEOREM 3. *Let r, d, and q be as above and let $V_0 = \mathcal{S}_{d,r}(2^{-q}) \cap L^2(\mathbf{R})$. Then there is some integer $\tilde{q}$ and some orthogonal multiresolution analysis $(\tilde{V}_p)$ such that*

$$V_{\tilde{q}} \subset \tilde{V}_0 \subset V_{\tilde{q}+1}.$$

5.1. Example: $\mathcal{S}_{3,2}$—C^1 cubic scaling functions.

In this case, $L = 2$, $q = 0$, and V_0 is generated by the two B-splines associated with the knot sequences $(-1,-1,0,0,1)$ and $(-1,0,0,1,1)$. As noted in Remark 3.1, we will need $\mathcal{W}$ to have dimension at least 2. Now $\mathcal{A}(V_1) \ominus \mathcal{A}(V_0)$ is two dimensional, which is not large enough to allow for any freedom in the choice of $\mathcal{W}$. Observe that $\mathcal{A}(V_2) \ominus \mathcal{A}(V_1)$ is $6 - 2 = 4$ dimensional. Using Mathematica to aid in the calculations, we find a two-dimensional $\mathcal{W}$ that satisfies equation (1). This gives a total of six scaling functions and six wavelets, as shown in Figure 3.

Appendix A. Interpolation values for scaling functions and wavelets.

TABLE 1
Interpolation values for piecewise-linear scaling functions (at the quarter-integers) and wavelets (at the eighth-integers) of the example in §3.1.

$\tilde{\phi}^1$	$\tilde{\phi}^2$	$\tilde{\phi}^3$	$\tilde{\psi}^1$	$\tilde{\psi}^2$	$\tilde{\psi}^3$
0	0	0	0	0	0
			$165 - 44\sqrt{5}$	$-3 - 2\sqrt{5}$	$6 + 4\sqrt{5}$
$3 + 2\sqrt{5}$	1	131	$242 - 352\sqrt{5}$	$-4\sqrt{5} - 6$	$12 + 8\sqrt{5}$
			$-1265 + 836\sqrt{5}$	$10\sqrt{5} + 15$	$-30 - 20\sqrt{5}$
$-18 - 12\sqrt{5}$	2	$150 - 96\sqrt{5}$	1716	$24\sqrt{5} + 36$	$-72 - 48\sqrt{5}$
			$-1265 - 836\sqrt{5}$	$25 + 2\sqrt{5}$	$-38 + 4\sqrt{5}$
$5 + 18\sqrt{5}$	1	$-381 + 160\sqrt{5}$	$242 + 352\sqrt{5}$	$-84\sqrt{5} - 82$	$92 + 120\sqrt{5}$
			$165 + 44\sqrt{5}$	$54\sqrt{5} - 117$	$254 - 36\sqrt{5}$
132	0	0	0	264	0
				$-117 - 54\sqrt{5}$	$-369 - 125\sqrt{5}$
$5 - 18\sqrt{5}$				$-82 + 84\sqrt{5}$	$78 + 118\sqrt{5}$
				$25 - 2\sqrt{5}$	$53 + 17\sqrt{5}$
$-18 + 12\sqrt{5}$				$36 - 24\sqrt{5}$	$12 - 36\sqrt{5}$
				$-10\sqrt{5} + 15$	$5 - 15\sqrt{5}$
$3 - 2\sqrt{5}$				$4\sqrt{5} - 6$	$-2 + 6\sqrt{5}$
				$2\sqrt{5} - 3$	$-1 + 3\sqrt{5}$
0				0	0
Norms					
$4\sqrt{231}$	$\frac{2\sqrt{3}}{3}$	$32\sqrt{33} - 10\sqrt{165}$	$44\sqrt{462}$	$8\sqrt{231}$	$28\sqrt{42} + 4\sqrt{210}$

TABLE 2

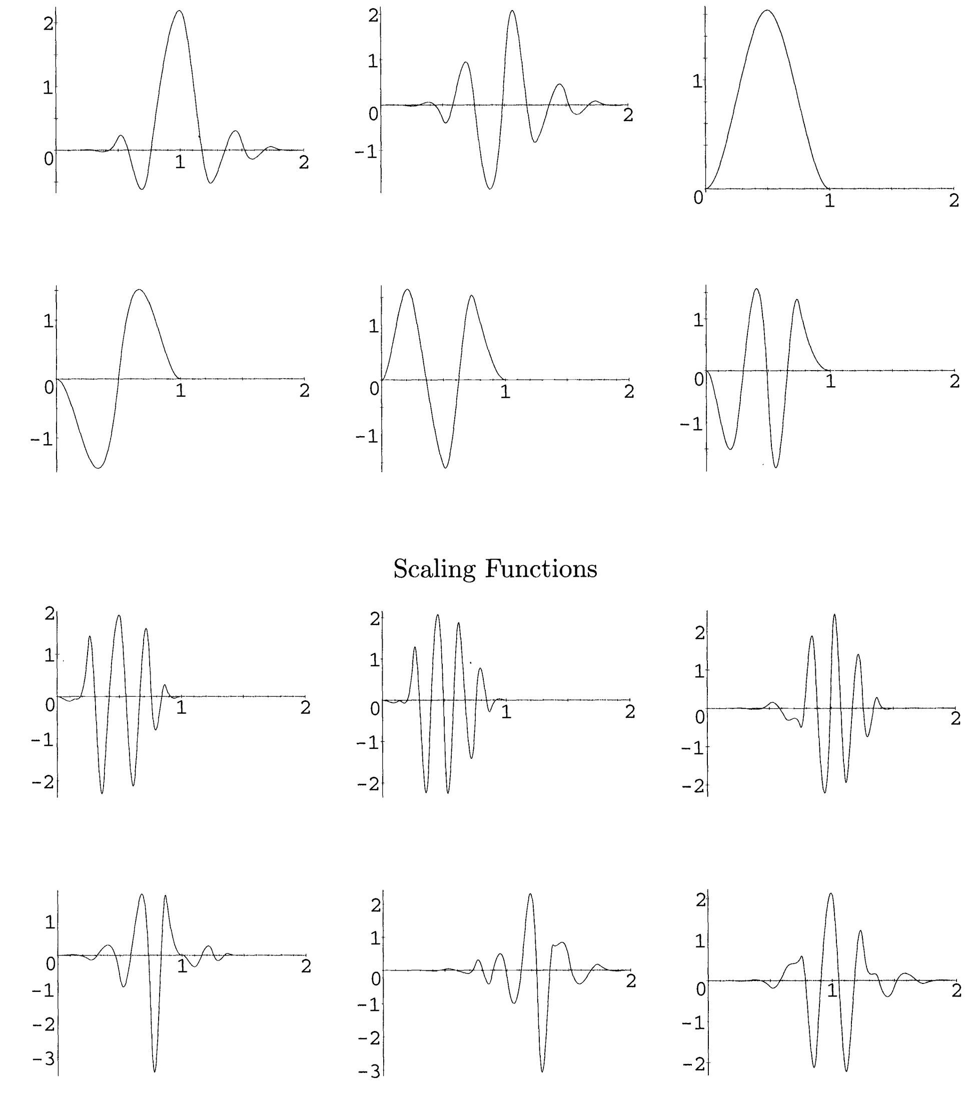
Interpolation values for piecewise-linear symmetric scaling functions (at the eighth-integers) and wavelets (at the sixteenth-integers) of the example in §3.2. *For* $\tilde\phi^1$, $\tilde\psi^3$, *and* $\tilde\psi^4$, *only half of the points are given. The others points may be computed using the fact that* $\tilde\phi^1$ *and* $\tilde\psi^3$ *are symmetric while* $\tilde\psi^4$ *is antisymmetric.*

$\tilde\phi^1$	$\tilde\phi^2$	$\tilde\phi^3$	$\tilde\phi^4$	$\tilde\psi^1$	$\tilde\psi^2$	$\tilde\psi^3$	$\tilde\psi^4$
0	0	0	0	0	0	0	0
				$-93\sqrt{14}+294$	$98-31\sqrt{14}$	$-1079+288\sqrt{14}$	$127\sqrt{14}-476$
$1079-288\sqrt{14}$	1	5	$175-20\sqrt{14}$	$-58\sqrt{14}+84$	$-308+66\sqrt{14}$	$-2158+576\sqrt{14}$	$-952+254\sqrt{14}$
				$21\sqrt{14}+182$	$266-37\sqrt{14}$	$1307-304\sqrt{14}$	$588-131\sqrt{14}$
$-2386+592\sqrt{14}$	2	10	$98+24\sqrt{14}$	$228\sqrt{14}-224$	$336-12\sqrt{14}$	$4772-1184\sqrt{14}$	$2128-516\sqrt{14}$
				$101\sqrt{14}-518$	$-266+37\sqrt{14}$	$-280\sqrt{14}+965$	$-125\sqrt{14}+420$
$1421-312\sqrt{14}$	3	-1	$175-20\sqrt{14}$	$-154\sqrt{14}-308$	$-364-42\sqrt{14}$	$-2842+624\sqrt{14}$	$-1288+266\sqrt{14}$
				$-13\sqrt{14}-406$	$-98+31\sqrt{14}$	$-2105+360\sqrt{14}$	$-980+145\sqrt{14}$
$684-48\sqrt{14}$	4	-12	0	0	$672-24\sqrt{14}$	$-1368+96\sqrt{14}$	$-672+24\sqrt{14}$
				$13\sqrt{14}+406$	$-98+31\sqrt{14}$	$5737-1464\sqrt{14}$	$-387\sqrt{14}+1596$
$-2105+360\sqrt{14}$	3	-1	$-175+20\sqrt{14}$	$154\sqrt{14}+308$	$-364-42\sqrt{14}$	$-5334+1648\sqrt{14}$	$226\sqrt{14}-168$
				$-101\sqrt{14}+518$	$-266+37\sqrt{14}$	$8139-1208\sqrt{14}$	$-201\sqrt{14}+2548$
$-350-400\sqrt{14}$	2	10	$-98-24\sqrt{14}$	$-228\sqrt{14}+224$	$336-12\sqrt{14}$	$3436+608\sqrt{14}$	$396\sqrt{14}+1232$
				$-21\sqrt{14}-182$	$266-37\sqrt{14}$	$-10411+1792\sqrt{14}$	$433\sqrt{14}-2884$
$2341+48\sqrt{14}$	1	5	$-175+20\sqrt{14}$	$58\sqrt{14}-84$	$-308+66\sqrt{14}$	$-6082-1696\sqrt{14}$	$-554\sqrt{14}-2968$
				$93\sqrt{14}-294$	$98-31\sqrt{14}$	$-2553+816\sqrt{14}$	$235\sqrt{14}-3500$
$9576-672\sqrt{14}$	0	0	0	0	0	$19152-1344\sqrt{14}$	0
Norms							
$1368\sqrt{7}-672\sqrt{2}$	$\frac{4\sqrt{3}}{3}$	$\frac{4\sqrt{21}}{3}$	$28\sqrt{21}-7\sqrt{6}$	$168\sqrt{7}-42\sqrt{2}$	$56\sqrt{21}-14\sqrt{6}$	$2736\sqrt{7}-1344\sqrt{2}$	$672\sqrt{14}-336$

TABLE 3

Interpolation values for piecewise-cubic C^1 scaling functions of the example in §5.1 *and their derivatives at the quarter-integers.*

ϕ^1	0	0.01145260744	0.1941789118	−0.2276642461	2.188816056	−0.5189133117	0.1301883312	0.04941362556	0
$\phi^{1\prime}$	0	0.2283230026	4.395387791	14.37750265	0	1.209857886	−7.265110200	−1.032301156	0
ϕ^2	0	−0.02531744710	−0.2967038040	0.1905122953	0.6577284145	−0.7845469243	0.1871122709	0.07121519464	0
$\phi^{2\prime}$	0	−0.5135739653	−7.525014752	−24.97260290	45.02381581	2.403631242	−10.46198835	−1.487792616	0
ϕ^3	0	0.8204126541	1.640825308	0.8204126541	0	0	0	0	0
$\phi^{3\prime}$	0	4.922475925	0	−4.922475925	0	0	0	0	0
ϕ^4	0	−1.280868846	0	1.280868846	0	0	0	0	0
$\phi^{4\prime}$	0	−5.123475383	20.49390153	−5.123475383	0	0	0	0	0
ϕ^5	0	1.456996003	−1.563883833	1.456996003	0	0	0	0	0
$\phi^{5\prime}$	0	−10.35032509	−5.872827930	−8.442724283	0	0	0	0	0
ϕ^6	0	−1.025146541	−0.5127445138	1.276503733	0	0	0	0	0
$\phi^{6\prime}$	0	18.39424679	−50.39758685	−12.10638475	0	0	0	0	0



Scaling Functions

Wavelets

FIG. 3. *Piecewise-cubic C^1 orthogonal scaling functions (from right to left and top to bottom: $\tilde{\phi}^1(\cdot - 1)$, $\tilde{\phi}^2(\cdot - 1)$, $\tilde{\phi}^3$, $\tilde{\phi}^4$, $\tilde{\phi}^5$, $\tilde{\phi}^6$) and wavelets (from top to bottom and right to left: $\tilde{\psi}^1$, $\tilde{\psi}^2$, $\tilde{\psi}^3(\cdot - 1)$, $\tilde{\psi}^4(\cdot - 1)$, $\tilde{\psi}^5(\cdot - 1)$, $\tilde{\psi}^6(\cdot - 1)$).*

Appendix B. Matrix-dilation coefficients. Suppose that (V_p) is an orthogonal multiresolution analysis generated by scaling functions $\phi^1, \ldots, \phi^r$ minimally supported on $[-1, 1]$. Let $\Phi = (\phi^1, \ldots, \phi^r)^{\mathsf{T}}$. Since $V_0 \subset V_1$, it follows that Φ must satisfy a two-scale dilation equation with matrix coefficients:

$$\Phi(x) = \sqrt{2} \sum_{n=-2}^{1} H_n \Phi(2x - n), \tag{23}$$

where each H_n is an $r \times r$ real matrix. The support and orthogonality of the scaling functions imply the given summation limits. In Tables 4–6, we give H_{-2}, H_{-1}, H_0,

TABLE 4

Matrix-dilation coefficients for the piecewise-linear scaling functions and wavelets of the example in §3.1.

H_{-2}	0	$-\sqrt{154}(3+2\sqrt{5})/3696$	$\sqrt{14}(2+5\sqrt{5})/1232$
	0	0	0
	0	0	0
H_{-1}	$-\sqrt{2}(3+2\sqrt{5})/44$	$\sqrt{154}(67+30\sqrt{5})/3696$	$\sqrt{14}(-10+\sqrt{5})/112$
	0	0	0
	0	0	0
H_0	$\sqrt{2}/2$	$\sqrt{154}(67-30\sqrt{5})/3696$	$\sqrt{14}(10+\sqrt{5})/112$
	0	$3\sqrt{2}/8$	$\sqrt{22}(-4+\sqrt{5})/88$
	0	$\sqrt{22}(32+7\sqrt{5})/264$	$\sqrt{2}(-5+4\sqrt{5})/88$
H_1	$\sqrt{2}(-3+2\sqrt{5})/44$	$\sqrt{154}(-3+2\sqrt{5})/3696$	$\sqrt{14}(-2+5\sqrt{5})/1232$
	$\sqrt{154}/22$	$3\sqrt{2}/8$	$\sqrt{22}(4+\sqrt{5})/88$
	$-\sqrt{70}/22$	$\sqrt{22}(-32+7\sqrt{5})/264$	$-\sqrt{2}(5+4\sqrt{5})/88$

K_{-2}	0	0	0
	0	$(3+2\sqrt{5})\sqrt{154})/3696$	$-(2+5\sqrt{5})\sqrt{14}/1232$
	0	$-\sqrt{7}(1+\sqrt{5})/336$	$\sqrt{77}(-1+3\sqrt{5})/1232$
K_{-1}	0	0	0
	$(2\sqrt{5}+3)\sqrt{2}/44$	$-\sqrt{154}(67+30\sqrt{5})/3696$	$(10-\sqrt{5})\sqrt{14}/112$
	$-\sqrt{11}(1+\sqrt{5})/44$	$\sqrt{7}(29+13\sqrt{5})/336$	$\sqrt{77}(-75+17\sqrt{5})/1232$
K_0	0	$\sqrt{77}(-2+\sqrt{5})/264$	$\sqrt{7}(13-6\sqrt{5})/88$
	$\sqrt{2}/2$	$\sqrt{154}(-67+30\sqrt{5})/3696$	$-(10+\sqrt{5})\sqrt{14}/112$
	0	$\sqrt{7}(-29+13\sqrt{5})/336$	$-\sqrt{77}(75+17\sqrt{5})/1232$
K_1	$13/22$	$-\sqrt{77}(\sqrt{5}+2)/264$	$-\sqrt{7}(13+6\sqrt{5})/88$
	$(3-2\sqrt{5})\sqrt{2}/44$	$\sqrt{154}(3-2\sqrt{5})/3696$	$(2-5\sqrt{5})\sqrt{14}/1232$
	$\sqrt{11}(1-\sqrt{5})/44$	$\sqrt{7}(1-\sqrt{5})/336$	$-\sqrt{77}(3\sqrt{5}+1)/1232$

and H_1 for the three examples given in §§3.1, 3.2, and 5.1.

The wavelets $\Psi = (\psi^1, \ldots, \psi^r)^T$ constructed from the scaling functions also satisfy a two-scale dilation equation of the form

$$\Psi(x) = \sqrt{2} \sum_{n=-2}^{1} K_n \Phi(2x - n). \tag{24}$$

These matrices are also given in Tables 4–6.

Acknowledgments. The second and third authors thank Gil Strang for his helpful suggestions and encouragement and Amos Ron for pointing out the simplifications in the multiresolution conditions when the scaling functions are compactly supported. We especially thank one of the anonymous referees for a careful reading and for numerous corrections.

TABLE 5

Matrix-dilation coefficients for the symmetric piecewise-linear scaling functions and wavelets of the example in §3.2.

H_{-2}	0	$-\sqrt{6}(-2\sqrt{2}+\sqrt{7})/144$	$-\sqrt{6}(14\sqrt{14}-65)/1008$	$-\sqrt{6}/84$
	0	0	0	0
	0	0	0	0
	0	0	0	0
H_{-1}	$\sqrt{2}/28$	$\sqrt{6}\left(-2\sqrt{2}+\sqrt{7}\right)/144$	$\sqrt{6}\left(14\sqrt{14}+79\right)/1008$	$-13\sqrt{6}/84$
	0	0	0	0
	0	0	0	0
	0	0	0	0
H_0	$\sqrt{2}/2$	$\sqrt{6}\left(-2\sqrt{2}+\sqrt{7}\right)/144$	$\sqrt{6}\left(14\sqrt{14}+79\right)/1008$	$13\sqrt{6}/84$
	0	$3\sqrt{2}/8$	$3\sqrt{14}/56$	$-\sqrt{14}/14$
	0	$\sqrt{14}/8$	$-9\sqrt{2}/56$	$3\sqrt{2}/14$
	0	$\sqrt{2}\left(4\sqrt{7}+\sqrt{2}\right)/24$	$-\sqrt{2}\left(-4+\sqrt{2}\sqrt{7}\right)/24$	0
H_1	$\sqrt{2}/28$	$\sqrt{3}\left(4-\sqrt{14}\right)/144$	$\sqrt{3}\left(65\sqrt{2}-28\sqrt{7}\right)/1008$	$\sqrt{6}/84$
	$\sqrt{42}/14$	$3\sqrt{2}/8$	$3\sqrt{14}/56$	$\sqrt{14}/14$
	$-3\sqrt{6}/14$	$\sqrt{14}/8$	$-9\sqrt{2}/56$	$-3\sqrt{2}/14$
	0	$-1/12-\sqrt{14}/6$	$-(2\sqrt{2}+\sqrt{7})/12$	0

K_{-2}	0	0	0	0
	0	0	0	0
	0	$\sqrt{6}\left(-2\sqrt{2}+\sqrt{7}\right)/144$	$\sqrt{6}\left(14\sqrt{14}-65\right)/1008$	$\sqrt{6}/84$
	0	$\sqrt{6}\left(-4+\sqrt{14}\right)/144$	$\sqrt{6}\left(28\sqrt{7}-65\sqrt{2}\right)/1008$	$\sqrt{3}/42$
K_{-1}	0	0	0	0
	0	0	0	0
	$-\sqrt{2}/28$	$-\sqrt{6}\left(-2\sqrt{2}+\sqrt{7}\right)/144$	$-\sqrt{6}\left(14\sqrt{14}+79\right)/1008$	$13\sqrt{6}/84$
	$-1/14$	$-\sqrt{6}\left(-4+\sqrt{14}\right)/144$	$-\sqrt{6}\left(28\sqrt{7}+79\sqrt{2}\right)/1008$	$13\sqrt{3}/42$
K_0	0	$-\sqrt{6}\left(-2\sqrt{2}+\sqrt{7}\right)/36$	$-\sqrt{6}\left(2\sqrt{14}+1\right)/36$	$\sqrt{6}/6$
	0	0	$-2\sqrt{2}/7$	$3\sqrt{6}/14$
	$\sqrt{2}/2$	$-\sqrt{6}\left(-2\sqrt{2}+\sqrt{7}\right)/144$	$-\sqrt{6}\left(14\sqrt{14}+79\right)/1008$	$-13\sqrt{6}/84$
	0	$\sqrt{6}\left(-4+\sqrt{14}\right)/144$	$\sqrt{6}\left(28\sqrt{7}+79\sqrt{2}\right)/1008$	$13\sqrt{3}/42$
K_1	0	$\sqrt{6}\left(-2\sqrt{2}+\sqrt{7}\right)/36$	$\sqrt{6}\left(2\sqrt{14}+1\right)/36$	$\sqrt{6}/6$
	$2\sqrt{6}/7$	0	$-2\sqrt{2}/7$	$-3\sqrt{2}/14$
	$-\sqrt{2}/28$	$\sqrt{6}\left(-2\sqrt{2}+\sqrt{7}\right)/144$	$\sqrt{6}\left(14\sqrt{14}-65\right)/1008$	$-\sqrt{6}/84$
	$1/14$	$-\sqrt{6}\left(-4+\sqrt{14}\right)/144$	$-\sqrt{6}\left(28\sqrt{7}-65\sqrt{2}\right)/1008$	$\sqrt{3}/42$

TABLE 6

Matrix-dilation coefficients for the C^1 piecewise-cubic scaling functions and wavelets of the example in §5.1.

H_{-2}	0	0	-0.0002208624742	-0.001924773722	-0.003288244991	-0.001947295571
	0	0	0.0005480992544	0.004856715145	0.008496591936	0.005147569977
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
H_{-1}	0.03702321115	0.02327794583	0.070258082635	0.3134392642	0.1926557751	0.06689140585
	-0.05789731181	-0.04178352399	-0.09405362704	-0.4199017293	-0.2020095784	-0.03215729927
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
H_0	0.5	0	-0.01275849911	-0.1152392333	0.254509594	-0.144924649
	0.07876397049	0.25	-0.01928041017	-0.1741474187	0.3578324616	-0.1761829744
	0	0	0.3365384615	0.1901521826	0.1136121823	0.07235657604
	0	0	-0.4553644374	-0.15625	-0.07802085054	-0.04900446818
	0	0	0.3801437869	-0.2233988914	-0.0859266686	-0.05492068733
	0	0	-0.1794572274	0.4142757111	0.156362129	0.09789851989
H_1	0.04186152385	-0.03847597749	-0.001137235601	0.0103188138	-0.02303513663	0.01434079321
	0.06311599521	-0.05809141319	-0.001718550297	0.01559368852	-0.03481655981	0.02168109372
	0.3748202832	0	0.3365384615	-0.1901521826	0.1632992753	-0.1259679427
	-0.03419477048	0.1085355724	0.4553644374	-0.15625	0.1015281078	-0.06864716186
	-0.3474452217	-0.03110245943	0.3428605425	0.04787803702	-0.1847362238	0.1397329126
	-0.03303819395	-0.2669053	0.2019668468	0.1235190581	-0.3007025809	0.2081776516

TABLE 6 (*cont.*)

K_{-2}	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	−0.0003450330824	−0.003069463328	−0.005399552433	−0.003288096492
	0	0	0.003677598183	0.03314000602	0.05933009478	0.03671227884
	0	0	−0.0002078481375	−0.001876605867	−0.003368376617	−0.002089116052
	0	0	0.0001926474622	0.001660807040	0.002792288660	0.001627445103
K_{-1}	0	0	0	0	0	0
	0	0	0	0	0	0
	0.03130736939	0.02388558244	0.04706225017	0.2101944353	0.08540876058	0
	−0.1540863084	−0.1701046849	−0.07718035836	−0.3484642221	0.5489690651	0.7074414504
	0.007173437053	0.008891783345	0.0007343402982	0.003523980761	−0.04349900544	−0.04602920876
	−0.03995938295	−0.02391005881	−0.07939816359	−0.3541468171	−0.2303288536	−0.08854025638
K_0	0	0	−0.02519015248	−0.2275254027	−0.4086088923	−0.2535449777
	0	0	−0.01623364889	−0.1466286491	−0.2633307378	−0.1634003680
	0	0.9112504122	0.00005668466137	0.0005119958472	−0.1797897710	0.2863652683
	0	0.002424886694	0.00001844956122	0.0001666429418	−0.05851745969	0.09320534726
	0	−0.2035583274	−0.05337191372	−0.4820738207	0.1967490848	0.7817811565
	0.6982782210	−0.05639870192	0.02262092952	0.2043201597	−0.4452067894	0.2472916741
K_1	0.6248862379	0	−0.001364675259	0.02429880239	−0.3319782450	0.4637528492
	0.02712562128	−0.8703879210	−0.002093878939	0.01024598177	0.1811356129	−0.3016012983
	−0.001149229936	0.0005230835228	0.000005181478463	−0.00004527100431	0.00006042145614	0
	−0.0003740480676	0.0001702517267	0.000001686452768	−0.00001473467681	0.00001966580247	0
	0.1704375211	−0.1592437059	−0.004756707896	0.04316890544	−0.09656524863	0.06030056198
	−0.07418827490	0.06820633491	0.002016324222	−0.01829536575	0.04084291440	−0.02542850505

REFERENCES

[1] G. BATTLE, *A block spin construction of ondelettes, part* I: *Lemarié functions*, Comm. Math. Phys., 110 (1987), pp. 601–615.

[2] C. DE BOOR, R. A. DEVORE, AND A. RON, *The structure of finitely generated shift-invariant spaces in $L_2(\mathbf{R}^d)$*, J. Funct. Anal., 119 (1994), pp. 37–78.

[3] ———, *On the construction of multivariate (pre)wavelets*, Constr. Approx., 9 (1993), pp. 123–166.

[4] C. K. CHUI AND J. Z. WANG, *A cardinal spline approach to wavelets*, Proc. Amer. Math. Soc., 113 (1991), pp. 785–793.

[5] I. DAUBECHIES, *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[6] S. DEMKO, W. F. MOSS, AND P. W. SMITH, *Decay rates for inverses of banded matrices*, Math. Comp., 43 (1984), pp. 491–499.

[7] G. C. DONOVAN, J. S. GERONIMO, D. P. HARDIN, AND P. R. MASSOPUST, *Construction of orthogonal wavelets using fractal interpolation functions*, SIAM J. Math. Analysis, 27 (1996), pp. 1158–1192.

[8] J. S. GERONIMO, D. P. HARDIN, AND P. R. MASSOPUST, *Fractal functions and wavelet expansions based on several scaling functions*, J. Approx. Theory, 78 (1994), pp. 373–401.

[9] J. S. GERONIMO, E. M. HARRELL II, AND W. VAN ASSCHE, *On the asymptotic distribution of eigenvalues of banded matrices*, Const. Approx., 4 (1988), pp. 403–417.

[10] T. N. T. GOODMAN, S. L. LEE, AND W. S. TANG, *Wavelets in wandering subspaces*, Trans. Amer. Math. Soc., 338 (1993), pp. 639–654.

[11] T. N. T. GOODMAN AND S. L. LEE, *Wavelets of multiplicity r*, Trans. Amer. Math. Soc., to appear.

[12] D. P. HARDIN, B. KESSLER, AND P. R. MASSOPUST, *Multiresolution analyses and fractal functions*, J. Approx. Theory, 71 (1992), pp. 104–120.

[13] L. HERVÈ, *Multi-resolution analysis of multiplicity d: Application to dyadic interpolation*, Comput. Harmonic Anal., 1 (1994), pp. 299–315.

[14] R. JIA AND C. A. MICCHELLI, *Using the refinement equation for the construction of pre-wavelets* II: *Powers of two*, in Curves and Surfaces, P. J. Laurent, A. le Méhauté, and L. L. Shumaker, eds., Academic Press, New York, 1991, pp. 209–246.

[15] R. JIA AND Z. SHEN, *Multiresolution and wavelets*, Proc. Edinburgh Math. Soc., 37 (1994), pp. 271–300.

[16] W. LAWTON, S. L. LEE, AND Z. SHEN, *An algorithm for matrix extension and wavelet construction*, Math. Comp., to appear.

[17] P. G. LEMARIÉ, *Une nouvelle base d'ondelletes de $L^2(\mathbf{R}^n)$*, J. Math. Pures Appl., 67 (1988), pp. 227–236.

[18] C. MICCHELLI, *Using the refinement equation for the construction of pre-wavelets* VI: *Shift invariant subspaces*, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., 356 (1992), pp. 213–222.

[19] M. REED AND B. SIMON, *Functional Analysis*, Academic Press, New York, 1980.

[20] A. RON, *Factorization theorems of univariate splines on regular grids*, Israel J. Math., 70 (1990), pp. 48–68.

[21] L. L. SCHUMAKER, *Spline Functions: Basic Theory*, John Wiley, New York, 1981.

[22] G. STRANG AND V. STRELA, *Short wavelets and matrix dilation equations*, IEEE Trans. SP, 43 (1995), pp. 108–115.

[23] P. P. VAIDYANATHAN, *Multirate Systems and Filter Banks*, Simon and Schuster, New York, 1993.